Acknowledgements

I like to thank my supervisor professor Andreas Jacobsson at the department of Mathematical Statistics at Lund University, for all the help he provided during this thesis. I also like to thank Jens Ålebring, for providing measurements(data) and instructions on how the microscope function, this have given me a better incite on the possibilities and limitations of the microscope. Aboma Merdasa I like to thank for providing measurements done by him in the field, Givanni Soldi for help with understanding the ellipsoid constraints and how they work.

Abstract

There has been much work on classification of malaria infected blood; in resent time, a method using LED-based microscopy has been developed with the goal of reducing time and cost. The education level needed to make such decisions is also reduced using this microscope. This is mainly done to help the developing countries in the fight against malaria and develop these countries competence in the field of multispectral analysis. The LED-microscope used, was constructed during a workshop including scientists from Lund and 6 developing countries in Africa, so there is identical equipment in the field in these countries. This could be a useful complement to the pathologists in the field. The LED- microscope uses 13 different wavelengths and 3 ways to illuminate the sample (reflecting, scattering and transmitting). Each combination is used, all in all 39 different pictures of the sample. An automatic process based on this would be a great help to simplify the detection of malaria.

The goal of this thesis is to analyze different methods of classifying malaria using data from the LED-microscope. The data was collected by already existing, and under development, methods using the LED-microscope. The main two statistical methods used in this thesis to do the classification are: First Fisher's Linear discriminant to reduce the dimensions with minimal information loss. Second Ellipsoidal constraints to formulate an optimization problem that is then rewritten as a convex optimization problem, which is then solved. Then the classification is analyzed to conclude if the samples contain malaria, and to find suitable thresholds for the classifiers, every analyzed method is evaluated. Even when this is mainly software optimization it can impact the work needed to construct new microscopes to make it more efficient. Also contained in this thesis are some examples from analysis of both fresh and old malaria-infected blood samples to see the difference between the different methods. The methods is quite general and can with little extra work be applied to different data sets, when the microscope is used to gather data from other sources than blood samples.

Contents

List of Figures

List of Tables

1 Background

1.1 Problem/Malaria

In the developing countries malaria is a massive problem; nearly half of the world's population is in danger of being exposed to the parasite. Yearly about 250 million people get infected and of these nearly one million die [1]. Much research is done in the area of trying to prevent the spread of and to detect the disease. Malaria spread is done by mosquitoes that bite an infected host, drinking their blood. When the mosquito then bites a healthy victim it injects a bit of "saliva" to stop the blood from coagulate. Together with the saliva there is a risk that the parasite gets into the victims blood. Malaria can be detected by looking at red blood cells (RBC) to see if it contains any malaria parasites. Today the main detection of malaria is done by microscopy, by chemical colouring the blood samples and then have a pathologist look at it, to try to decide if it contains malaria.[2] The chemical reacts with the malaria parasites and makes it more visible. As the pigments reacts with the parasites, it helps the pathologist to see them. Of late there has been some research on how to make this in a more automatic fashion where neither chemicals will be used nor a cunning person need to look at it. Some of this work has focused on using a LED-microscope, i.e. a microscope that can illuminate the blood sample with singled out wavelengths to try to see if the parasite is distinguished by some of specific chosen wavelengths distributed in the range of 375-940nm. Also to make it more efficient, and since humans cannot see all wavelengths, a computer is the one to make the decisions. Not only do the microscope use different wavelengths it also illuminates the blood sample from different directions. Using reflecting, transmitting and scattering light to see how the different wavelengths react to the parasite and it's surroundings[3]. So far the actual creation of the microscope and deciding how to exactly do the measurements has been done and some try outs of different methods have been tried. Methods varying from just looking at a picture or graph to a Singular value decomposition (SVD) approach. One of the goals with this project is to make a fully automatic process, where just one blood sample is needed. The vision is that the microscope takes the pictures then the only thing needed is to hit a button that says malaria and then get a green or a red light. Optimally then the microscope could be used to detect other diseases or infections as well. Or a specialised version of the microscope, specialized in such a way that when there is known what pictures are the best to use for detecting malaria, one can either build a microscope that only can do just that or maybe just slim the calculation process. This Thesis have the goal to try different statistical methods to detect or separate plasmodium falciparum (malaria) from a blood sample.

1.2 The Collaboration

This is part of a larger ongoing project that was founded in Collaboration between scientist from Ghana, Cote D'Ivoire, Mali, Senegal, Kenya, and Sri Lanka and the division of Atomic Physics, Lund University. The Collaboration was funded by the International Science Programme (ISP). [4] This collaboration started as the means to develop and apply the microscope and the multispectral microscopy from the microscope created by Mikkel Brydegaard [5].

There are some other work done by this Collaboration to try to help in the fight against malaria. One example is using a similar idea as the LED-microscopy but instead of using LED for lighting one uses the indirected light. This can be for trying to detect the mosquitoes that are the carriers of the malaria parasites in one of it's stages. Mosquitoes are one of the big transporters of malaria between other mammals and humans. Measures of how the mosquito fly is a good way to try to map malaria spread. Some work have been done i this field of trying to detect insects with indirect illumination. One paper that describes this a bit further is "rare Events in Remote Dark Field Spectroscopy: An Ecological Case study of Insects [6]. Another use for the same setup as for the malaria detection are the detection of pollen and skin cancer [7]. Or the detection of polluted water from far away. All these methods of detection can use an method for separation, it might be that the existing ones is good enough but it is always good to to explore some alternatives.

In this workshop the LED-microscope where created and was very versatile as to try it in different areas to see how it works. And then later on decide if a simpler constructed version of the microscope is to detect for example malaria can be constructed for this sole purpose or in combination with some other blood born disease. This can be done by removing hardware parts of the microscope fixating other or by having dedicated software that only do one thing. This should mainly be for the purpose of simplifying the usage, constructions and or costs. And also what other hardware can be added so that the microscope can be used to try some nearly related detection methods.

2 Method

2.1 The LED-microscope

The LED-Microscope was developed in collaboration with scientists from 6 developing countries (Ghana, Cote D'Ivoire, Mali, Senegal, Kenya, and Sri Lanka) and the division of Atomic Physics, Lund University. This collaboration started with a two weeks workshop in University of Cape Coast (Cape Coast, Ghana) where the microscopes were constructed. This led to other workshops that have continued until today's date. The goal of the collaboration was the construction of a microscope, educate and explore the usage of multi spectral microscopy. There is a great need for this kind of research in the developing part of the world. [3] [4] The LED-Microscope that can be seen in figure 1 were constructed using a sufficient but cheap regular metallurgical microscope, where all the lighting was replaced with LED-lights able to emit 13 different wavelengths or spectral bands in three geometries; reflection (S1), scattering (S3) and transmittance (S2). The ocular was replaced by a camera that could perceive light in all the different wavelengths. The camera do not perform as good with all wavelengths as can be seen in figure 2, showing the cameras sensitivity to read the different wavelengths. The 13 smaller peeks in that figure is the emission from the different LED lights. The large curve in the graph is the camera sensitivity over the different. One aspect that also where sought after was to eliminate chromatic aberration. Because different wavelength behaves differently when passing through glass this would make the focus point to be different for the different wavelengths and therefore the pictures would not overlap perfect. Why this is important has to do with how the data space is constructed and will be explained later. To solve the problem with chromatic aberration all the optics of the microscope were replaced with mirrors because the reflection of different coloured light reflects at the same angle but when light passes though a lens different wavelength refract differently. In this case the same focus can be used for all the pictures and do not need to be adjusted between the different spectral bands lengths. This eliminates the need to line up the pictures or to scale them to get a perfect overlap. If, however, there microscope had been using lenses, that are cheaper than mirror oculars, the focus point would have moved when changing light [4]. The microscope and/or the methods it uses, can be used to do many other analyzes, due to that it is very generally build in its set-up. The reason for that is for it to be able to approach a big variety of problems. There are some thoughts on how to make the microscope more specific for use in a special case. For example the microscope can be used to try every different wavelength and geometry for detection of something i.e. malaria. Then when the detection has been done some analysis can be performed to see how much one is able to scale down without a significant loss of information. If the market is big enough one can then build a microscope that is just to do exactly those measurements and also fixate as much in the hard- and software as possible make it easy to use.

The 13×3 pictures are taken in a sequence and over a specific area. The area should be one that has a single layer of RBCs. The sequence shot is done because it is of the essence that all pixels in the different measurement are aligned when the space of measurements is constructed. This is performed to avoid moving the sample during microscopy. All the settings will have to be done beforehand and then the microscope will take all the pictures needed. If the pictures are aligned then there will be no problem with trying to align them afterwards, this can otherwise be a somewhat hard task, and removing the need for it is much simpler then the problem it causes. Reference pictures are taken for elimination of mechanical differences. The references are dark and bright, the dark references are done by using no light from the diodes. The dark reference is for catching outside lights or other disturbances. The bright reference uses no sample in the microscope and for the scattering geometry there is an evenly scattering opal. During the bright reference sampling the light is on. A further expatiation on how this is done can be seen in the paper "Versatile multispectral microscope based on light emitting diodes"[3]. This process is done for each specific part of the sample that is of interest to collect data from. The picture is taken at a part of the blood sample with a single layer of RBC to get no or as few as possible overlaps of RBC. This because the detection of malaria becomes much harder for the detection algorithm if the RBCs are

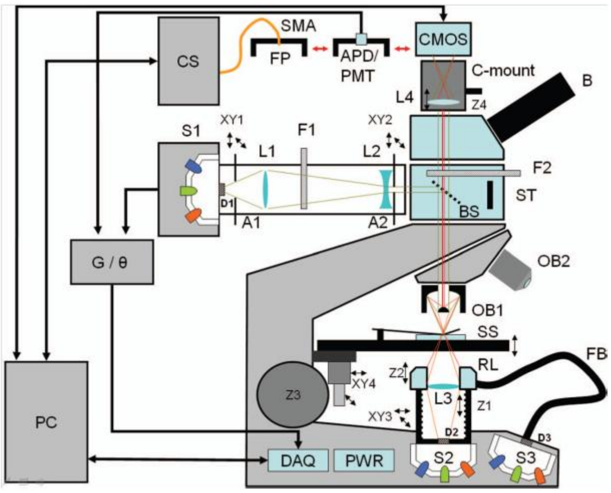


Figure 1:
Overview of the LED-microscope, The microscope have 3 illumination sources one for each way(reflecting(S1), scattering(S3) and transmitting(S2)) to illuminate the sample stage(SS). Each light source can produce 13 different wavelengths from the UV to the NIR region. Here to avoid chromatic aberration a reflective objective is used(OB1). The camera used for the detection is a monochromatic (CMOS) camera. This picture is taken from [3]. P 2

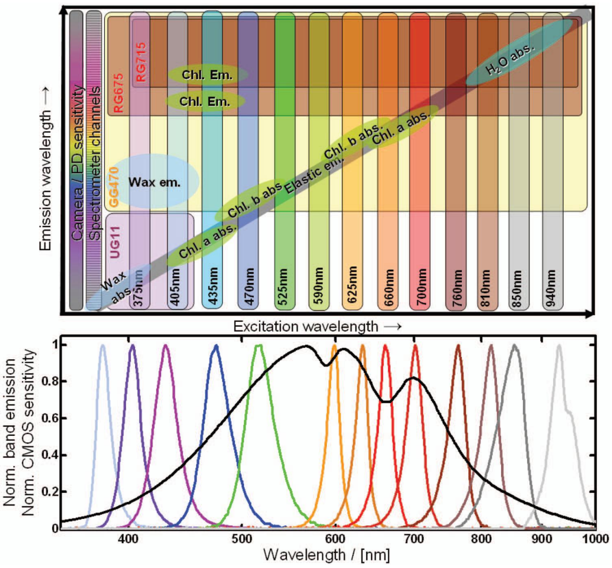


Figure 2: The lower part of the image describes the camera sensitivity (Black wide curve)in the different light regions, the 13 peaks represent the different wavelengths emitted by the diodes. This picture is taken from [3]. P 3

overlapping. This is due to the fact that the light is affected by the RBCs and the light from an parasite might change considerably if it have to pass through several other RBCs. As opposed to how the methods is done by a pathologist where in the counting of the RBC while affected by the chemical colouring is done in much thicker layer of RBCs because with the pigment the parasites are clearly visible.

2.2 Data

The data for the analysis is in the form of photos from the microscope described above. The pictures should be taken at a point in the blood sample that contains only a single layer or RBCs. To note also is that the malaria parasites can be detected with warring level of difficulty in the RBC. This depends on where in the cell it is located and how far it is in it development. The placement is considered to be of two main types depending on where the parasites are located. They are located either in the almost empty inner part of the RBC or close to or inside the wall. Hence forth they will be referred to as "in-cell" parasites respective wall parasites. In this thesis the focus is on the parasites in the inner parts of the RBCs. Due mostly to the lack of parasites in the cell walls in the sampling data. They also differ from the "in-cell" parasites, this differences depends on that the cell wall will add a disturbance of quite large significance, making the parasite harder to detect. And when these parasites will look different and the goal is to separate out the parasites it is best not to mix the two types of parasites. Due to the fact that there are more of the "in-cell" parasites then the wall parasites the focus will be on finding the "in-cell" parasites.

The blood samples used in this method is from Jens Ålebring [7] and from Aboma Merdasa[4]. J. Ålebring's measurements was taken with the decided standard but the blood sample was a bit old so it affect some of the optic behaviour and therefore differ somewhat from a photo taken with the same standard but with a fresh sample. J. Ålebring samples had been stored for many years and had then been frozen this will affect the possibility to compare this with other fresh samples. But the method for detecting and separating the infected parts is the same even if the thresholds and weights will not be. The biggest problem with this data is that its differs from fresh data and even other old data. This makes it harder to do any comparison good between blood samples. The samples dry up over time and to be frozen and transportation also can have effect on the sample especially sins this sample was without cover glass.

A. Merdasa's photos was taken from a fresh blood sample down in Africa [4], but on the other hand this was before the standard of how to take photos was decided. Therefore some differences between photos occur. The greatest problem when comparing photos is that the height of the placement of the part in the microscope illuminating with scattering light not decided. This part can differ a lot more then the pictures taken with reflecting or transmitting light as they stay more fixed with angles and such. And also the photos were taken by A. Merdasa in the beginning of his thesis when he was down in Africa so the sample quality differ between the different photos.

It was concluded that the current data would be sufficient to do evaluation of different approaches. However before trying to do any classifications in Africa a standard method for taking the data would have to be in place. Then some collection of new data can be done to adjust the methods for detecting malaria according to this standard. But the method to perform the classification could be the same. Just that the training data would be new.

2.2.1 Processing the data

When the data comes from the camera it is in 12 bits and actually has a larger resolution than what is possible for the optics in the microscope. The picture also contains salt and pepper noise. The fact that the camera sensitivity is not equally good everywhere and some disturbance light might come in to it also cause some problems. This makes the wavelength where the camera is not as sensitive to need longer exposure to give a good picture or a brighter illumination. This will have the effect that salt and pepper noises in these regions will be greater, and over all the signal to noise ratio (SNR) will be lower, giving us a bigger variance of those samples. If the difference in these dimensions would be small one might suspect that these regions will not be of much use.

There are also some other disturbances like light not coming from the camera, even thou the photos are taken in such a manner that to reduced that light, there is always a possibility that some unwanted light might find its way into the sample. Here it is assumed that this light will be in a short time frame static. There can also be some disturbance over the camera or particles on the mirrors. The way to minimize these errors has been by taking a dark and a bright reference and use them to normalize the microscope. The purpose is to be able to remove static interference so that this does not follow to the data that is then to be analysed. This is done mainly to get the possibility to compare results from different microscopes and measurement time of the same microscope. This is also performed to get close to the same result if the same sample was to be measured at different times. It can also be applied by using known samples and their known behaviour to detect parasites in other samples.

The dark reference and the bright one is captured along with the 13×3 different pictures. Then the pictures are normalized with the references to remove any mechanic bias. Due to the resolution problem the data resolution is reduced using a binning process with size 4×4 i.e. each 4 by 4 square is reduced to one new pixel. This is mainly done to come into the theoretical resolution of the optics. Since it makes little sense to analyse data with higher resolution than can be observed.

The data is now formed as a 13×3 or 13×2 dimensional space with respect to each pixel. For the 13×2 dimensional the measurement is performed without the reflecting geometry and thereby increasing the signal to noise ratio by 2. This is due to the fact that the camera is situated behind a 50% reflecting mirror and when the light from the source hits the mirror half of it is reflected and goes to the sample. The other half goes strait through the mirrors and is absorbed. The reflected part then hits the sample and the light that reflects back up hits the half reflecting surface where once again it is split into two beams: One that goes to the camera and one that goes back towards the light source. The inside of the microscope is built to absorb any stray light, so here there is 2 losses of light at half the light. This means the signal is down to a forth. For scattering and transmitting light parts they just move through the mirror once so then they lose half of the light. Giving if one were to remove the reflecting part the SNR would double, giving a stronger signal for the transmitting and scattering geometries. [7]

Let's assume that there is a 39 dimensional space to cover all the data from the photos. Then a point x_i in that space is built up by taking the values of a pixel at a specific coordinate in each of the different pictures. Here it is of utter importance that all the pictures are aliened and in the same focal point since otherwise this point would not describe the same physical spot at the blood sample with all its values. This is why when taking the pictures one take them all in sequence as to not have to move the sample or worrying that it has been moved during the process. The mirrors instead of lenses insure that the focal point and there by the magnification stays the same. This can be seen as taking the 39 pictures and align them on top of each other. Then choosing a point and draw a line through all the pictures and then chooses the values of x_i as to be the value where the line cut the pictures.

Let there be a total data set

$$\mathbf{x} \in \mathbb{R}^d \tag{1}$$

that is describes as a matrix with the columns as the points x_i, where d is the number of pictures used, i.e. the dimension of the space. Further using a validation set that describes what parts in the blood sample that are malaria parasites. This validation set has been done in the regular method with chemical colouring and a pathologist to know what part of the data that is malaria. Then a pixel that is a malaria parasite can be chosen, the point x_i describes that malaria parasite in the 13×3 dimensional space. A parasite exists on a few pixels only. Due to that it is quite hard to decide what pixels that actually are malaria even with the knowledge, of which RBCs that are infected, from the validated pictures by the pathologist. A problem that can occur is that if a pixel thought to be malaria is chosen, actually not is malaria. This problem produces some errors when trying to make the classifier. This is due to the fact that these outliers might impact the whole procedure a lot.

Different data sets can be created by choosing, for example, all the points that are malaria and put them in a set $\mathbf{x}_{(pa)}$ and let that be the set of malaria parasites. In the same manner sets of clean inner cells can be chosen as $\mathbf{x}_{(in)}$ and a set containing RBC wall segments as $\mathbf{x}_{(wa)}$. This can be performed for all the number of classes that might fit. For the creation of the classifiers i.e. the mechanics that chooses what cluster a point later on will belong to, the selection is done by hand and not automatic. This makes room for the human error but the training data must be chosen somehow. One other problem while choosing the pixels that are to represent the malaria, due to the small size, is that there are pixels that are partly malaria and partly something else. For example the inner part of the cell. Those pixels might also put the classification in some trouble when the pixels values probably will lie somewhere between the classes. So while making the classification algorithm, the points that represents this kind of pixels should be avoided as much as possible from being in the classification.

Each dimension refers to a wavelength and either transmission, scattering or reflecting light. Figure 3 shows a picture with scattering light of wavelength 435nm. Figure 4 shows a regular photo of the blood sample after coloured with chemicals to make the parasite more visible.

Not all of the dimensions gives as good pictures as the one in Figure 3. In Figure 5 the same RBC can be seen from all the different wavelengths. As noticed, in some of the pictures the RBC can hardly be seen but in other pictures, what seems to be a parasite is quite clearly visible.

A space in 39-dimensions is build by taking the pixels with the same position in the 13×3 pictures and adds them as a point in the space. I.e a point is built up by the different pixel values along a line that cut the pictures in the same coordinates. It is these points that are to be classified. By confronting the validated data set and looking at the different points in the space it is now possible to do a rough pic out of what pixels that are malaria and what pixels are not.

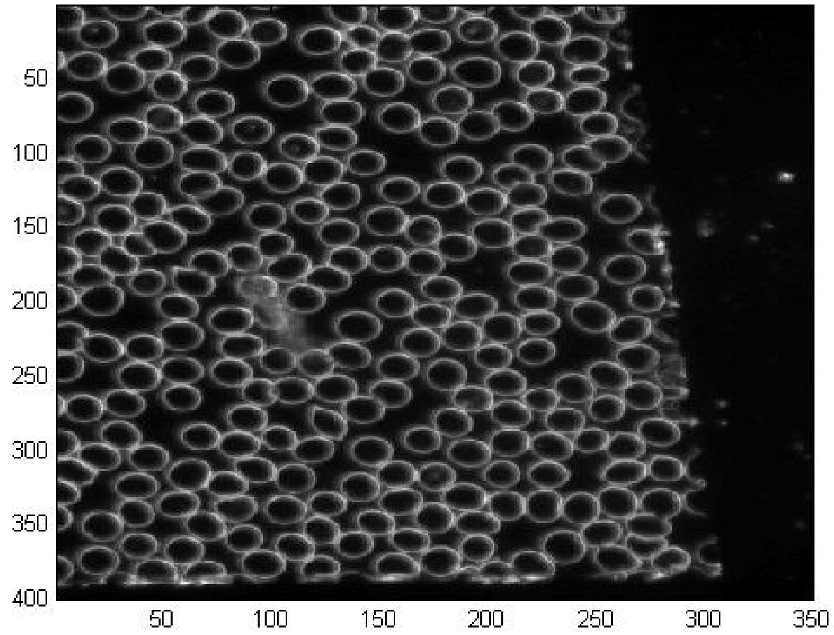


Figure 3: Photo of a blood sample at 435nm wavelength and scattering light.

3 Algorithm

3.1 Methods from earlier reports, SVD

There have been other methods trying to analyse the information from the microscope. One where for each geometry the two pictures/wavelengths that had the most contrast while looking at an infected blood cell and healthy one, was used to create a contrast index by dividing the two pictures points. From this the pair that had the greatest index were chosen. For each of these pairs a threshold where chosen to get a value as to decide if this could be malaria or not. With this a 3 dimensional space is created, one dimension for each geometry. If the data point is on the malaria side of each of the three thresholds it can be considered to be malaria. But if anyone of the three dimensions fail, it is considered not to be malaria. [4]

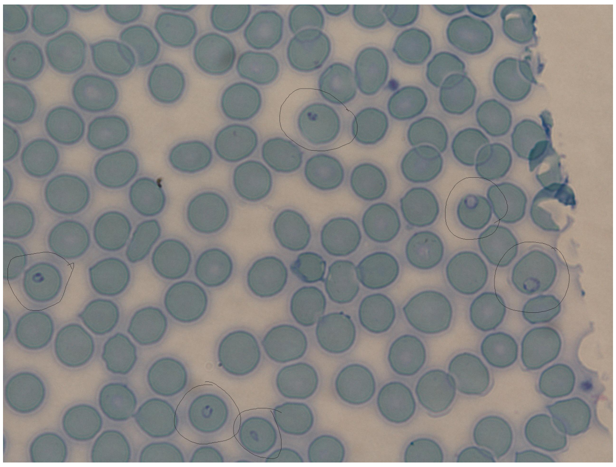


Figure 4: Photo of validated data with chemically identified malaria parasites. The circles around a blood cell marks an identified parasite infected cell. Here the edges of the parasites are clearly visible due to chemical colouring.

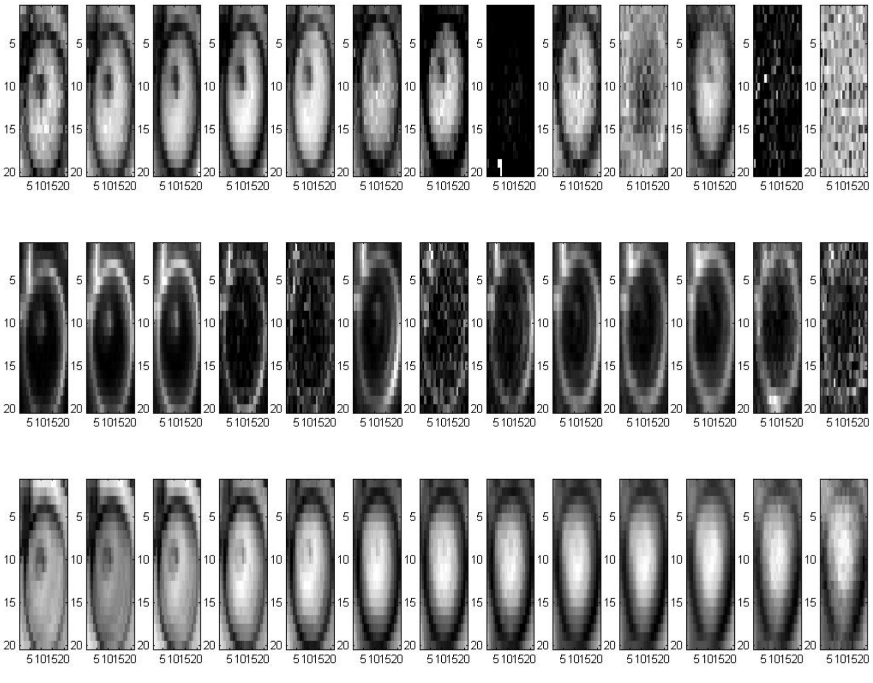


Figure 5: 13 × 3 pictures of each wavelength and geometry. Each column is a specific wavelength and each row represent a geometry. As can be seen some of the geometries have a really high noise level while others looks really fine.

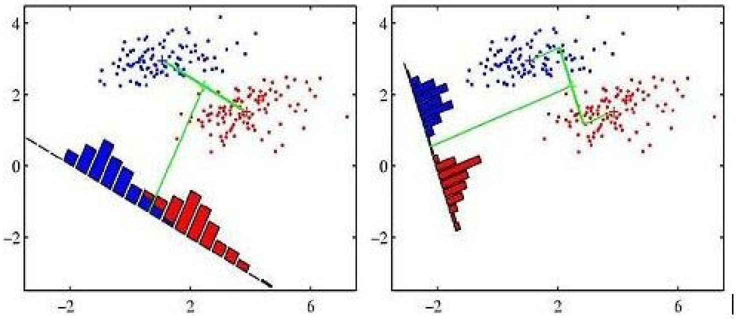


Figure 6: The left picture shows a standard projection on the mean of two classes which gives a considerable overlap of the classes after the projection. The right picture describes the projection done by using Fisher's Linear discriminant to calculate the optimal projection direction. The classes is much more separated in the right picture then they are in the left. This clearly shows the advantage of using Fishers method. Figures from C. M. Bishop's book [8]

Another approach that was used was the use of singular value decomposition (SVD). In this the variance of the images is analysed to decide from which parts of the data the information comes from. All the wavelengths were concatenated with the rest from that geometry (i.e. transmittance, reflectance and scattering). From analysing the eigenvalues, after the decomposition with the SVD, the conclusion was that 75% of the variance were from 3 of the 39 pictures used. [3]

3.2 Fisher's Linear Discriminant

In Fisher's method the idea is to take the space that is spanned by the features given by each picture and project it down on a smaller space. In such way that the loss of information is as small as possible. This is done by finding the cut of the space so that the clusters are as far apart as possible. If this is done in a normal fashion by finding the cluster centres of the two classes and then make a line that passes though both of the centres and then project onto this line there is a great risk of an unnecessary big overlap between the two clusters. The example in figure 6 is enough to be certain that this is actually the case. The figure describes two clusters that have a ellipsoid like distribution appearance so when just projecting on the line connecting the centres there is a significant overlap as can be seen in the left plot. In the right plot while using Fishers method the two clusters are completely separated. This is due to that a line can be found which clearly separates the clusters. The idea is to find the projection that minimizes the information loss when reducing the dimensions, this in the sense of maximal separation. To note if both the classes would have a Gaussian distribution, equal in all directions, the mean method and Fisher's would have been the same.

One drawback with this method is that if the wanted data, which is to be separated, is surrounded by or have clusters/points on both sides. Then there is no way to do a linear cut in the space. Thereby it will not be a way to find a good projection, and this will probably lead to a significant overlap of the clusters. Instead a way to separate the surroundings into more clusters/classes can be found. Instead of dividing the data into the malaria data and not malaria data, the data is divided into; malaria, inside of RBC, wall of RBC and outside of RBC(glass). In some cases there are also some other parts of the picture that are not accounted for, but they should be way off from any clusters. When this is not the case a fifth or sixth cluster might be considered to overcome this.

One way to get around the problem with unknown parts outside of the RBCs could be to only focus on the RBC. This can be done by finding the inside of the RBCs and only compare the points that are inside of a RBC. That is the area of interest since that is where the malaria parasite is located with the exception if the parasite is in the cell wall. Points looking like parasites outside of the cells are of no interest, because the parasites do not exists there, and can therefore be excluded. But this would mean some extra work, so first the focus will be to see if it's possible to separate the classes with the only concern that the data is from a blood sample taken by the microscope. To do this the method of Fisher's linear discriminant is used. This is actually not a discriminant per say but a way to find the projection or the cut that separates the clusters the most. Then thresholds can be added to adjust where the line separating the clusters will be. It is first after that it is considered a discriminant.

The following is results from [8]. For two classes the goal is to project the sample space down to only one dimension using the projection

$$\mathbf{y} = \mathbf{w}^T \mathbf{x} \tag{2}$$

where **x** is the input vector in D-dimensions, **y** is the result vector and **w** is the weights that decide how the projection is done to achieve minimal loss. **w** is to be constraint of unit length, otherwise the distance between the classes after the projection could be made arbitrary large. The same projection is used on more than 2 classes but then the projection will be down on a hyper plane because **w** will be a matrix. The **x** here is the data set described in equation (1) or any data that is to be tested. Since (2) is linear it follows that if the data is well separated after the projection, it was also well separated before. It might be much simpler and computationally less demanding to find a separation in the projected set **y** since it is of lesser dimension. To keep in mind regarding projections, also at this projection (2) there is information loss. But due to the smart way of choosing the projection this loss is minimized. Although there is always a risk that clusters well separated in the higher dimensions not are well separated after the projection. This is particularly the case if it do not exists a linear cut between the classes. The lack of a linear cut can be for different reasons, either the parasites are inside/(mixed to) of the class that it should be separated from and then no separation is possible. Or the class one like to separate is bent around the class that is to be separated. In the later case this might be solved with using a non linear approach to solve the separation problem. Or it might be possible to split the second class in to smaller classes and then try to find a linear cut between all of the new classes and the one that is to be separated.

To find the best separating cut i.e. the projection that minimize the information loss regarding the separation of K classes the mean vector $\mathbf{m}_k$ for each class is calculated

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in C_k} \mathbf{x}_n \tag{3}$$

and also the mean for all the values used in the separation process is to be used

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N} \sum_{1}^{K} N_k \mathbf{m}_k \tag{4}$$

Here N_k is the number of points in respectively class. One thing to keep in mind is that, if for example only two classes is used to find the separation, this will lead to a good way to separating these classes after a well chosen threshold that depends on the margin of error. There is a problem when trying to classify a point, if it should not belong to any of the data that was used to create the weights **w** then there is a risk that point could be classified as either. This could cause much damage. The reason for this is that when a point is not in the exclusion set there is always a risk that this point will not be well separated from the classes. Ideally a point should be far off from any of the classes and be classified as neither, but testing has shown that this is not always the case. Therefore the data used to creating the separation should be the same that later are to be classified. That means that a classifier that only separated $\mathbf{x}_{(pa)}$ from $\mathbf{x}_{(in)}$, that is really good at that, might be really bad if one also try to separate the walls. Therefore before creating the weights is important to know how well specified the input data will be. If it is just inner parts of the blood cell or the whole blood sample.

The covariances between and within the classes is to be used to decide what angle the projection is going to have as to minimize the information loss or maximize the separation. The within class and between class covariance matrixes is calculated. The within class covariance matrix is calculated by summing up all the covariances as

$$\begin{aligned} S_W &= \sum_{k=1}^{K} S_k \qquad (5) \\ &= \sum_{k=1}^{K} \sum_{n \in C_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T \qquad (6) \end{aligned}$$

where and $\mathbf{m}_k$ is given by (3). The between classes covariance is calculated by

$$S_B = \sum_{k=1}^{K} N_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T \tag{7}$$

In the above equations (4) - (7) K is the number of classes that should be separated, N is the total number of points in the space. N_k is the number of points in a specific class k and C_k is all the points in class k.

There are many approaches to decide how to optimise the weight function in (2). One of them is to use the covariances to decide the weight matrix **w**. The columns in **w** can be set to the eigenvectors with the largest eigenvalues of

$$S_W^{-1} S_B \tag{8}$$

[9].

The number of eigenvectors to use depends on the number of classes that is to be separated. The number eigenvectors can be at most one less than the number of classes. This is due to the fact that there are only so many non-zero eigenvalues because S_B consists of the sum of K matrixes of rank 1 because they are the outer product of two vectors. Max K-1 of them can be linearly independent due to the constraint formed by equation (4)[9].

A further explanation to can be found in Bishops book[8] pages 191-192.

When the matrix **w** is created it is of importance that the choices of the points in the training data are exact. Meaning that the points should all be correctly chosen for the classifier and that there should be a minimum of outliers. Outliers or wrongly chosen points could otherwise make the projection skew in a manner of speaking. Thresholds are used to decide the where the region for each cluster will be, this to try to classify the less well defined points and regulate the accepted ratio of wrong classification.

When the projection in equation (2) is applied to the data, a space of dimension one less than the number of classes is created. In this space it is hopefully be easier to see if the clusters are separated or not. The more separated they are the smaller is the risk for wrongly made classifications. When the data point $\hat{\mathbf{x}}$ wanted to classified is inserted into (2) the projected point $\hat{\mathbf{y}}$ should be close to the cluster projection it belongs to, if the data points belongs one of the expected cluster. Other points from this that are a bit more far away from any clusters is left to classify.

Here **x** is a matrix build up by columns, where each column represents a point in the space containing the data that is to be projected. This gives after applying the projection (2) a new matrix **y** where each column now is the respective point after the projection.

3.2.1 Testing the implementation

To test the implementation a test data set **x** is constructed by simulating different clusters. First the number of classes K, the amount of feature points and the amount of data N is chosen. The center value is chosen at random for each coordinate of the points. Then, with a chosen variance, points are randomly simulated until a satisfying amount of data has been simulated. When all classes are simulated they are put into the algorithm described in section 3.2 to get the projection matrix **w**. Secondly new points are simulated by the same distribution as the data that was used to create the projection matrix. These points are now projected and if the implementation is correct, these new points should be nicely separated. As can be seen in figure 7 an example of data after projection it can be seen well separated.

To visualize the points after the projection is done, depending on how many dimensions it now has. The dimension is always one less than the number of classes used to separate the points into. It is easy to visualize if the dimension is one or two even three, but then it gets a little bit harder. One way is to have multiple images with one, two or three dimensions in each. A simpler solution is to skip the visualization of the clustering altogether, but with some risk of missing it, if it did not perform so well in this step and instead go strait for the classifications. After all points (or the points chosen to) in a picture, is classified, they can be represented by just a colour or by right or wrong. The last way of visualization requires that a picture-set is classified and not lists consisting of points from respectively class.

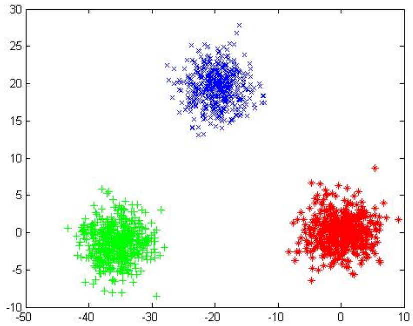


Figure 7: This figure shows how projection can look like after three randomly created classes, has run through the implementation of the Fisher's liner discriminant stated above, here the classes is shown with either blue 'x', green '+' or red '*'. And here they are very well separated.

3.2.2 Two classes

For two classes the goal is to project the sample space down to only one dimension using the projection (2). By finding a suitable threshold, aiming for that the data that is to be separated from each other, is on different sides of the threshold as far apart as possible.

First the mean of each class is calculated and the between (S_B) and within (S_W) class matrices are calculated as in (7) and (5). The projection matrix(vector) can then be retrieved from the eigenvector with the largest eigenvalue of the matrix $S_W^{-1}S_B$. To be noted is that when only two classes are used there is only one eigenvector that is non zero. All the points can then be classified using the projection. A threshold is chosen and all values larger than this threshold are classified as one class and all values smaller than the threshold as the other class. If the classes is well separated the threshold should be chosen somewhere between them, which is done quite easily. Problem to chose where to set the threshold occurs if the classes are overlapping. Then the potential damage must be considered and what consequences that can occur if the classifier is wrong in one way or the other. I.e classifying 5% of malaria cells as not malaria might be acceptable due to the fact that the patient will be classified as infected anyway. On the other hand if it renders the patient to be classified as healthy, a false negative result, the cost could be large. If the classifying judge to many patient as sick when they are not, it would increase the treatment costs and the risk that the parasites become more resistant towards the treatment.

3.2.3 More than two classes

If there are more than two classes that are to be separated, the algorithm is modified so it takes the given amount of classes. It then returns an $N \times M$ matrix, where N is the number of features and M is one less than the number of clusters. Apply the projection then a space that is one less than the number of clusters is retrieved. For example with only malaria and non malaria tests, a one-dimensional space is found and then uses the threshold as above. If there are three clusters a two-dimensional space is found as can be seen in Figure10. On example data, used for checking the consistency of the algorithm, it is easy to see which cluster belongs where, but with real data this gets harder. To decide which cluster the points are closest to, there are some different approaches. One is just to look in every direction to see which cluster center point is closest, i.e. manhattan or squared distance. Since the clusters have different spread in different directions this can be somewhat misleading if the spread is not taken into account. To fix this one can use the Mahalanobis distance [10]. This means that the width or the variance in different direction decides how one should calculate the distance so the outskirts of the cluster in all directions is at the same distance from the center. If the variation is small, a point is further away than if the variation is high. In figure 8 the distance from the cluster center to the red circles would be $\sqrt{2}$ if using a regular Euclidean distance. Using a Mahalanobis distance instead the distance around 1 respective 20 are found. Therefore the Mahalanobis distance is much preferred if the clusters are not equally distributed in all directions.

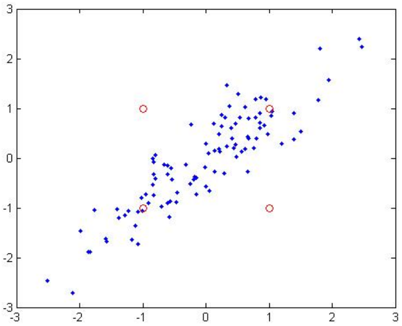


Figure 8: The distance from the cluster center, built up by the blue points, to the red circles would be around $\sqrt{2}$ if using Euclidean distance. If however using Mahalanobis distance, the distant to the upper right and lower left would be around 1 and around 20 to the other two circles

The same equation 2 is used for more than two clusters only here **w** is an matrix instead of an vector. So the projection is no longer on a line but rather on the on a space of dimension $d = K - 1$ where K is the number of classes. There is still a possibility to project down on a line but then there will be problems when trying to discern the different clusters. What is done is that when the between and within covariances is calculated in the manner described in equation 3 to 7. Now when the **w** is to be chosen it is done so by the d largest eigenvalues from equation 8 and not just the largest as in the case for only two classes. The eigenvectors corresponding to those eigenvalues are used as the columns of **w**. So when projecting each point will now have d values because of this.

Using the mahalanobis distance to decide what cluster a projected point $y = (y_1 y_2 ... y_{K-1})$ belongs to first the mean μ_k of each of the clusters gained after the projection is calculated by

$$\mu_k = \frac{1}{N_k} \sum_{n \in C_k} \mathbf{y}_{k,n} \tag{9}$$

An covariance matrix $S_{(y)k}$ is calculated for each projected cluster of $\mathbf{y}_k$ by

$$S_{(y)k} = \sum_{n \in C_k} (\mathbf{y}_n - \mu_k)(\mathbf{y}_n - \mu_k)^T \tag{10}$$

The distance D_k to each of the classes cluster centres is calculated as follows

$$D_k(y) = \sqrt{(y - \mu_k)^T S_{(y)k}^{-1} (y - \mu_k)} \tag{11}$$

where y is the projected point x that the distance will be calculated for. When the distance to each class has been calculated the point is classified as to belong to the class where the distance is shortest. Also the distance can be adjusted with a threshold used to decide if the point is below the distance threshold th_k for a certain class it will be classified as that class. This is so that one might regulate the classification of an point towards or away from a specific cluster as to get the classification right.

3.2.4 Excluding false positives

After the corresponding clusters are found the problem to decide if the points classified as malaria actually is malaria or false positives occurs. One way to solve this problem is to adjust the parameters that say how to calculate the distance to the centres in such a way that the threshold for malaria moves. If assumed that the positive malaria points should be more than one point at the same spot, points that not are in the vicinity of any other malaria point can be excluded. One obvious thing for human eye is that a point outside of RBC, that indicate malaria, is false positive since the parasite must be in the cell, but this is not obvious for the computer. One way to get around that is to first find all the RBCs and only look at them, either only on the inside or both on the inside and in the wall.

In the data there is too few parasites in the cell walls to do anything with statistic relevance. Much of the data was not overlapping; this also is a problem as it reduces the significance of the results.

3.3 Ellipsoids/ellipse

The idea with the Ellipsoid approach is to find ellipsoids around the classes/clusters in such a way that the ellipsods can be used to separate the different classes [11][12].

Two ellipsods is constructed both with the same center and directions, but the later one is larger by a factor ρ. The goal is that the inner ellipsoid is to enclose all the data for that class and the exterior is to exclude all the other data not in the class. In all cases it might not be possible to do this distinction and then there might be some overlap. This can cause the value of ρ to be smaller than 1. The higher the value of ρ is the better is the separation. When the classes not are separable one might introduce slack variables that have weighted penalty functions. This is done to allow for some points to be on the wrong side of an ellipsoid, but the more wrong the placement is the higher is the cost. Although allowing for some wrongly placed points the distance to the outer ellipsoid still is forced to be ρ times the inner ellipsoid . The inner ellipsoid is constructed so that it covers most of the essential points while having as few other points inside, or some of its own points outside of the inner ellipsoid.

An ellipsoid is defined as

$$\varepsilon_d(\mu, P) = \{x \in \mathbb{R}^d | (x-\mu)^T \mathbf{P}(x-\mu) \leq 1\} \tag{12}$$

where μ is the center of the ellipsoid and **P** is a matrix describing the size and shape of the ellipsoid. **P** is a symmetric, positive semidefinite matrix.

The problem to solve with strict separation is stated as

$$\begin{array}{rc} \text{maximize} & \rho_k \\ \text{subject to} & (x_i-\mu_k)^T \mathbf{P}_k (x_i-\mu_k) < 1, \\ & \forall i : c_i = k \\ & (x_i-\mu_k)^T P_k (x_i-\mu_k) \geq \rho_k, \\ & \forall i : c_i \neq k, \\ & P_k \succeq 0 \\ & \rho_k \geq 1. \end{array} \tag{13}$$

Here the first row is the ellipsoid and it states that for a given point in our class, stated by row two, that point should lie within the ellipsoid. The value after multiplying the point to both sides of the ellipsis is supposed to be less than one. In the same manner the third and fourth row states that any point outside of the class should be outside of the outer ellipsoid. I.e. the distance should be greater than ρ. The fifth row tells us that the matrix **P** to perform this and build up the ellipsoid has to be positive semidefinite. This means that for any given point **x** that

$$x^T \mathbf{P} x \geq 0 \tag{14}$$

and this is quite an obvious constraint due to the fact that in the data space any given point can occur and a negative distance to the ellipsoid is not possible. One thing that can be noted about **P** is that it is symmetric, so that the ellipsoid looks the same on both sides of the center. Otherwise it would not have been an ellipsoid.

This problem can be stated as a SDP if it is changed to a convex problem and then solved by, for example using cvx package in Matlab, for solving the SDP. This is done by a homogeneous embedding technique. Meaning that the dimension of the data is increased by one. This leads to the increase of the dimension anywhere. With addition of the slack variables, the problem will be stated as the following convex problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \eta_{ik} \\ \text{subject to} & z_i^T \Phi_k z_i \leq 1 + \eta_{ik}, \\ & \forall i : c_i = k \\ & z_i^T \Phi_k z_i \geq \rho - \eta_{ik} \\ & \forall i : c_i \neq k, \\ & \Phi_k \succeq 0 \\ & \eta_{ik} \geq 0, i = 1, ..., M. \end{array} \tag{15}$$

where

$$\Phi_k = \begin{pmatrix} \mathbf{P}_k & \mathbf{q} \\ \mathbf{q}^T & r \end{pmatrix} \tag{16}$$

.

Where $\mathbf{P} \in \mathbb{R}^{d \times d}$, $\mathbf{q} \in \mathbb{R}^d$ and $r \in \mathbb{R}$. This gives the increase of dimension for the matrix as stated above.

Here the variables to optimize are Φ_k and η_{ik}. The embedding is done in such fashion that one add a 1 to the end of the data so that it can be viewed as an intersection by an ellipsoid and the hyperplane. I.e $z_i \in \mathbb{R}^{d+1} = (x_i \in \mathbb{R}^d, 1)$. The soft margins of the optimization problem i.e. the embedding, is needed due to the fact that the collected data, as a difference to the test data, is not always well separated. This will also allow that if there are any outliers they will not harm as much as they would if there were fixed margins. But then again outliers and wrongly chosen points can still harm a lot and are preferably avoided.

Now solving this problem for a data set will give a matrix E describing the ellipsoid under the given constraints.

3.3.1 Is the point inside the ellipsoid?

To determine if the point is inside the ellipsoid, the point matrix is multiplied to the matrix describing the ellipsoid from both sides. This will give a vector with one value for each point, where the value determine if the point is inside the ellipsoid or not. If the value is less than 1 the point is inside that ellipsoid and otherwise the point is outside of the ellipsoid.

To calculate the distance the follow equation can be used, if E, the matrix describing the ellipsoid was found using the strict method without using the embedding.

$$\mathbf{x}^T \mathbf{E}_s \mathbf{x} = distance \tag{17}$$

Applying the embedding to the point so $\mathbf{z} \in \mathbb{R}^{n+1} = (\mathbf{x} \in \mathbb{R}^n, 1)$ and then the distance is calculated by

$$\mathbf{z}^T \mathbf{E}_e \mathbf{z} = distance \tag{18}$$

where **E** is the matrix describing the ellipsoid enclosing the cluster of the interesting points. Here E_s is without the embedding and E_e is describing the ellipsoid in the larger space. And **z** is either a vector describing an embedded point or a matrix built up by these vectors. (17) will give the distance from the ellipsoid to the point, where the value 1 will be on the ellipsoid and smaller values will be inside the ellipsoid. The value 0 will be in the center of the ellipsoids and any value greater than 1 will give a reference to how far away from the ellipsoid the point is. [ref[ref giovani]] This value can later be used to try to decide what cluster/ellipsoid an unidentified point is closest to.

The value can also be used to determine how far away from the ellipsoid the point is. This is good for deciding which cluster a point outside all ellipsods belongs to. If the value is to high for every ellipsoid, the point is considered to be an outlier and is not taken into account. This is a good way to deal with unpredicted appearances in the blood sample. This could still happen close to or in an ellipsoid but on the other hand this is not likely, or becomes less likely the more clusters and features there are.

To get the matrix describing the ellipsoid, the CVX plugin in Matlab is used to solve the SDP described in equation 15. This takes two classes and tries to get as good matching ellipsods around them as possible.

3.3.2 Multiply Classes

For deciding multiple clusters a somewhat simplified method is used. In this case, the method for two ellipsods is used. Meaning that the cluster of interest to the ellipsoids, is placed in one group and all the other clusters are placed in another. Then the algorithm for two ellipsoids is used and the process is repeated until there are ellipsoids for all clusters. The drawback is that there is a risk of two clusters, in the joint group, overlapping the interesting area. This could result in that the ellipsoids is less precise than if the multi ellipsoid method had been used. It was chose to do a simplified method when creating the ellipsoids for the classifier. Instead or creating an algorithm that takes all the classes separately and try to do all the ellipsoids at one time, this will be a somewhat more complex setup with the constraints. The choice was made to use the same optimization problem as for two classes. In difference from Fisher's Linear discriminant that can get problems if the sum of the opposing data is on different sides of the cluster that is to be separated from the rest. The Ellipsoid method handles this problem well, due to the fact that here inner points and outer points are chosen to lie on either side of the two ellipsoids for the separation. So to get this the exact same optimization problem that is equation 15. This time the c_i on row four represents all the points/classes not in the current class, then this is run for each class. Another problem with the ellipsoid method is that it takes a lot of computational power to evaluate. That is a problem since one of the goals is that all shall run smoothly on rather cheap equipment, so that it can be used throughout the development countries. Therefore it's a good thing if the reduction of the dimensions of the space can be large, i.e. the space has as few dimensions as possible to still do a good classification.

In choosing this approach the aim was to get the ellipsoid to have as few false positives inside their reach, this is for all the classes. The interesting part here is to decide if there is malaria or not. In this case if something get wrongly classifies as something other, as long as it does not involve malaria, it can be accepted in a larger extent, than if malaria get wrongly classified. So there was some tests with just using the malaria alone and not put it together with anything else when computing the ellipsods for the other classes. But to get the optimal result the multi SDP method should be used. This part is considered an area for future improvements.

4 Data analysis

4.1 Fisher's Linear Discriminant Classification

The method using Fisher's linear discriminant works, as can be seen in figure 9, quite well for dividing healthy and infected parts of the inside of RBC. The red part is 30 pixels represented by parasites and the black, in the histogram, is the 23775 healthy inner pixels of blood cells. One problem is that 30 malaria pixels are quite few in comparison. As can be seen in the figure there are some overlap between the two types of pixels. This part of the test only contains healthy and infected pixels and does not take into account that there are other types of pixels in a blood sample.

If the cell walls are included as seen in figure 10 it can be seen that there is a considerably larger overlap between the malaria parasites (red stars) and the cell walls (blue circles), than it was when only comparing the inside of the cells. Again this can depend on the small amount of pixels of malaria infected RBCs in the blood sample.

If no other method to separate the RBC is used, there is a need to test everything in a picture. The picture can contain a lot of other things, which could wrongly be classified as malaria. But for every new cluster that is added to the separation problem the dimension of the space, which is projected on, must be increased by one. The dimension is always one less than the number of clusters, one try to classify an unknown pixel to, and this makes the problem more difficult.

4.2 Classifying unknown points

When the projection is found there is the problem of how to classify an unknown point to a specific cluster.

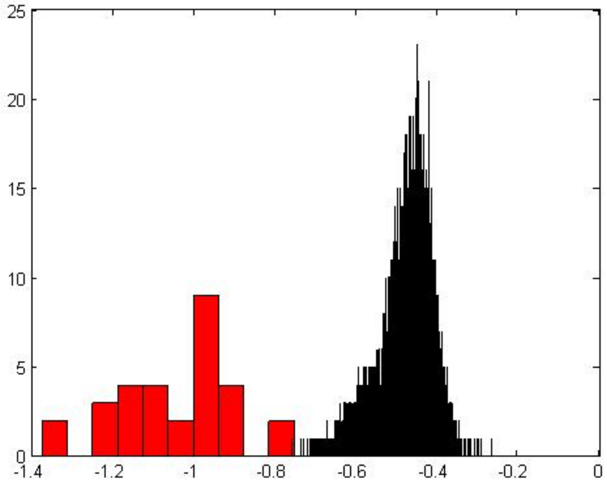


Figure 9: using Fisher's linear discriminant to separate only malaria parasites (red, left) and healthy RBC insides (black, right)

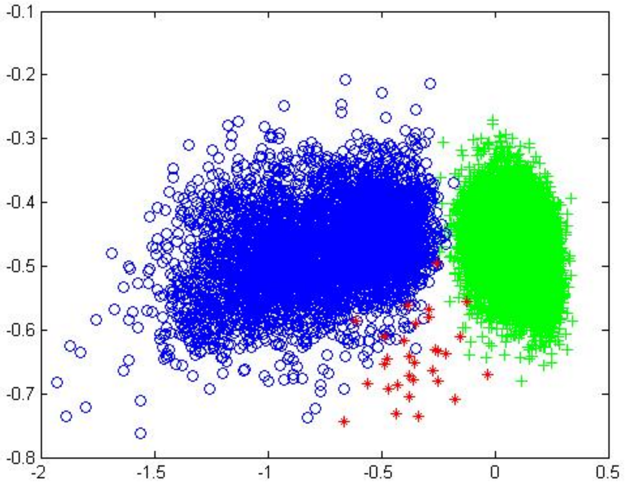
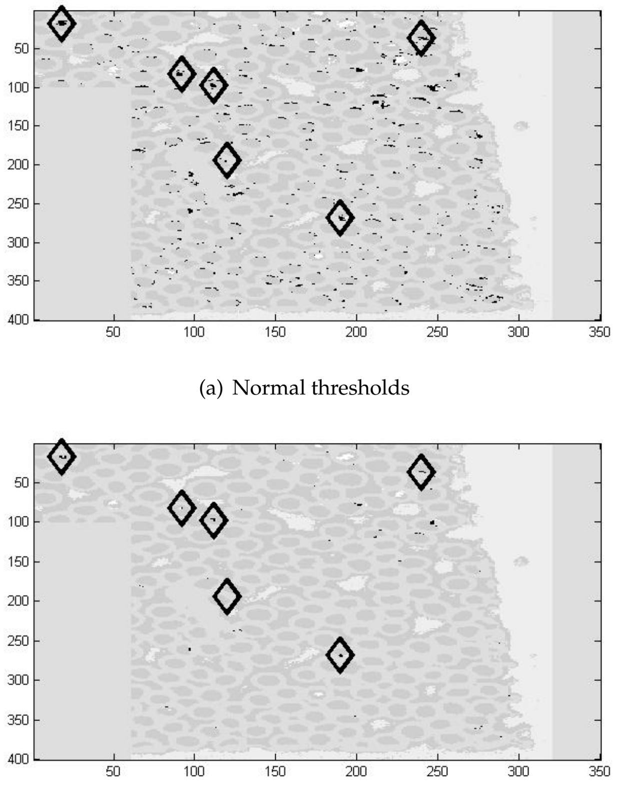


Figure 10: using Fisher's linear discriminant to separate malaria parasites (red $*$), healthy RBC insides (green $+$) and cell walls (blue o) The inner part of the cell separates well from the other two, but they are somewhat overlapping.

Here the Mahalanobis distance is used and as can be seen in figure 11(a). There are still many false positive results in the classification. In this particular example the clusters are: malaria - blue, cell walls - light green, unknown things in blood cells - red, background - orange, inside of healthy cell - light blue. The classification of background and inside of cells is quite much overlapping each other, as can be seen. This is due to that much of the background has been classified as inside, but for this classification this problem is of minor interest. The main problem here is that cell walls look quite like malaria parasites and some of the cell wall parts has been classified as malaria parasites. It can be seen that the cell walls tend to be larger on the lower part of the picture. This indicates that there can be some irregularity throughout the pictures and therefore it is a good idea to try to normalize even inside the same picture, to see if that makes any difference.



(a) Normal thresholds

(b) Thresholds: Malaria = 2, healthy = 1/2, wall = 2/3, background = 1/2

Figure 11: Fisher's linear discriminant has been used with Mahalanobis distance. This picture shows a separation attempt with 5 classes; Black is suspected malaria, gray in descending colour is "incell", cell walls, background and unidentified stuff in RBC. In the upper figure the normal threshold is used this means that the separation is done in the middle between each cluster. In the lower picture the cluster radius have been multiplied with a value to shrink or expand the cluster in comparison to the rest. In the upper picture there is a lot of false positive but all the malaria infected RBCs is found. In the lower picture however the infected cell close to the wall is missed. There are still some pixels that are false positives but they can later be sorted out when a certain amount of positive pixels is required.

For the ellipsoids the distance was quite easily calculated as above but the problem was to normalize the center for the healthy cells and thereby try to get a better measurement between samples.

Since the ellipsoid algorithm is fairly heavy in its computation it was impossible to use all the data points for the walls and inner points. One approach that was followed was to take the mean of a selection of points and then use 1000 new point after dividing all the starting points into a 1000 groups and thereby create the new points by their mean. This removes any outliers that would falsely lay in the clusters but it also makes the variance for the cluster point smaller.

The reason why the algorithm classifies a lot of points as malaria when the variance gets smaller is due to the fact that the point becomes more distant from the other clusters, because their ellipsoids are smaller.

This can be solved by dividing the distance with a constant to make it closer to the cluster center. In this case the constant that regulates the distance to the non-malaria cluster centres was chosen to 3 respective 4. This makes the algorithm more likely to classify an uncertain point between the classes as non malaria and it was seen that this reduced much of the wrongly classified points along the edges of the sample. But on the other hand it removes some of the pixels that is malaria as well. The figure 13 shows a zoom on one of the malaria infected cells and its chemically validated set. The ring of suspected pixels that occurs here is inside of the outer ellipsoid, with $\rho = 2$, for malaria, but when the threshold is changed most of it comes outside and is classified as "in-cell" of a healthy cell. So the approach with just changing the thresholds might not work as well for method with ellipsoidal constraints as it do for Fishers method. But if not the threshold is changed a lot of wrong classifications are made. Here one might suspect that because this "ring" is so close to the "in-cell" class and hard to get as malaria the pixels buildning it up is probably a mix of "clean" parasite pixels and "in-cell" pixels.

In figures 12(a), 12(b) and 12(c) the different results when dividing the distance with different constants can be seen. In the pictures the light blue indicates malaria parasites, the red shows the malaria from the training data, orange is where the training data is undiscovered. Light blue next to or in the vicinity of red or orange pixels is good, but the other ones are still false positives. In figure 12(a) the distance has not been changed so here the problem with the false classification can be seen clearly in figure 12(b).

A problem that occurred was that both Fisher's linear discriminant and ellipsoid methods were very sensitive to small shifts of cluster centres. When evaluating another picture there is a problem with the classification due to the fact that now the points are projected a bit off or outside of the ellipsoids. A first normalization where the picture is shifted, so that the cluster center for the background/inside of healthy cells is set to origin, before the projection is performed. This gives a better estimation for the other classes cluster center. Although if the center is too much shifted, compared to another, there is still a problem to detect malaria. Here it is of great importance that the mechanics when the photos are taken is as alike as possible.

4.3 Finding RBC

To find the RBC in a sample and thereafter finding the inside of them.

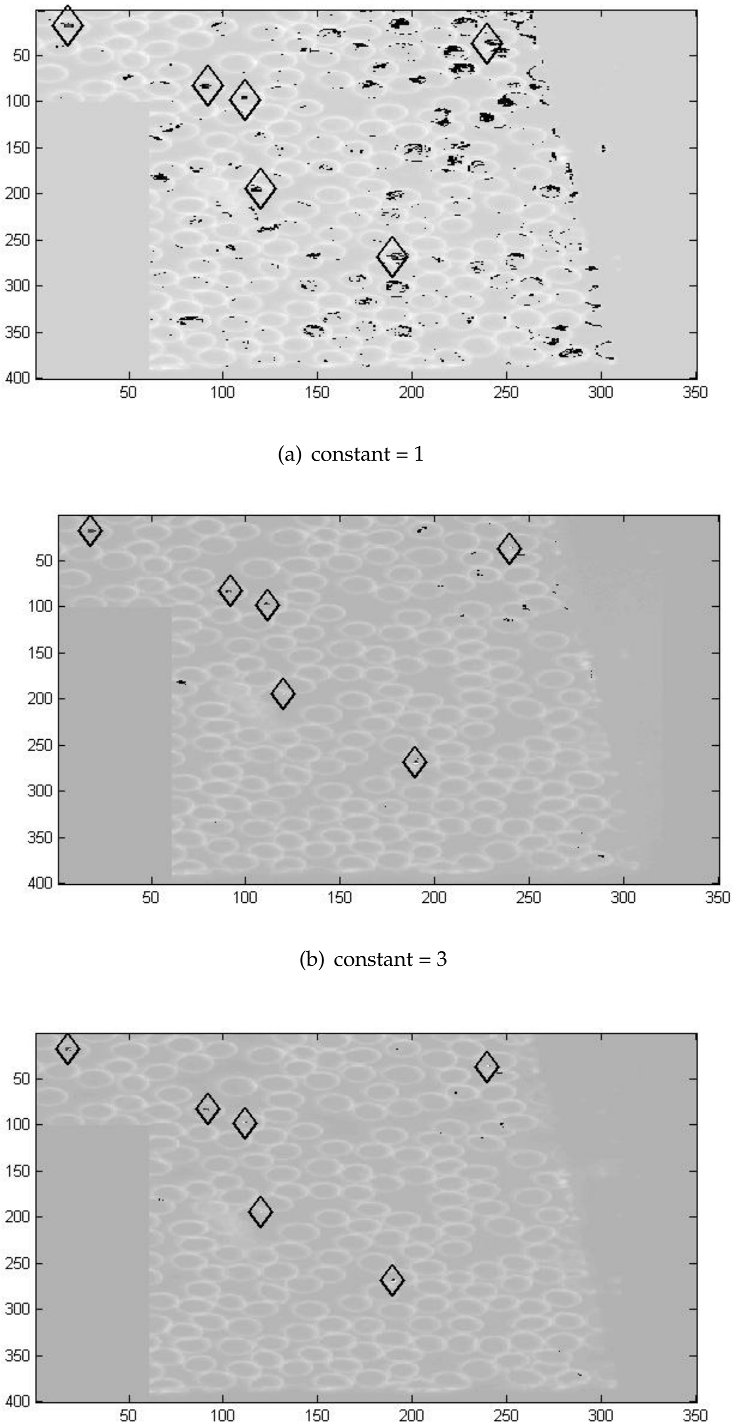
One method is to do a simple threshold separation by the walls and take all the areas that fulfil certain criteria. Criteria such as, shape, area and size. One problem is that the malaria parasite in such simple method looks a lot like the walls, so there is a risk that cells containing malaria is excluded if the criteria are too harshly chosen.

This is an research area that can be improved, to find a better way to find RBC centres/insides. If the centres are found in a satisfyingly good way, they can be used to form a circle which will be used as the test area. The problem with the data outside the blood cells will then be non-existing. A clear advantage with using this method is, that this method needs fewer clusters to focus on when trying to decide if there is malaria or not. On the other hand, if the method to find the RBC has such drawbacks/weaknesses, then it might not be preferable after all. So it all comes down to the question of what method is the most efficient.

4.4 Further Normalization

The first thing performed trying to get the pictures as similar as possible, is to take a white and dark reference for them to exclude as much as possible of the light and measurement errors that is due to flaws in the camera, light equipment and outside light.

Secondly trying to get the centres to coincide, is to move, for example the center for the healthy RBC inner part, to the origin. This is done by taking the mean of that center from all the data before trying to decide which cluster it belongs to. In that way one normalises each image such that the healthy RBC

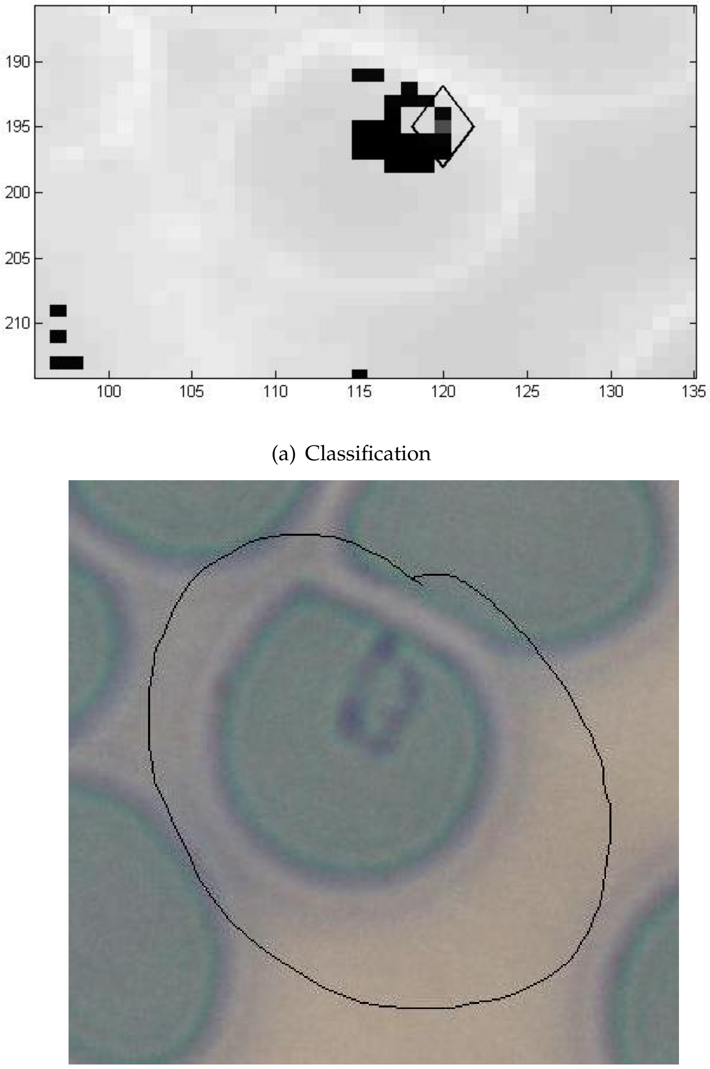


(a) constant = 1

(b) constant = 3

(c) constant = 4

Figure 12: Trying to classify points in blood sample using ellipsoid method. To reduce the number of points, the points represent the inner cells and are the mean of more points. This reduces the size of the ellipsoid and therefore change how the distance is calculated. To compensate for this the distance, given by the function with the "meaned" points, is divided by a constant. In figure 12(a) the constant is 1, in figure 12(b) the constant is 3, And in figure 12(c) the constant is 4.



(a) Classification

(b) Validation

Figure 13: This is zoom in figure 12(a) and the validation set, Here the classifier have found the ring formation of the parasite and classifyed is as malaria. Figure 12(a) have a lot of false positeves some of these can be seen in the neighborhood around this blood cell. The gray pixel in the black rhombus is part of the training set used to train the classifier, but all the rest have been classified as malaria. But when

match perfectly. This will hopefully get the other clusters centres closer to one another when trying to detect the malaria infected RBC.

Also this approach can be used on a single image if it is suspected that it does not behave the same at different spots of the image. Then the picture can be separated into different grids where the data is normalized in each grid.

But to be able to do this normalization one have to be rather certain which cluster that actually is the healthy RBC inner part. Here there are some different approaches to do this. Firstly it can be assumed that this center should be fairly close and therefore first remove the training data center for the healthy RBC and thereafter take the one that is closest to origin and remove that as well. Secondly one can make the assumption that there are probably a lot more healthy RBC and therefore if one is able to determine data point only in the RBC, one can assume that the cluster with most points is the one of interest. To actually find the cluster one might need to use some clustering method, for example K-means. K-means does for a given number of clusters choose the center at random, then for some distant function add the points to the closest center. After that it recalculates the center and redoes the process until no points change cluster.

5 Results

One theory by A. Merdasa [4] was that the centres of the infected RBC behave differently than the centres of the healthy ones. One approach, under the circumstance that the centres could be found in an efficient way, is to take the 9 pixels containing the centres and then feed them into the algorithms. In this sample data from A. Merdasa there are 452 RBC. 20 of them are defined as healthy, 20 is sure to be infected and the rest is unknown. Using these centres gives an initial good definition, if looking at them after separation, where all the cells has been used as the classifier, the algorithm using the ellipsoids can do a good distinction. In figure 14 the separation of the inner parts of infected and healthy RBCs seems promising. Classify all the centres in the whole picture, using only the parts inside of the RBCs, the result seen in figure 15(a) seems promising as well. But since when this classifier used all the known data, to be as good as possible, there are none of the validated blood cells left to check if the classifier is correct. Therefore a new classifier will be constructed, using a random selection of healthy and infected RBC centres. A drawback with this classifier is that if the whole data is used as input the result will be chaotic. Due to the fact that the classifier only takes into account the inner parts of the cells. The outer parts and cell walls will cause some troubles and cannot be correctly classified and therefore it is hard to get some relevant information from this, as can be seen in figure 16.

Choosing to use 10 random of each of the infected and healthy cells respectively to create the ellipsoids and the left over cells are used as to validation, the result is surprisingly good. In figure 18 only seven malaria pixels out of 90 were unclassified, with the threshold of 1 and all the healthy ones were classified correctly. Due to the random selection some other runs sometimes gave less and sometimes more non classified infected pixels, where 17 unclassified was the highest number. Still they were inside the outer ellipsoid so there was no risk for false classification. This is a good result but the drawback is that this is under the assumption that there exist a stable and good way to find the center of the RBCs.

While picking the centres by hand and compare them is easy, in comparison to doing it automatic, there is still the problem to see if the comparison between pictures and with less images/information works. One approach is to find a good way to pick out the centres but a more appealing way of doing this is to be able to run the entire data through the process and classify the malaria without the need to do any prior selection of what data to test. This will make the method more general and able to apply to a greater variety of problems.

Using the same classification but reducing down to only the 6 dimension instead of 39. Using the 6 had the greatest contrast according to A Merdasa [4]. These 6 are: the scattering 435nm and 625nm, reflection 435nm and 700nm and transmitting 435nm and 590nm. In the approach of looking only at non malaria and malaria centres, with reduced dimensions, this gave somewhat lesser results. This can be seen in figure 19 where it can be seen that some of the healthy points are inside of the outer circle of the infected malaria cells. But since they are all inside of the inner ellipsoid of the healthy cluster it would be classified correctly with higher threshold.

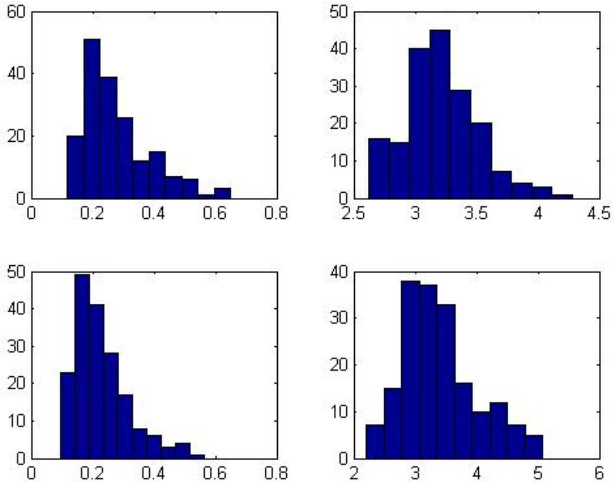
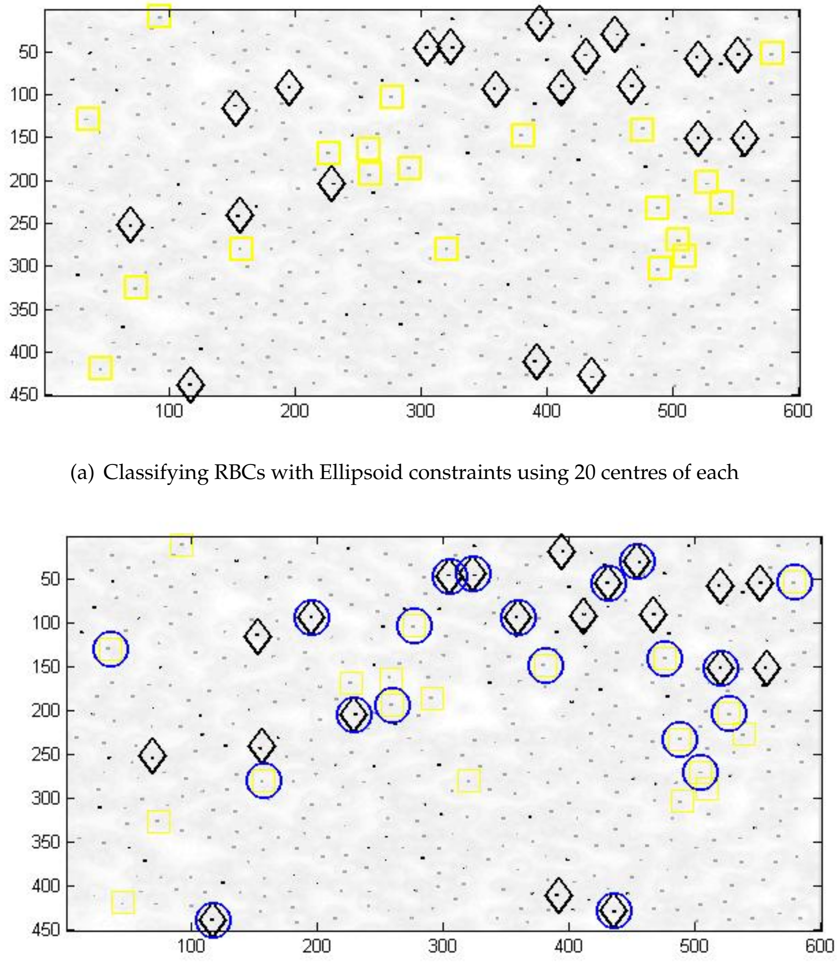


Figure 14: Using ellipsoid on the centres of the RBC. Here the upper left histogram shows the distance to the ellipsoid center for infected blood cells to the infected center. The upper right is the distance from a healthy RBC to the infected center. The lower left picture is from healthy cells to healthy centres. The lower right is the distance that infected cells have to the healthy center. It can be seen that the centres are well separated. Here the rho of two was used while creating the ellipsoids.

Trying this approach using Fisher's method gives good results here as well. While using the same 6 dimensions as above the separation of the healthy and the infected RBCs can be seen in figure 20 with clearly separated classes.

Classifying with four dimensions to see if the result gets better when trying to classify a whole picture. The added dimensions are the class containing the walls of the RBCs and the class describing the background. The point is now to reduce the leakage that was observed in picture 16. This is done by selecting points from the cell walls and from the background as this are the other big parts that a blood sample contains of. This is done by selection 20 points of the cell walls and 20 points from the background. These points are treated like the infected and health inner parts, and the 9 pixels around those points is chosen to be included in the respective class. More points can be chosen but here the choice was to have the same amount of all classes. The addition of more points gives a better classification but as the validation is done for the same picture to many points may hurt the usability of the result. The initial result looks rather good as can be seen in figure 17. There is still some small chunks of black outside the RBCs but most of them are small. The big blocks of black outside of the known infected cells are all inside of cells that could very well be infected cells. For this classification the ellipsoids for all but the malaria classes have the radius increased by a factor of three. This is done as to push the threshold for the classification closer to the malaria cluster so that most outlier should be classified as not to be malaria. If looking at the classifier in figure REF it is possible to observe a clear separation between the clusters belonging to the different clusters. One approach would be to All the point $\hat{\mathbf{x}}_{(pa)}$ classified as malaria is stored away for further analysis.

When the reduction of the dimension is done as an addition to the 4 classes classifier of the whole picture the choice of the wavelength that had the greatest contrast do not work as good as had been hoped: this is not that surprisingly thou because these dimension was chosen as to separate the healthy from the infected cell centres. As can be seen in figure 21 here the lower plot shows the classifier for the 6 dimensions where the third plot describes the walls. This is separated but not as good as for the 39 dimensions. These results in a risk of wrongly classify walls. This can be seen in figure 22 here the wrong classified walls are clearly seen as black segments along the border of the RBCs. The conclusion is that either the use of more wall points is used in the training or some other dimensions are chosen that are better at classifying walls, that such dimensions exists is obvious due to the fact that that the separation



(a) Classifying RBCs with Ellipsoid constraints using 20 centres of each

(b) Classifying RBC with Ellipsoid constraints using a selection of 10 random centres of each

Figure 15: Here the use of ellipsoids to try to separate the center parts of the RBC have been used. $\rho = 2$ was used in the creation of the ellipsoids. The black parts indicate an infected cell and the grey parts indicates healthy ones. The black robs show which centres were used in the training and the yellow squares show the healthy cells that was used in the training. But in the upper picture all the known data was used to do the classifier so this does not say much. The lower picture however the classifier is created using 10 randomly chosen centres from each of the healthy and the infected cell. Here the algorithm performs nearly as well

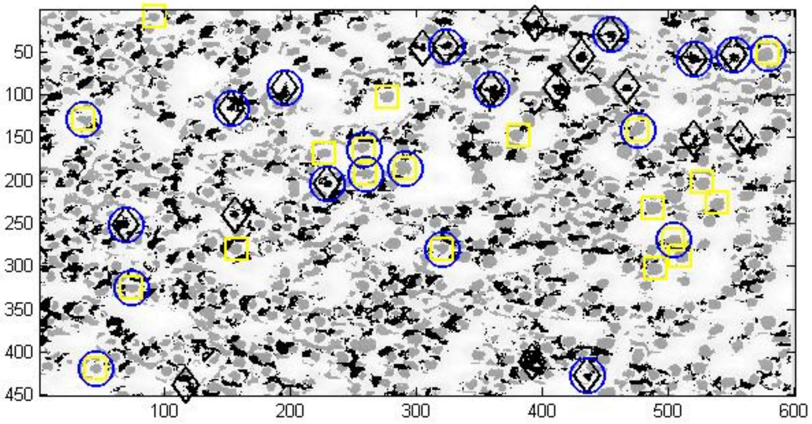


Figure 16: Here the use of ellipsoids to try to separate the center parts of the RBC have been used. $\rho = 2$ was used in the creation of the ellipsoids. Here red parts indicate a healthy center. The yellow '+' show which centres were used in the training and the green '*' shows the infected cells that was used in the training and light blue parts has been classified as Malaria. The difference from figure 15(a) is that here the whole data set has been used while classifying and as can be seen, the picture is a mess and it is hard to decide if indicators of malaria actually are that.

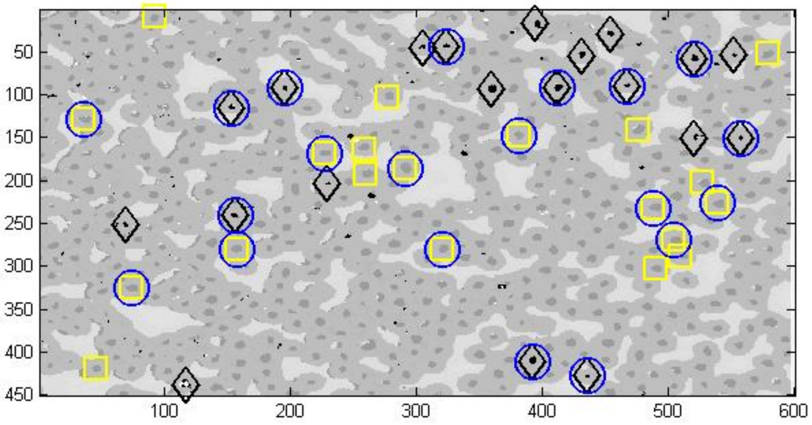


Figure 17: Here the use of ellipsoids to try to separate the malaria infected RBC from the healthy ones. $\rho = 2$ was used in the creation of the ellipsoids. The black parts indicates an infected cell, the grey parts indicates healthy ones cell walls and back ground in descending color in the gray scale. The black robs show malaria infected validated cells. The Blue circles show which centres was used in the training and the yellow squares shows validated healthy cells. Here the cell walls and the background were added as two classes. The classification works much better with this then without. And there are much less false positives outside of the cells.

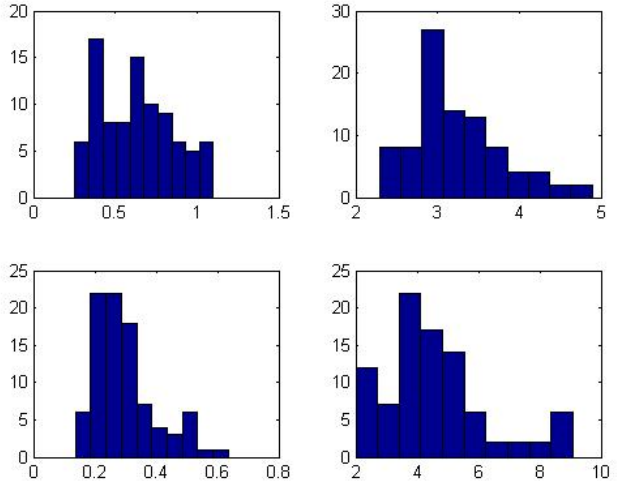


Figure 18: Using ellipsoid on the centres of the RBC. Here the upper left histogram shows the distance to the ellipsoid center for infected blood cells to the infected center. The upper right is the distance from a healthy RBC to the infected center. The lower left picture is from healthy to healthy. The lower right is the distance that infected cells have to the healthy center. As can be seen the centres are well separated. Here the ρ of two was use while creating the ellipsoids. Using the threshold of one to classify the points to belong to either ellipsoid, the result with this random selection was surprisingly good. Only seven of the 90 malaria pixel was outside of the threshold but still within the radius of the outer ellipsoid. Meaning that at first is not classified as either malaria or not, but with lesser threshold it would have been correctly classified. All the healthy centres was correctly chosen and neither healthy or infected centres was wrongly classified

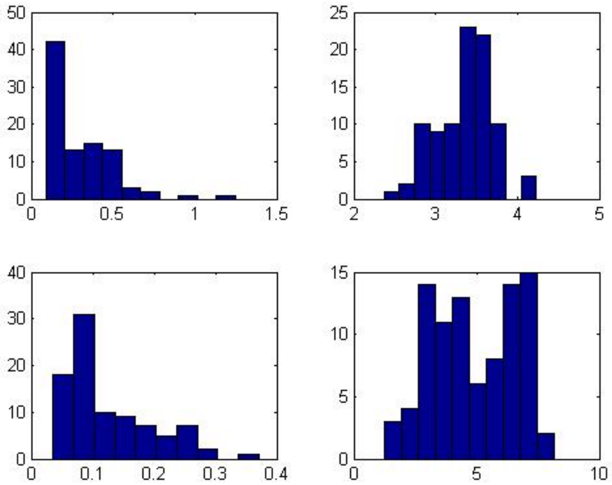


Figure 19: Using ellipsoid on the centres of the RBC. Here the upper left histogram shows the distance to the ellipsoid center for infected blood cells to the infected center. The upper right is the distance from a healthy RBC to the infected center. The lower left picture is from healthy to healthy. The lower right is the distance that infected cells have to the healthy center. As can be seen the centres are well separated. Here the ρ of two was use while creating the ellipsoids. Using the threshold of one to classify the points to belong to either ellipsoid, the result with this random selection of 10 centres for creation and 10 for validation of malaria and healthy RBCs respective. Here only 6 of the dimensions are used, and it can be seen that 3 of the points that are supposed to be healthy are under the threshold of 2 but still over 1 for being classified as Malaria. But since all healthy still are under the threshold, to be classified correctly, this would not have given a false positive alarm anyway.

with 39 dimensions succeeded well as can be seen in figure 17.

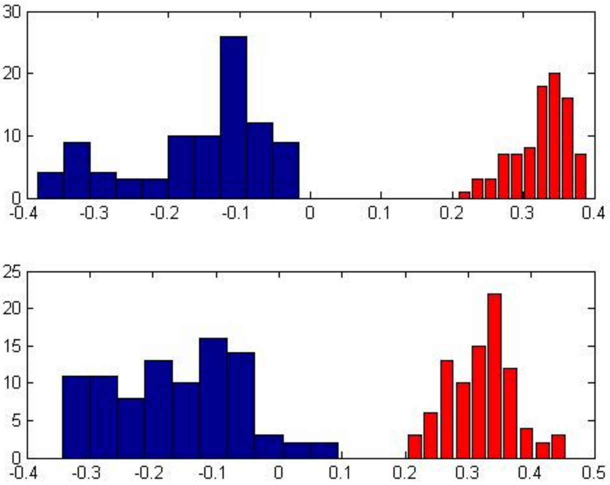


Figure 20: Here the Fisher's linear discriminant has been used to in the upper figure create a classification projection using 10 random infected and healthy inner parts of RBCs. The lower figure describes the result after the projection is applied to a test set, containing of 10 healthy and infected centres. The inner part of each RBC contains of the 9 pixels closest to the calculated center.

When all the points are classified is should be decided if the potential malaria infected points $\hat{\mathbf{x}}_{(pa)}$ in the data probably is malaria. This can be done by looking in the neighbourhood of each point and see if there are many more point there classified as malaria if so this can be considered an infected RBC. If however there is only a few or no points in the vicinity of this point it can be reclassified as probable false alarm. This number can be chosen depending on how secure one must be not to imply malaria on an uninfected person. Then again if the threshold is high and there still are at least a few point classified as malaria then the sample can be deemed to be infected with really high probability. However if the sample do not indicate any malaria at all at a high threshold then the threshold is lowered to see if there is any hits now. If not then the sample is probably not infected. Then we have the region where we have some middle values. This separation is good due most of the time a fast conclusion can be drawn. Otherwise some more investigation can be used and if after that there still is confusion whether or not this is malaria then some other method may have to be used. Then again the middle section can be skipped by simply set the a single threshold at a fix point. For this a value of ten is a good benchmark.

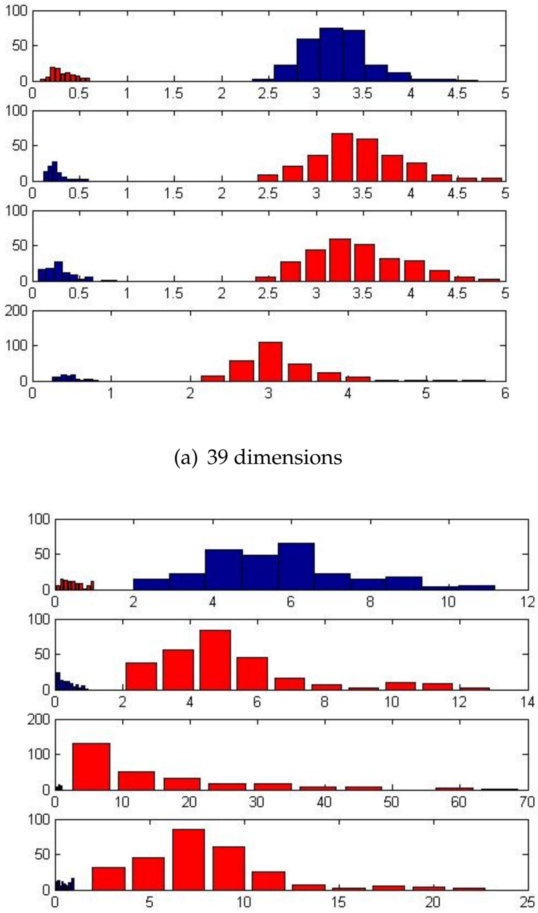
To get a measurement on the performances of the classifications one can use ROC-curves.

The ROC (Receiver Operating Characteristic) curve is constructed by plotting the true positive rate versus the false positive rate. That means in this case; the amount of found malaria infected cells divided by the total amount of infected cells. This value is then plotted against the number of found uninfected blood cells divided by the total amount of uninfected blood cells. Given the use of different thresholds.

This method gives a measurement on where to choose the threshold for best results, also taking the risk of false positive results into account. It is also a good way to compare different methods for classifying malaria with each other.

In figure 23 a ROC-curve gained when using the Fisher's linear discriminant method is displayed. It quite rapidly finds the way to include most of the malaria infected red blood cells. Although since the uninfected RBC are so many more then the infected ones, the cost for falsely classify only a small fraction of the healthy cells could deem the whole result incorrectly as malaria. For example if 3 RBC of a 1000 are infected by malaria then the patient would be classified as infected. Then one can not have that 1 in 1000 of the healthy RBC is wrongly classified as infected, because that would be a too great risk of wrong classification.

Two ROC-curves describing the result from using the ellipsoid approach can be seen in figure 24 and



(a) 39 dimensions

(b) 6 dimensions (contrast)

Figure 21: This shows the distance between the classifiers for the 39 dimensions, upper picture and the 6 dimension classifier using the 6 wavelengths with greatest contrast between healthy and infected RBC centres, lower picture. The left part of the picture shows the class that is to be separated and the right the sum of the separated parts. In the lower picture it can clearly be seen that the separations are not as good as in the upper picture. This is due to the fact that de dimension is reduced but mostly due to the fact that de dimensions used was chosen to separate only two of the classes well.

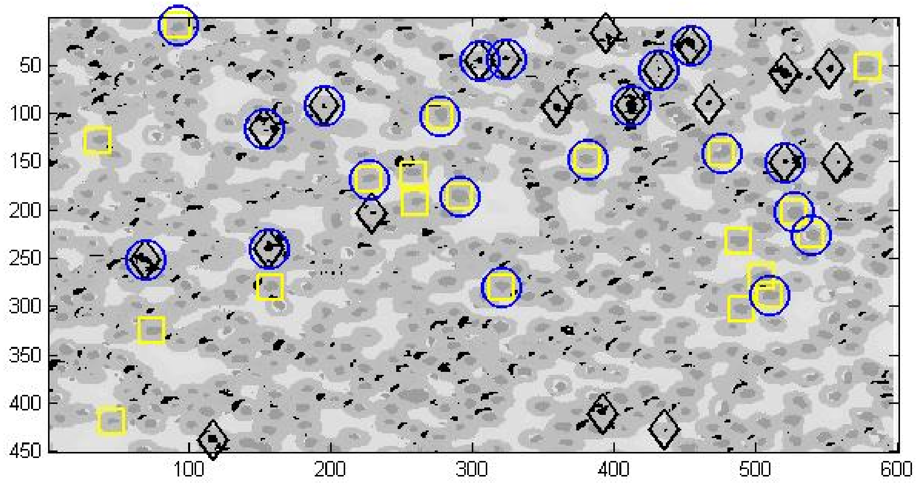


Figure 22: Here the use of ellipsoids to try to separate the malaria infected RBC from the healthy ones. $\rho = 2$ was used in the creation of the ellipsoids. The black parts indicates an infected cell, the grey parts indicates healthy ones cell walls and back ground in descending color in the gray scale. The black rhombs show the malaria infected known cells and the blue circles which centres was used in the training and the yellow squares shows the healthy cells that was used in the training. When the dimension is reduced as before using only the dimension with the greatest contrast then some false positive classification occurs this is due to the fact that even when using this reduction worked fine for only separating the two classes, which was why just them was chosen. Information loss regarding the other parts of the blood sample was overlooked and now shows in the form of false positives

figure 25. As one can see the curves differ somewhat but in general they look about the same.

The curves from the ellipsoid method are somewhat better just in the beginning and it is here the threshold would have to be. Because of the ratio between the infected and non infected RBC, this is the part of the ROC-curves that are interesting.

6 Discussion

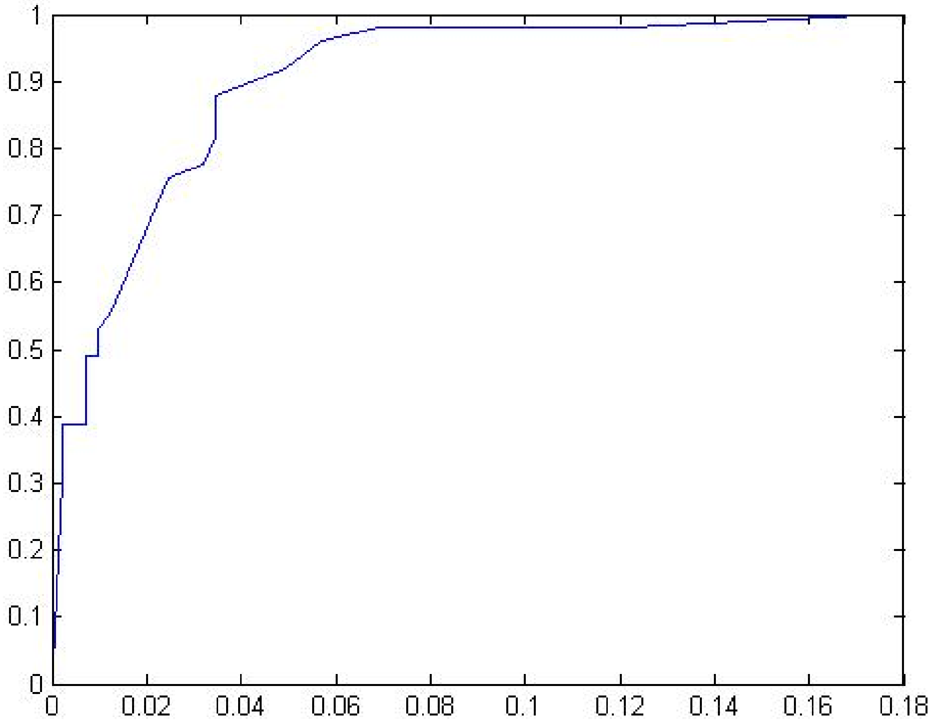
The method using Fisher's linear discriminant has a tendency to wrongly classifying points as in a higher rate than the ellipsoid method do. But on the other hand the Ellipsoid method miss and do not classify malaria points, in an higher extent then what Fishers do. This can have something to do with the thresholds chosen. But this in is not the only cause for the miss classifications. Fisher's discriminant worked best on the old data. Where it found most of the malaria infected RBC but there are still a lot of pixels wrongly classified so it do not work good.

The fresh data is much simpler to separate than the old data, all the methods given. This is mostly due to the fact that the scattering geometry performs really good at the center of the cells in the fresh samples, but the old and dried samples have lost most of this information. It is still possible to get a bit of information out of the old samples but it gives false positive results at a higher rate.

On further basis more data is needed to be able to get a good measurement of how good the approaches are while trying to decide what approach works the best.

In the future the focus on further testing should be on fresh data due to the fact that there is the data that can do any difference. The old data was much harder to get a classification out of and there is not much of a point to try to optimize that.

The implementation/methods used with the Ellipsoid constraints and the Fisher's approach are very general so therefore they should be able to separate other things then just malaria. If only a few dimensions are used the ellipsoid approach is better because it is somewhat computational heavy. There is also



Fi.jpg

Figure 23: This figure are describes the ROC-curve while using the Fisher's linear discriminant based algorithm to classify the blood sample. To note, since there is many more healthy than infected RBC, the interesting parts of the ROC-curve is in the left most part.

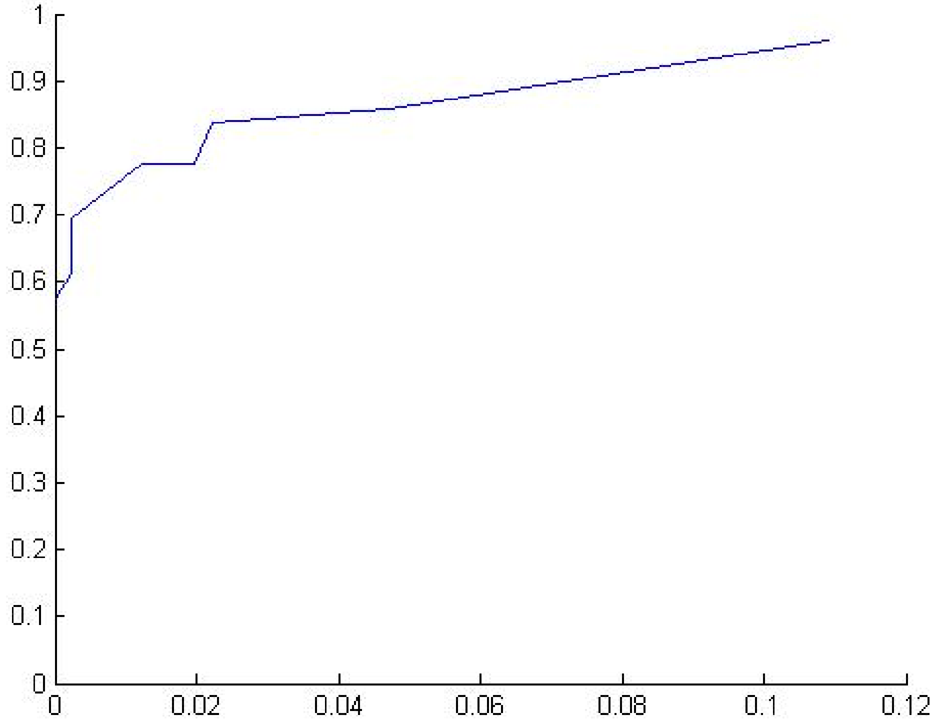


Figure 24: One ROC-curve given when the Ellipsoidal based algorithm was used. To note, since there is many more healthy than infected RBC, the interesting parts of the ROC-curve is in the left most part.

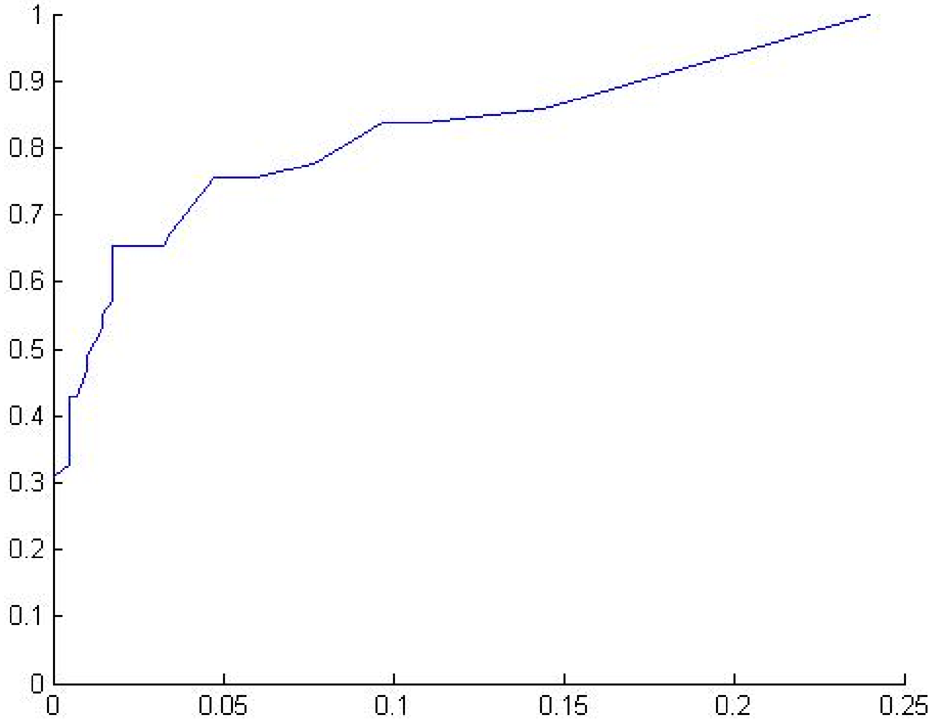


Figure 25: Another ROC-curve given when the Ellipsoidal based algorithm was used. To note, since there is many more healthy than infected RBC, the interesting parts of the ROC-curve is in the left most part.

the possibility of first using Fisher's linear discriminant then use the ellipsoid on that result instead of using the mahalanobis distance to do the classification.

When trying to classify the malaria parasites, they are rather small in the sense of pixels, this made the choice of the pixels hard when trying to get the correct pixels containing only the parasite. There is a risk that all pixels chosen could have been only partly a parasite. This could have made the training data somewhat inaccurate, and may be one explanation to why pixels on the border between cell insides and the cell wall were wrongly classified as malaria.

One big problem is that all the blood sample that have been used during this thesis have been from infected samples. On one side the comparison between malaria and not malaria have been on a blood cell level but there are still some issues that the uninfected blood cells come from a malaria infected blood sample. Also one cannot conclude from the testing here if there are any big differences between different persons. So there is no notion of that for example the blood group makes a different. So here is a big area that could be further explored.

If one like to continue this approach; I would recommend getting blood samples from a few different persons both infected and not. Then take photos from at least two different spots of each sample, but collect as much as one likes because after collection the samples should go through the process of regular malaria testing. This is done to get a reliable validation set. The problem is that after this process the sample cannot be used for any more data collecting with the LED-Microscope because it has been stained with chemicals.

I realized a bit late that the data I thought was sufficient probably was not. To get better results more data is required. Much of the data I have has not been validated, or more exactly the data and the validated data do not overlap as much as I would have liked. Making the choice of validated malaria infected cells in the old samples sparse. Even for the validated RBCs there was a problem with the choosing of what pixels to use for the training set, with as few as there were some of the pixels chosen was probably a mix of malaria parasite and cell. But in this case with the old data this is of more of method comparison and not so much for usability since the blood sample was too old. No values from these actual samples would be usable when trying to classify a new sample.

Problem with cluster centres being different in different pictures. Made the comparison between pictures harder, this will probably not be that a significant issue when the measurement standard is in place.

7 Conclusions

The classification of malaria works much better on fresh samples, though it's hard it is still possible to get separations from old samples. Due to the fact that for old samples there is no rush to do a classification, further research on this field do not need to be done, if the goal is to get a method that should give result shortly after the blood sample is taken. The focus should lie on trying to improve the method on fresh samples, which have been taken with the routine now in place for taking photos of malaria. This can be done by running the same process but on this data to calibrate the values. Test should be done with data from different microscopes to see how this effects the classification. Preferably it should be field tested in the different countries that have the microscope equipment.

Before any classifications in Africa can be made, a standard method for taking the data have to be in place. Then a collection of new fresh data can be done to get new training data for the algorithms. Collecting data from different location is needed if not a standard is in place. Even if the methods are standardized it would be better if there was an easy way to select the training data, so this could be done in field to update the classifier when more data was collected.

The Classification with the ellipsoids works better on the fresh sample and that method could be developed by looking into multi ellipsoidal classification, instead of using the current method.

References

[1] WHO, "Malaria," http://www.who.int/tdr/diseases-topics/malaria/en/.

[2] WHO Press, "Basic Malaria Microscopy Part I.," WHO Press, Geneva Switseland, 2010.

[3] M. Brydegaard, A. Merdasa, H. Jayaweera, J. Ålebring, and S. Svanberg, "Versatile multispectral misroscope based on light emitting diodes," , no. 2, October 2011.

[4] Aboma. Merdasa, "Multispectral Microscopy with Application to Malaria Detection," , no. 1, February 2010.

[5] M. Brydegaard, Z. Guan, and S. Svanberg, "Multispectral Broad-band multispectral microscope for imaging transmission spectroscopy employing an array of light-emitting diodes," *Am. J. Phys.*, vol. 77, no. 2, February 2009.

[6] A. Runemark, M. Wellenreuther, H. H. E. Jayaweera, M. Brydegaard, and S. Svanberg, "Rare Events in Remote Dark Field Spectroscopy: An Ecological Case study of Insects," 2011.

[7] Jens. Ålebring, "Multispectral LED based microscopy," , no. 1, September 2012.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[9] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*, Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[10] M. Hazewinkel, *"Mahalanobis distance", Encyclopedia of Mathematics*, Springer, New York, 2001 edition, 2001.

[11] Giovanni. Soldi, "Wireless Positioning using Ellipsoidal Constraints," , no. 1, June 2010.

[12] L. Xiao and L. Deng, "A Geometric Perspective of Large-Margin Training of Gaussian Models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 118 –123, Nov. 2010.