Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2019;5(1): 1-11

Original Article

Assessing Malaria using Neutral Zone Classifiers with Mixture Discriminant Analysis on 2D Images of Red Blood Cells

Shariq Mohammed* and Dipak K. Dey

Department of Statistics,College of Liberal Arts and Sciences, University of Connecticut, Storrs, Connecticut, USA.

ARTICLE INFO

ABSTRACT

Received 20.01.2018
Revised 15.04.2018
Accepted 06.10.2018
Published 01.02.2019

Key words:
Cell segmentation;
Conditional misclassification rates;
Feature extraction;
Receiver operating characteristic;
Sensitivity and specificity

Background and Aim: We aim to build a classifier to distinguish between malaria-infected red blood cells (RBCs) and healthy cells using the two-dimensional (2D) microscopic images of RBCs. We demonstrate the process of cell segmentation and feature extraction from the 2D images.
Methods and Materials: We describe an approach to address the problem using mixture discriminant analysis (MDA) on the 2D image profiles of the RBCs. The extracted features are used with Gaussian MDA to distinguish between healthy and malaria infected cells. We also use the neutral zone classifiers where ambiguous cases are identified separately by the classifier.
Results: We compare the classification results from the regular classifiers such as linear discriminant analysis (LDA) or MDA and the methods where neutral zone classifiers are used. We see that including the neutral zone improves the classification results by controlling the false positive and false negatives. The number of misclassifications are seen to be lower than the case without neutral zone classifiers.
Conclusion: This paper presents an alternative approach for classification by incorporating neutral zone classifier approach, where a prediction is not made for the ambiguous cases. From the data analysis we see that this approach based on neutral zone classifiers presents a useful alternative in classification problems for various applications.

Introduction

Malaria is one of the widespread diseases in many developing countries. With 3.3 billion people in 97 countries at risk and with an estimated 200 million cases and around 600,000 deaths, malaria remains a disease of global health importance [1]. The high costs associated with both the equipment required for laboratory procedures and manpower, makes it imperative to have fast and efficient ways for detection. Statistical techniques have commonly been employed for automatic image classification of blood cells to detect and distinguish between disease or infection category and control groups. In this paper, we discuss the performance of existing statistical parametric classifiers and demonstrate how the performance can be improved by considering extensions of such classifiers based on indecision strategies. Such algorithms can be used to improve the efficiency of the diagnosis procedure of malaria through the

* Corresponding Author: 215 Glenbrook Rd U-4120, Storrs, CT 06269, USA.

Please cite this article in press as: Mohammed S, Dey D.K. Assessing Malaria using Neutral Zone Classifiers with Mixture Discriminant Analysis on 2D Images of Red Blood Cells. J Biostat Epidemiol. *2019; 5(1): 1-11*

classification of two-dimensional (2D) image profiles of red blood cells (RBCs).

Pattern recognition and classification can be done using either a supervised or an unsupervised classification approach. In supervised classification (discriminant analysis), patterns in the input are identified as members of predefined classes, while unsupervised classification (clustering) methods partition a set of observations into subsets, so that observations in the same subset are similar in some metric and the subsets are separated from each other. Clustering and classification techniques can be broadly classified into the following categories as deterministic clustering (hard clustering) and model-based clustering (soft clustering). Deterministic clustering is computationally cheap with a straightforward interpretation, but it produces no probability statement (only produces a binary statement). Model-based clustering produces precise cluster assignment in terms of a probability statement; however, it is computationally more expensive.

Classification problem focuses on building a rule to assign class membership of an item based on the p-dimensional vector of predictors or features for n observations. Linear discriminant analysis (LDA), multiple logistic regression, nearest neighbor methods and classification trees are a few of many traditional statistical approaches of handling the classification problem. Most recent advances include methods with neural network classifiers, which can incorporate non-linear modeling assumptions and have the capability of handling large number of predictors. In this paper, we focus on mixture discriminant analysis (MDA), as proposed by [2] is an extension of Fisher's LDA. The classes are modeled as mixtures of Gaussian distributions in MDA, whereas LDA models each class as a single Gaussian distribution. Similar to LDA, MDA also does optimal subspace identification with additional functionality. MDA been used in a wide variety of applications. Gaussian mixture discriminant analysis has been used for single-cell differentiation of bacteria [3], process monitoring [4], digit recognition and connected digit recognition [5], improving tissue classification in Magnetic Resonance Imaging (MRI) [6].

Automatic classification of red blood cells was studied much earlier by [7] where the cells are classified based on their morphological characteristics and hemoglobin content. Significant attempts have been made to count malaria infected cells to demonstrate the presence of parasites in the blood. Most recently [8] used neural networks for classification with image moments as predictors. We propose the use of MDA with morphological characteristics of the cells as the predictors for our modeling. However, as with any statistical classifier, the predictions are prone to higher number of false positives and/or false negatives. We focus on addressing this problem by incorporating neutral zone classifier [9,10] strategies to decrease the conditional misclassification rate. Neutral zone classifiers do not assign predictions to ambiguous cases. Consequently, such cases can be referred for further scrutiny to make a more informed prediction.

In the next subsection, we describe the process of segmentation of the red blood cell from their image profiles. We describe mixture discriminant analysis, a model-based clustering algorithm in the methods section which shall be used to demonstrate classification performance of red blood cell and further be used to improve the prediction. The methods section also discusses neutral zone classifiers and how we plan to use it to improve the prediction of MDA. The results and conclusions of this analysis are discussed in the latter part of the manuscript.

Cell Segmentation and Feature Extraction: In a 2D image of red blood cells in the blood sample, we need to be able to segment out the pixels which correspond to only the red blood cell and not the plasma around it. Since an object can be easily detected in an image if it has enough contrast from the background, we use edge detection along with basic morphology for detecting the cell. The step by step procedure (as shown in Figure 1) is carried out for each image to do the segmentation (cell detection).

```
Algorithm: Cell segmentation
Step 1: Read image into MATLAB
Step 2: Detect entire cell
Step 3: Dilate the image
Step 4: Fill interior gaps
Step 5: Remove connected objects on border
Step 6: Smoothen the object
```

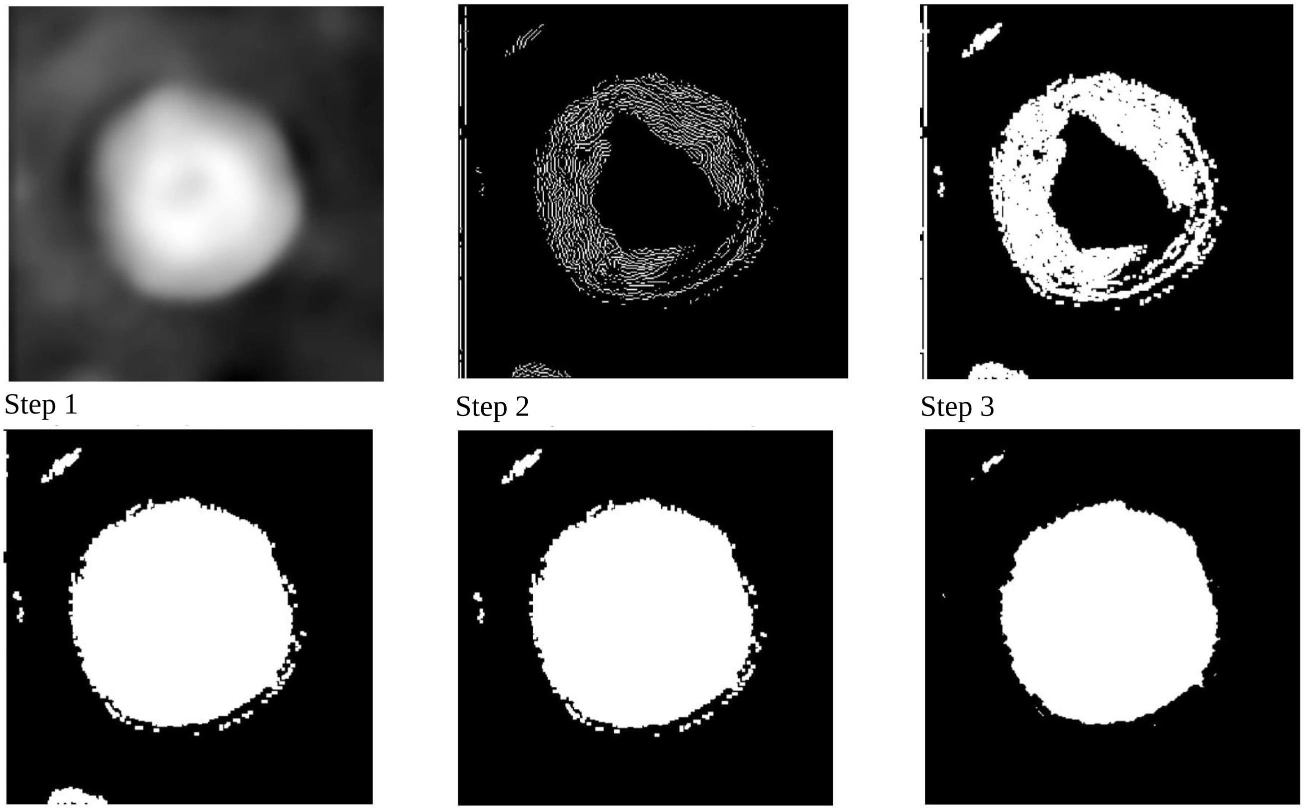
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6

***Figure 1**. Image segmentation of a healthy red blood cell.*

Step 2 (Figure 1(b)) aims to detect the pixels containing the cell in the whole image and utilizes the fact that the cell would differ in contrast from the background of the image. Once we obtain the lines of high contrast in step 2, we dilate the binary gradient mask using linear structuring elements in step 3 as shown in Figure 1(c). Step 4 (Figure 1(d)) fills the gaps in the interior of the cells using the dilated gradient mask. Step 5 (Figure 1(e)) removes any connected objects in the blood within the image or other cells in the background, as we are only concerned about one cell in the image. To make the segmented objected look more natural, we smooth the cell by eroding the image in step 6 (Figure 1(f)). The final segmented image looks as shown in Figure 2. To remove any residual objects, we calculate the centroid and the major axis length. We consider centroid of the image as the center of mass of the region obtained after step 6 and major axis length is considered as the length of the major axis of the ellipse that has the same normalized second central moments as the region. The Image Processing Toolbox from MATLAB is used to detect the cell in the image. Due to the irregular shapes of some of the RBCs, step 3 might not be very efficient. Different filling criterion could be employed in such cases. More details are given in the MATLAB documentation on cell segmentation and can be accessed at https://www.mathworks.com/help/images/detecting-a-cell-using-image-segmentation.html.

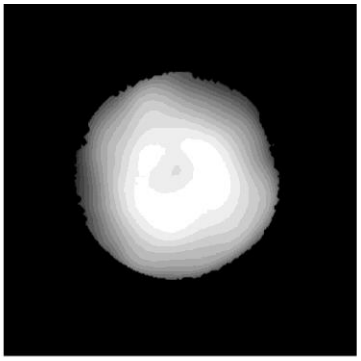

Figure 2. *Gray scale image of the segmented cell.*

For each of the segmented red blood cells we consider the following as covariates: mean phase, ratio of major axis length to the area of the cell, ratio of minor axis length to the area of the cell, ratio of the perimeter to the area of the cell, coefficient of variation, solidity, eccentricity, circularity and area. Major (minor, respectively) axis length is defined as the number of pixels on the major (minor, respectively) axis of the ellipse that has the same normalized second central moments as the segmented cell. Perimeter is defined the distance between each adjoining pair of pixels around the border of the segmented cell. Area is defined as the number of nonzero pixel values in each segmented image which is equivalent to the number of pixels occupied by the cell in each image. Mean phase is defined as the average pixel value in the image. Coefficient of variation is defined as the ratio of the standard deviation to the mean of the phase values in the segmented image. Solidity is defined as the proportion of pixels in the convex hull that are also in the segmented cell. Eccentricity is computed as the eccentricity of the ellipse that has the same second moments as the segmented cell, where eccentricity for the ellipse is defined as is the ratio of the distance between the foci of the ellipse to its major axis length. Circularity is defined as the ratio of major axis length to the minor axis length.

One of the advantages of these predictors is that, they are rotation and translation invariant. This is a very desirable property, as the orientations of the blood cells doesn't affect the complexity of the classification problem. The three covariates solidity, eccentricity and circularity as we have defined above, broadly try to capture the deviation of the shape of the segmented cell from a perfect circle. This provides an interesting insight as the red blood cells are known to have the shape of flat disc. We aim to capture any departure from this behavior.

Method

Mixture discriminant analysis: Let $\mathcal{D} = \boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ denote the training data set where each $\boldsymbol{x}_i \in R^p$ is the observation i with true and unique membership $y_i \in 1, \dots, K$. In general, discriminant analysis assigns an unlabeled p-dimensional observation $\boldsymbol{x}$ to one of the K known unique classes by estimating a function $f(\boldsymbol{x}) = y$, which determines the class in terms of y. Using $\mathcal{D}$, a classifier is trained to select the most probable class label for $\boldsymbol{x}$ as

$$\hat{y} = \hat{f}\,(\boldsymbol{x}) = \max_{\mathrm{k}} p\,(y = k|\boldsymbol{x}) = \max_{\mathrm{k}} \pi_k\, f_k(\boldsymbol{x})$$

where $f_k(\boldsymbol{x})$ is the class-conditional probability density function and π_k is the prior probability of class membership for class k, such that

$$0 \le \pi_k \le 1 \forall k \in \{1,2,\dots,K\} \text{ and } \textstyle\sum_{k=1}^{K} \pi_k = 1.$$

This is called the Bayes classifier.

In mixture models, as described by [11], given the data $\mathcal{D}$ the likelihood for the model with K components is given by

$$\mathrm{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_K, \pi_1, \pi_2, \ldots, \pi_K) | \mathcal{D} = \prod_{i=1} \sum_{k=1}^{K} \pi_k \, f_k(\boldsymbol{x}_i | \boldsymbol{\theta}_k) \ ,$$

where f_k is the density of the component k, β_k are its corresponding parameters and π_k is the probability that an observation belongs to the component k, $\forall k \in \{1, 2, \ldots, K\}$. Generally, f_k is assumed to follow a multivariate normal density ϕ_k

$$\phi_k \, (x_i | \boldsymbol{\theta}_k) = \frac{\exp(-\frac{1}{2}(\, x_i - \mu_i)^T \sum_k^{-1}(\, x_i - \mu_i)}{(2\pi)^{\frac{p}{2}} \, | \sum_k |^{\frac{1}{2}}} \quad (1)$$

Where the mean and covariance parameters are given by $\boldsymbol{\theta}_k = (\mu_k, \sum_k)$. This indicates that clusters or classes centered at μ_k characterize the data generated by multivariate normal distribution. The covariance matrix $\sum_k$ determines the geometric features of the clusters. This can also be used to impose constraints between clusters. The most commonly used structures for $\sum_k$ are given by $\sum_k \ = \lambda I$ or $\sum$. the former indicates that all the clusters are spherical and of the same size, and the later indicates that the geometry of all the clusters remains identical but it need not be spherical. The parameter estimation for mixture models can be done using the Expectation-Maximization (EM) algorithm by defining a latent variable corresponding to each observation, as the indicator function of the clusters.

Under the assumption of a constant covariance matrix, $\sum_k = \sum \ \forall \ k \epsilon \{1, 2, \ldots \mathrm{K}\}$ and that the maximum likelihood estimates of μ_k and $\sum_k$ are obtained, Fisher's linear discriminant analysis is given by the conditional Bayes classifier. If we consider an unconstrained$\sum_k$, then the resulting method is called the standard quadratic discriminant analysis (QDA). Mixture Discriminant Analysis is an extension of LDA and uses mixture of normal distributions for the density estimation of each class. Often, linear decision boundaries are insufficient for classification. Furthermore, a single Gaussian distribution might be too conservative assumption in characterizing a single class. With these points as motivation, [2] proposed mixture discriminant analysis. Let

$$f_k(\boldsymbol{x}) = \sum_{\{r=1\}}^{R_k} \pi_{kr} \, N_p(\boldsymbol{x}_i | \mu_{kr} \, , \sum \,) \quad (2)$$

be the probability density function of a finite Gaussian mixture density with R_k components, where the mixture density of the component r has the prior probability π_{kr} such that $0 \leq \pi_{kr} \leq 1$ for all $r \in \{1, 2, \ldots, R_k\}$ and $k \in \{1, 2, \ldots, K\}$, and $\sum_{r=1}^{R_k} \pi_{kr} = 1$ for each k. $\sum$ is assumed to be identical across all classes and subclasses. Figure 3 gives an example of Gaussian mixture discriminant analysis with two classes ($K = 2$), where the class on the left has two subclasses ($R_1 = 2$) within and the class on the right has three subclasses ($R_2 = 3$) within it.

As an extension to [2], a more general method was proposed by [11], where the number of subclasses within each class could be different along with having different covariance matrix for each class. For the application of mixture discriminant analysis, we use the *mclust 5* R package [12]. One of the major questions in mixture modeling is how to decide on the number of subclasses or the covariance structure being the same across all the classes. Model selection can be performed using information criteria, such as the Bayesian Information Criterion (BIC) [13] or the integrated complete-data likelihood criterion [14]. Formal hypothesis testing can be used to determine the optimal number of mixture components. A more recent review of this can be found in [15]. Both these model selection criteria have been incorporated in the *mclust 5* R package, and we employ these for our classification of red blood cells.

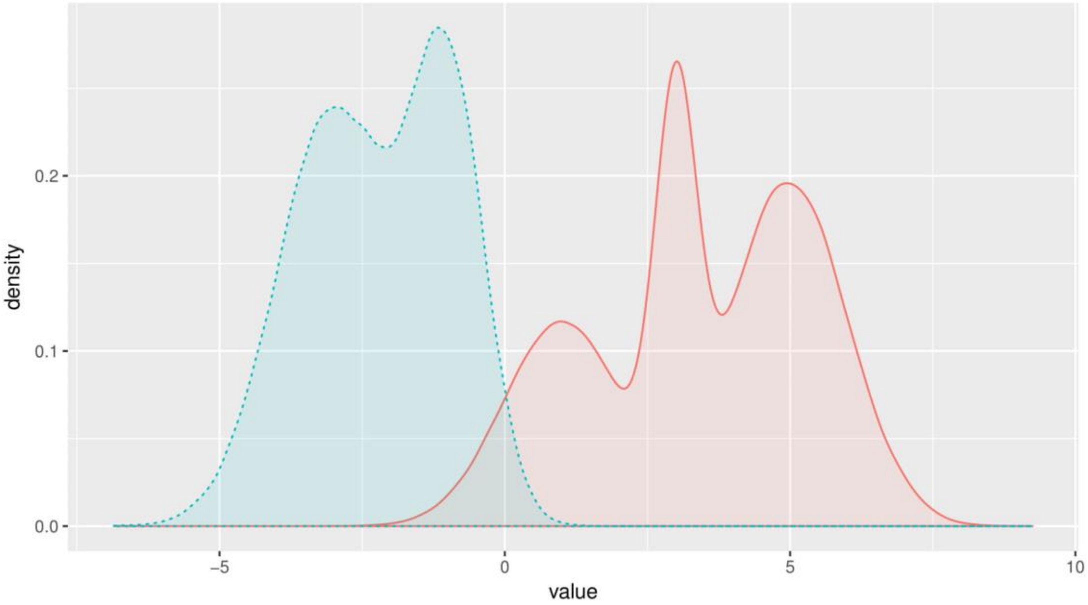

Figure 3*. Example with two and three subclasses.*

Neutral Zone Classifier: Most statistical classifiers make class predictions based on the value of the corresponding latent variable and/or the estimated probability of the possible classes. In a two-class classifier, a single threshold value is used to assign the predicted class. This forces us into assigning one of the two possible classes for each test case for class prediction. For example, let us consider a case where we employ logistic regression and would like to make a prediction based on the estimated probability and decide to use a threshold c as a cutoff probability to decide between the classes 0 or 1. For any $\epsilon > 0$ we would be forced to assign the class as 1 (0, respectively) for any estimated probability $c + \epsilon$ ($c - \epsilon$, respectively). That is, if the threshold is 0.5, we would be assigning the class as 1 in both the cases where the estimated probability is 0.51 or 0.99.

Using a single threshold value leads to producing higher number of false positives and false negatives, when the value of the latent variable or the estimated probability is closer to the corresponding threshold being used. To address this problem, [9] propose an alternative thresholding strategy called the neutral zone classifiers by modifying the two class Bayes classifier. As an extension, [10] incorporated a neutral zone region into the classifier so that ambiguous cases falling into the neutral zone can be further investigated before making a classification decision. Classifiers can thus be made useful as the conditional misclassification rates can be controlled. In the rest of this section, we briefly describe a neutral zone classifier with appropriate thresholding to obtain the corresponding class predictions.

Let $h(\boldsymbol{x_i})$ be the decision statistic being used to decide the class prediction for the subject i. Let the possible classes belong to $\{0,1\}$ and

$$\hat{y}_i = \begin{cases} 1 & if\ h(x_i) \geq c_1 \\ N & if\ c_0 < h(x_i) < c_1 \\ 0 & if\ h(x_i) \leq c_0 \end{cases}$$

where $c_0 < c_1$ and N denotes the decision of not assigning either of the classes during prediction. The choice of c_0 or c_1 could be based on the cost of indecision [9] or can be computed by simultaneously controlling the false positive and false negative rates [10]. For the latter case, α and β are fixed and c_0 and c_1 are computed such that $P(\hat{y}_i = 1|y_i = 0) = \alpha$ and $P(\hat{y}_i = 0|y_i = 1) = \beta$. Both these equations are solved by identifying the conditional distribution of $h(\boldsymbol{x_i})$ given $y_i = 0$ or $y_i = 1$. In the classifiers where a specific form is not observed for class conditional distributions of $\boldsymbol{x_i}$, the receiver operating characteristic (ROC) curve is used to compute c_0 and c_1. Once c_0 and c_1 are obtained, all the test cases which fall in the neutral zone could then be further analyzed before deciding on the predicted category.

Results

Thickness information of the red blood cells are obtained by phase-contrast imaging. Phase is obtained from the object hologram procured by digital holographic interferometric microscopy (DHIM) as described in [16]. We have 301×301 images of 27 red blood cells of which 13 cells were from healthy subjects and 14 cells from malaria-infected subjects. Each pixel represents the phase and is proportional to the *thickness* of the red blood cell at that pixel [16]. We would like to classify these 2D images of RBCs as either healthy RBC or malaria-infected RBC.

After the segmentation of the cells and the feature extraction as proposed in the earlier sections, each subject is represented by a 9-dimensional vector $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{i9})$ where x_{ij} represents the j^{th} covariate for i^{th} subject for $i = 1, 2, \ldots, 27$ and $j = 1, 2, \ldots, 9$. In this case, $K = 2$ as we have two classes: healthy and malaria infected. The stars plot in Figure 4 represents a row corresponding to a subject in the data set. The stars plot on the left panel corresponds to the healthy subjects and the one in the right panel corresponds to malaria infected subjects. Each star in the star plot contains a sequence of equiangular spokes where each spoke represents one of the predictors. The length of the spoke is proportional to the magnitude of the predictor for the data point, relative to the maximum value of the covariate across all data points.

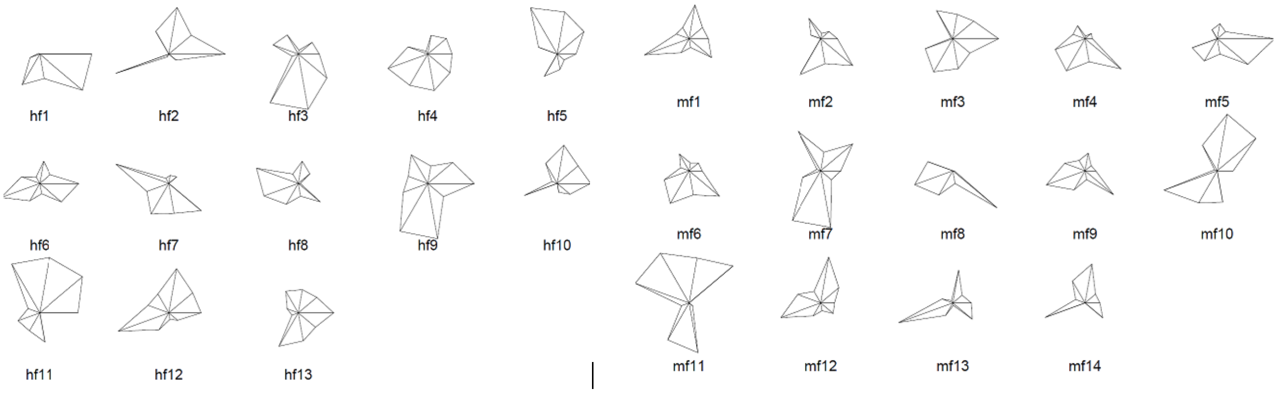

Figure 4. *Stars plot of predictors for healthy and malaria infected subjects.*

We now fit classification models through MDA using the R package *mclust 5*. We do the leave-one-out cross-validation (LOOCV) approach and estimate $\hat{y}_i^{(-i)} = \hat{f}^{(-i)}\,(\boldsymbol{x}_i)$, where $\hat{f}^{(-i)}$ is the classifier based on all the observations except the observation i. Prediction is done for both LDA and MDA, with LDA being considered as a special case of MDA. For MDA, model selection to choose the number of subclasses and the structure of covariance within the classes is based on BIC as described in [12]. For each subject in the LOOCV, we do a model selection. The average number of subclasses for the healthy class after the LOOCV was 3 and 4 subclasses for the malaria infected class.

***Table 1**: Confusion matrix corresponding to LDA and MDA*

	Predicted class (LDA)		**Predicted class (MDA)**	
TrueClass	Healthy	Infected	Healthy	Infected
Healthy	11	2	12	1
Infected	3	11	5	9

We now want to try and decrease the misclassifications from both LDA and MDA. Clearly the threshold used on the class posterior probabilities for prediction linear discriminant analysis was 0.5. It was the same for mixture discriminant analysis but, we first compute the posterior probabilities of the classes first from the posterior probabilities of the subclasses and then use the threshold value. To improve the model performance, we use the models fit through both the discriminant analyses and incorporate the neutral zone classifier strategies. We first compute the ROC curve using the class posterior probabilities and the true labels for the subjects.

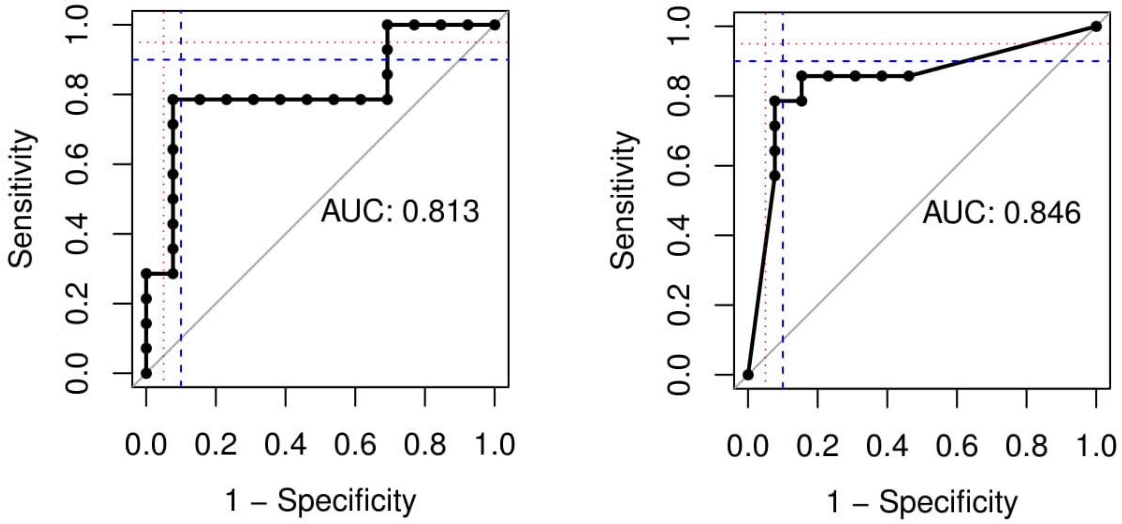

***Figure 5**. ROC curves for plotted for both LDA (left) and MDA (right). In both the plots, dotted (red) lines indicate a conditional misclassification of 5% and dashed (blue) lines indicate a conditional misclassification of 10%.*

Using Figure 5, we try to obtain the thresholds c_0 and c_1 to decrease the conditional misclassification rates. The results of the neutral zone classifier with both LDA and MDA are shown in Table 2. The first two rows show a confusion matrix corresponding to LDA and MDA, when the false positive rate (FPR) is fixed as $\alpha = 0.10$ and the false negative rate (FNR) is fixed as $\beta = 0.10$. We see that the number of misclassifications has reduced from 5 subjects to 4 subjects using LDA and 6 subjects to 4 subjects using MDA. However, using neutral zone classifier in both these cases reduces the number of subjects on which a definitive decision of prediction is made. We see that 8 subjects are left out using LDA and 5 subjects using MDA. Similarly, when the FPR is fixed at $\alpha = 0.05$ and the FNR at $\beta = 0.05$, the third and fourth rows in Table 2 show that the misclassifications reduce to 2 and 3 for LDA and MDA, respectively. However, the number of subjects falling into the neutral zone increases as the FPR and FNR are controlled by lower bounds. Also note that, since we have only 27 subjects, 5% or 10% control on the error rates bounds number of misclassifications as 1 or 2, respectively.

Table 2*: Confusion matrices corresponding to neutral zone classifiers along with LDA and MDA. The first two rows represent results when FPR and FNR are both controlled at 10%. Results in the third and fourth rows correspond to controlling FPR and FNR at 5%. NA indicates no decision being made.*

		Predicted class (LDA)			**Predicted class (MDA)**		
	True class	Healthy	Infected	NA	Healthy	Infected	NA
α = 0.10	Healthy	4	2	7	7	2	4
β = 0.10	Infected	2	4	1	2	11	1
α = 0.05	Healthy	4	1	8	7	1	5
β = 0.05	Infected	1	11	9	2	8	4

Discussion and Conclusions

In this paper, we have addressed the problem of automatic red blood cell classification by using statistical classification model, MDA, in which the granularity of the modeling is increased as the classes are modeled by mixture of Gaussian distributions. It is well known that the red blood cells look like biconcave disks with a flattened and depressed center, a dumbbell-shaped cross section, and a torus shaped rim on the edge of the disk. But the infected cells are deformed irregularly. On the 2D images of the red blood cells we first employ a cell segmentation approach to identify the RBC in the image. Once the RBC are segmented, we construct features from them which can then be used for further downstream analysis. Standard normality tests along with these geometric characteristics of the RBCs encourage the use of mixture of Gaussian distributions within each subclass of the healthy and malaria-infected cells. We demonstrated the performance of LDA and MDA. We then use neutral zone classifiers to improve the classification by relooking at those subjects whose prediction is not very certain from the modeling. Mixture discriminant analysis as implemented in *mclust 5* by [12], also does model selection in terms of selecting the number of subclasses and choosing the appropriate covariance structure. MDA also has the advantage of being a statistical model-based classification algorithm and not a hard-clustering method.

After such an analysis of the 2D images, the subjects which are referred to the neutral zone can be further studied to better understand the ambiguity in them. This method of classification also takes care of glaring errors in the data collection process such as, mislabeling the subjects. Referring subjects to further diagnosis before assigning a treatment due to a wrong prediction is also of utmost importance in many applications, specifically in the medical research. As the cells are obtained by centrifuging the blood, not all red blood cells of a subject infected with malaria, would be infected. This implies that we might have some healthy cell images from a malaria-infected subject. Incorporating neutral zone classifier can also help us to detect such situations by controlling the false negative rates. In a general scenario, we would try to classify multiple cells in a single image and then we can use the majority classification of the cells in the image as the assigned class of the subject.

When we look at the microscopic images of blood, there are large number of blood cells within the image. Instead of looking at each cell separately, we could use similar cell segmentation algorithms to retain all the cells within the image. We can then extract the features of each of these cells and use them for classification. We could classify each cell within an image using MDA. These classifications can now be used to make informed decisions about malaria. Also, MDA offers the additional choice of fixing the number of subclasses by choosing them according to prior knowledge of the application and the properties of the data at hand. This is a very desirable feature for the modeling approach. Mixture discriminant analysis could also be used for the automatic classification of white blood cells (WBC) using the microscopic images, as WBCs are known to exist in different structures based on their functional and physical characteristics [17].

Acknowledgments

The authors gratefully acknowledge *Dr. Arun Anand*, M S University of Baroda, India and *Dr. Bahram Javidi*, University of Connecticut, USA for providing us with the data. We also thank *Dr. Yuping Zhang*, University of Connecticut, USA for the inputs.

Conflicts of Interests

The authors declare that there is no conflict of interest regarding the publication of this article.

References

1. World Health Organization. World malaria report 2015. World Health Organization; 2016 Jan 30.
2. Hastie T, Tibshirani R. Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society: Series B (Methodological). 1996 Jan;58(1):155-76.
3. Schmid U, Roesch P, Krause M, Harz M, Popp J, Baumann K. Gaussian mixture discriminant analysis for the single-cell differentiation of bacteria using micro-Raman spectroscopy. Chemometrics and Intelligent Laboratory Systems. 2009 Apr 15;96(2):159-71.
4. Choi SW, Park JH, Lee IB. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. Computers & Chemical Engineering. 2004 Jul 15;28(8):1377-87.
5. Haeb-Umbach R, Geller D, Ney H. Improvements in connected digit recognition using linear discriminant analysis and mixture densities. IEEE International Conference on Acoustics, Speech, and Signal Processing 1993 Apr 27 (Vol. 2, pp. 239-242). IEEE.
6. Harris G, Andreasen NC, Cizadlo T, Bailey JM, Bockholt HJ, Magnotta VA, Arndt S. Improving tissue classification in MRI: a three-dimensional multispectral discriminant analysis method with automated training class selection. Journal of Computer Assisted Tomography. 1999 Jan 1;23(1):144-54.
7. Bacus JW. Method of and an apparatus for automatic classification of red blood cells. United States Patent US 4,097,845. 1978 Jun 27.
8. Tomari R, Zakaria WN, Jamil MM, Nor FM, Fuad NF. Computer aided system for red blood cell classification in blood smear image. Procedia Computer Science. 2014 Jan 1;42:206-13.
9. Jeske DR, Liu Z, Bent E, Borneman J. Classification rules that include neutral zones and their application to microbial community profiling. Communications in Statistics-Theory and Methods. 2007 Aug 7;36(10):1965-80.
10. Jeske DR, Smith S. Maximizing the usefulness of statistical classifiers for two populations with illustrative applications. Statistical Methods in Medical Research. 2018 Aug;27(8):2344-58.
11. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association. 2002 Jun 1;97(458):611-31.
12. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal. 2016 Aug;8(1):289.
13. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-

based cluster analysis. The Computer Journal. 1998 Jan 1;41(8):578-88.
14. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000 Jul;22(7):719-25.
15. McLachlan GJ, Rathnayake S. On the number of components in a Gaussian mixture model. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014 Sep;4(5):341-55.
16. Anand A, Chhaniwal VK, Patel NR, Javidi B. Automatic identification of malaria-infected RBC with digital holographic microscopy using correlation algorithms. IEEE Photonics Journal. 2012 Oct;4(5):1456-64.
17. Wheeler BS, Rebecca M. Exploring Medical Language: A Student-Directed Approach. Journal of Health Occupations Education. 1989;4(2):9.