

Blood corpuscles classification schemes for automated diagnosis of hepatitis

Luminița STATE * Iuliana PARASCHIV-MUNTEANU †

Nicolaie POPESCU-BODORIN ‡

January, 2009

Abstract

The paper proposes three methods for automated diagnosis of hepatitis based on shape recognition and classification of blood corpuscles. The processed data are binary images of blood samples taken from positive diagnostic and negative diagnostic patients. The idea followed in the paper is to identify T-lymphocyte response (Danne corpuscles) using approximations of shape coefficients as perimeter, area and circularity. The automatic diagnosis is accomplished through an analysis process on the class of the most circular corpuscles identified in the image. A series of experimentally derived conclusions are supplied in the final part of the paper.

2000 Mathematics Subject Classification: Medical applications (general) (92C50), Pattern Recognition (68T10), Image processing (68U10).

Key words and phrases: k -Means algorithm, Automatic diagnosis, T-lymphocyte, Image processing, Classification.

*Faculty of Mathematics and Computer Science, University of Pitești

†Faculty of Mathematics and Computer Science, University of Bucharest

‡Faculty of Mathematics and Computer Science, University Spiru Haret

1 Introduction

This paper is a case study on the automated diagnosis of hepatitis based on shape recognition and classification techniques. The methods proposed herein seek to obtain a positive/negative diagnostic message following the processing of a binary image (figure 1) representing the corpuscles found in the serum taken from the patient. They are based on the following working hypotheses ([2]):

- the corpuscles in the input image can be of three types: small and nearly circular (circa 22 nm across), large and nearly circular (T-lymphocyte or Danne corpuscles, circa 42 nm across), and non-circular corpuscles (circa 22 nm wide and 20-250 nm long);
- the presence of Danne corpuscles is associated with hepatitis.

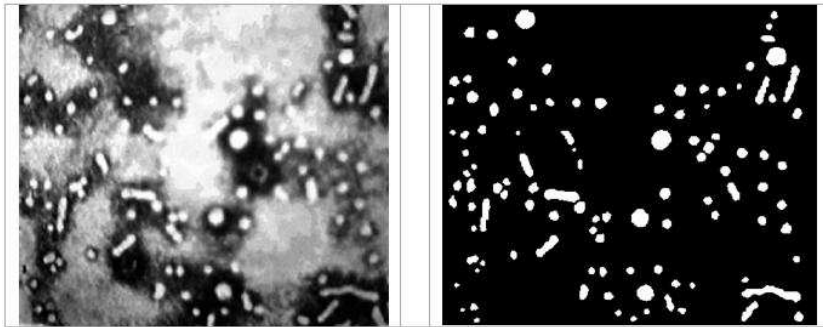


Figure 1: *Corpuscles within the serum of a hepatitis patient.*

The automated diagnosis methods presented in the following are briefly exposed in figure 2): all the available corpuscles are extracted from the binary image and approximations of area, perimeter and circularity are computed for each corpuscle. The almost circular corpuscles are separated from the rest on the basis of circularity coefficients. This separation can be done using either an explicit, more or less subjective threshold value obtained from observation and medical expertise, or in terms of a threshold value obtained directly from the data using a clustering algorithm. According to their area, circular corpuscles are further classified into two subclasses: the negative diagnostic class made up of corpuscles with a smaller area than the threshold, and the positive diagnostic class

consisting of corpuscles of area larger than the threshold.

Because the evaluation of circularity in terms of area and perimeter using the coefficient \mathbf{C} as a function of the area \mathbf{A} and perimeter \mathbf{P} of a corpuscle,

$$\mathbf{C} = \frac{4\pi\mathbf{A}}{\mathbf{P}^2}, \quad (1)$$

usually yields to significant error, in section 2 we introduce an estimate of the circularity and confirm its usefulness on experimental basis.

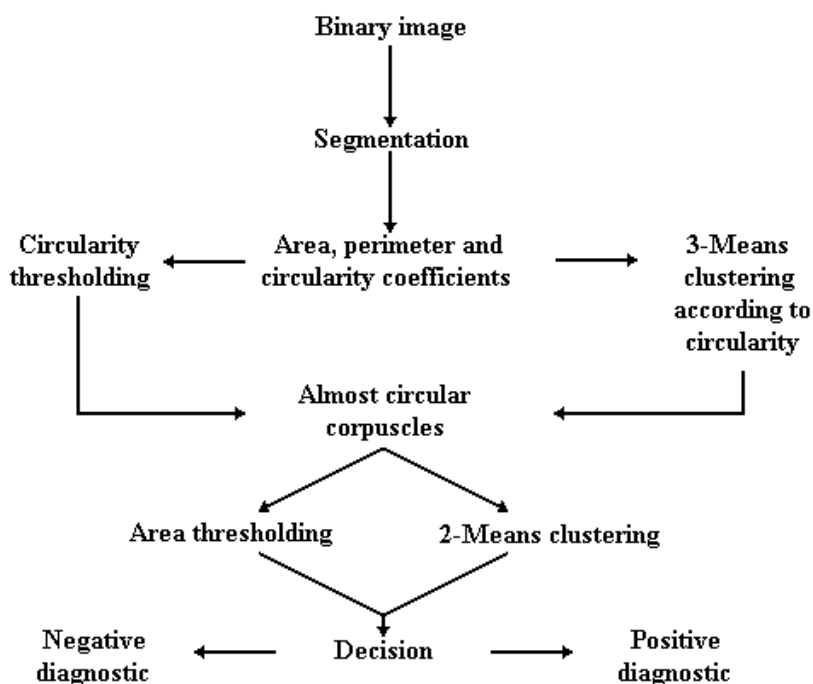


Figure 2: *Automatic hepatitis diagnostic.*

A deterministic approach based on explicit thresholding is briefly presented in section 3. Next, an approach based on the k -means algorithm is presented in some detail allowing the identification of almost circular corpuscles and the separation of this class in small and large circular corpuscles.

A subsidiary problem to the clustering results obtained through k-means algorithm is that the significance of determined clusters may vary according to the input image: for an image containing noise (corpuscles of null or near-null circularity), the first cluster (of the most non-circular corpuscles) will tend to contain noise only, whereas for an image that does not contain any noise, the first cluster will tend to include the longest corpuscles (with relatively low circularity, yet not close to zero).

2 Pixel-based estimation of corpuscle shape

The motivation for searching for a new computation formula to compute circularity coefficients is the fact that the area and perimeter of corpuscles in the input image should be evaluated exclusively on the basis of the information contained in the pixels. In case of digital images, the boundary of a sub-image (corpuscle) and its area become commensurable since they share the same physical support (given by a set of pixels). For a sub-image S of area \mathbf{A}_0 and perimeter \mathbf{P}_0 , the circularity coefficient is defined by

$$\mathbf{C} = \frac{\sqrt{\frac{\mathbf{A} - \mathbf{P}}{\pi}}}{\frac{\mathbf{P}}{2\pi}}, \quad (2)$$

where \mathbf{A} and \mathbf{P} are estimates for \mathbf{A}_0 si \mathbf{P}_0 : \mathbf{A} is the total number of pixels in \mathbf{S} and \mathbf{P} is the number of pixels within the contour of \mathbf{S} , where the contour is given by the set of those pixels belonging to S which are placed in the neighborhood of the complementary of S . Therefore $(\mathbf{A} - \mathbf{P})$ is an estimate of the inner area.

Consequently, $\sqrt{\frac{\mathbf{A} - \mathbf{P}}{\pi}}$ and $\frac{\mathbf{P}}{2\pi}$ can be taken as empirical radii of S computed on the basis of inner pixels and perimeter, respectively.

The previous empirical estimates are consistent with the theoretical concepts as defined in topology based developments in image processing ([13], [14], [15]).

3 A Deterministic Morphological Approach

The methodology presented in section 1 for the diagnosis of hepatitis on the basis of the circularity coefficient yields the following computational procedure.

Algorithm DMA:

1. **Input:**
 - a binary image of sampled serum;
 - a circularity threshold;
 - an area threshold corresponding to the positive diagnostic (i.e. the case when at least one Danne corpuscle is detected);
2. Identification and extraction of all corpuscles within the input image;
3. Computation of the estimates of perimeter, area and circularity for each corpuscle according to formula (2);
4. Check for nearly-circular corpuscles using the selected circularity coefficients and the circularity threshold. In case no nearly-circular corpuscles are detected an error message informing about the poor quality of the image has to be generated.
5. Separate the set of all nearly-circular corpuscles into two subclasses according to the selected area threshold: the former entailing negative diagnostic and the latter containing the abnormally super-sized nearly circular corpuscles, i.e. the physical support of a positive diagnostic (in case a Danne corpuscle is detected the patient is suspected of hepatitis);
6. In case the subclass of corpuscles of size larger than the threshold contains at least an element the positive diagnostic is decided, otherwise the negative diagnostic is accepted.
7. **Output:** The decision concerning the diagnostic positive/negative or an informing message about the absence of circular corpuscles in the analyzed image is generated.

Two main questions concerning the usefulness of DMA algorithm should be answered. On the one hand, a question arises about the quality of formula (2) of expressing the corpuscles circularity. Experimentally it is established that the values obtained using (2) mostly belong to $[0, 1]$, very seldom values larger than 1.1 are obtained. Values larger than 1 indicate that the value of the perimeter approximation is smaller than the actual value. This is due to the fact that the contours of sub-images are imprecise and it can happen that some pixels actually belonging to the contour are allotted either to inner or to outer parts of the sub-image.

Fortunately, this effect occurs in cases of rather large and almost circular corpuscles, therefore values larger than 1 of the circularity coefficient given by (2) do not affect the diagnostic decision.

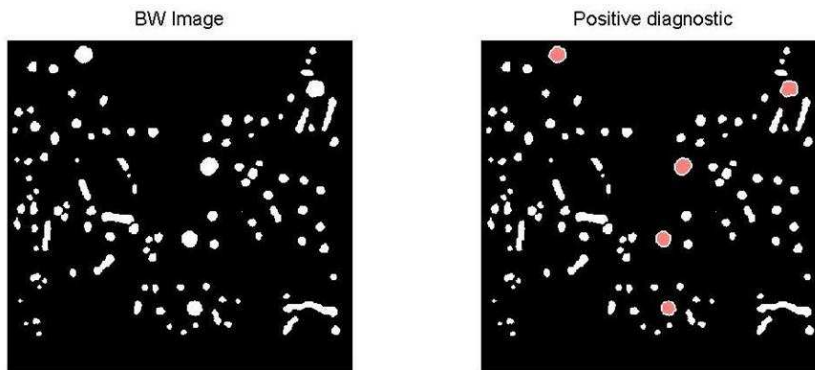


Figure 3: *A positive diagnostic.*

On the other hand, the problem is to obtain a suitable value for the threshold used to separate the almost circular corpuscles in the image. This problem can be solved either heuristically using the medical expertise or, in case a significantly large database of images whose diagnostics are confirmed is available, it can be solved on statistical basis. There still remains the problem of inferring suitable threshold value when neither medical expertise nor large size databases of confirmed images are available. This problem can be handled many ways. One possibility is to find out the natural grouping tendency of corpuscles in the tested image and to derive a threshold value by some sort of averaging technique. This idea is followed in the next two sections where the k -Means algorithm is

used to cluster the set of corpuscles in the images.

The results of some tests performed on the algorithm DMA are presented in figure 3.

4 The k -Means Based Automated Diagnosis Algorithm (k -MBADA)

The celebrated k -means clustering algorithm was introduced by MacQueen in 1967 [7]. Basically, starting with a set of k randomly selected "seeds" taken as initial cluster centers, the k -means algorithm initiates an iterative clustering process where, by reclassifying the elements according to the current set of centers, at each step, the variability within clusters is reduced, and the cluster centers are directed toward stable positions. The algorithm stabilizes when no significant changes of center positions are detected.

An adapted version of the k -means algorithm that computes the clusters according to the corpuscle circularities is presented as follows:

k-means procedure:

Input: k desired number of clusters ;

$\mathcal{X} = \{ \mathbf{X}_i \in \mathbf{R} \mid i = \overline{1, N} \}$ the set of data to be classified;

\mathcal{C} the stopping condition;

the algorithm computes a classification of \mathcal{X} into k clusters;

d dissimilarity index;

Initializations: $m = 0$ the iteration index;

$\mathbf{Z}_1(0), \dots, \mathbf{Z}_k(0) \in \mathcal{X}$ initial centers randomly
selected from the set $\mathbf{X}_1, \dots, \mathbf{X}_N$;

repeat

$sw = false$;

for each element \mathbf{X}_i in the data set:

determine one of the closest centers $\mathbf{Z}_r(m)$, such that

$$d(\mathbf{X}_i, \mathbf{Z}_r(m)) = \min_{j=\overline{1, k}} d(\mathbf{X}_i, \mathbf{Z}_j(m)) ;$$

allot \mathbf{X}_i to the cluster $\mathcal{C}_r(m)$,

endfor

for each of the computed clusters $\mathcal{C}_j(m)$, $j = \overline{1, k}$:

update the cluster centers such that the function $J : \mathcal{X} \rightarrow [0, \infty)$

is minimized on each cluster,

$$J(\mathbf{Z}) = \sum_{\mathbf{X}_i \in C_j(m)} d^2(\mathbf{X}_i, \mathbf{Z});$$

endfor

if $\mathbf{Z}_j(m+1) = \mathbf{Z}_j(m)$, $j = \overline{1, k}$ then

$sw = true;$

endif

$m \leftarrow m + 1$

until sw or \mathcal{C}

Output: the clusters $C_1(m), \dots, C_k(m)$ of centers $\mathbf{Z}_1(m), \dots, \mathbf{Z}_k(m)$.

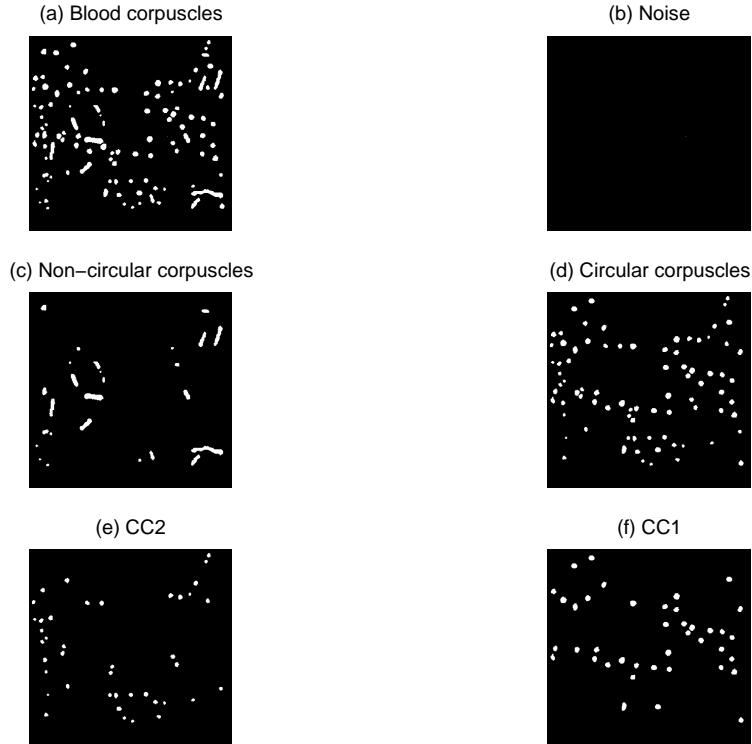


Figure 4: Clusters computed by 3-means applied with respect to circularity coefficients ((b),(c),(d)) followed by 2-means with respect to area applied to the cluster of circular corpuscles ((e),(f)). The blood sample (a) comes from a patient of negative diagnostic.

Note that the clusters produced by the *k*-means algorithm depend on the initial centers, and moreover, the stabilization is not guaranteed. The stopping condition \mathcal{C} can be imposed many ways. Typically, \mathcal{C} becomes *true* either when

$$d(\mathbf{Z}_j(m+1), \mathbf{Z}_j(m)) < \varepsilon, \text{ for all } j = 1, \dots, k,$$

or $d_m < \varepsilon$, where ε is a given positive number and

$$d_m = \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j(m)} d^2(\mathbf{x}_i, \mathbf{Z}_j).$$

The clusters obtained by applying the 3-means algorithm for data representing the values of the corpuscles circularity have the following meaning:

- the first cluster of corpuscles whose circularity are close to 1 corresponds to almost circular corpuscles identified in the image;
- the second cluster contains the non-circular corpuscles;
- the third cluster contains atypical corpuscles taken as noise.

In order to use the clusters computed by 3-means algorithm for diagnosis purposes, several questions arise:

- find a basis to express what it is meant by a large circular corpuscle;
- check whether large circular corpuscles exist in the image entailing positive diagnostic;
- find a simple yes /no diagnostic procedure.

The approach aiming to derive answers to the above questions is the following:

Step 1: Apply 2-means algorithm to the class of circular corpuscles to classify data according to the corpuscle areas. Let \mathcal{CC}_1 and \mathcal{CC}_2 by clusters of "large" circular corpuscles and "small" circular corpuscles respectively.

Step 2: Compute the selection means

$$\mu_1 = \frac{1}{|\mathcal{CC}_1|} \sum_{\mathbf{x}_i \in \mathcal{CC}_1} \mathbf{A}(i), \quad \mu_2 = \frac{1}{|\mathcal{CC}_2|} \sum_{\mathbf{x}_i \in \mathcal{CC}_2} \mathbf{A}(i) \quad (3)$$

and

$$\eta = \frac{\mu_1}{\mu_2}. \quad (4)$$

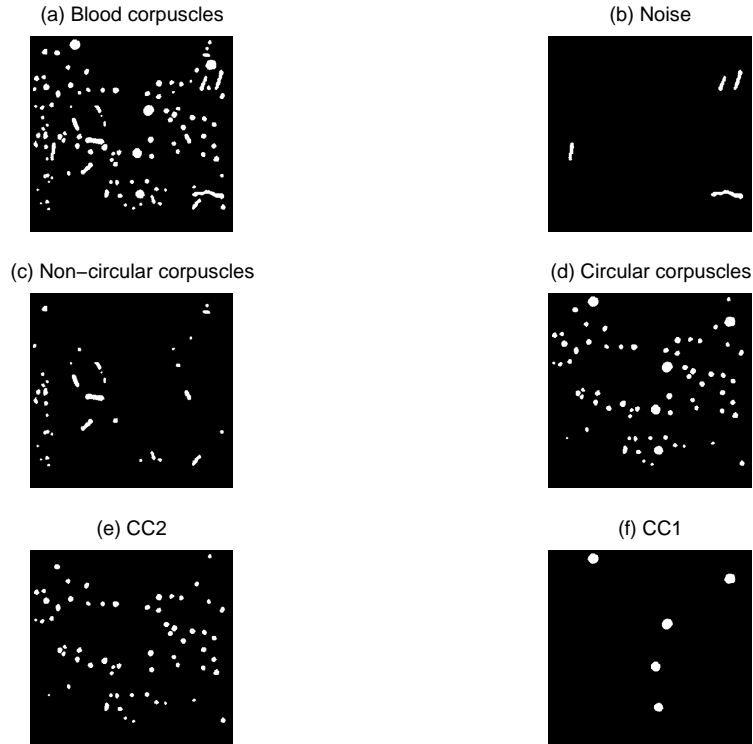


Figure 5: *Clusters computed by 3-means applied with respect to circularity coefficients ((b),(c),(d)) followed by 2-means with respect to area applied to the cluster of circular corpuscles ((e),(f)). The blood sample (a) comes from a patient of positive diagnostic.*

The tests performed on several images corresponding to positive and negative diagnostic point out that the value of parameter ratio is close to 4. This is an argument to justify that (2) provides good approximation of circularity. For the tests performed on the data displayed in figures 4 and 5 we get $\eta = 3.98$ and $\eta = 1.81$ in case of images of positive diagnostic and images of negative diagnostic respectively. Tables 1, 2, 3 and 4 present the first and the second order statistics corresponding to \mathcal{CC}_1 and \mathcal{CC}_2 .

This analysis yields the following yes/no/suspect diagnostic procedure:

Procedure k-MBADA:

Input: $\mathcal{X} = \{ \mathbf{X}_i = (\mathbf{A}(i), \mathbf{P}(i)) \mid i = \overline{1, N} \}$ the set of pairs area and perimeter of identified corpuscles;

Initializations:

the matrix $\mathbf{AP}(i, j)$ where $\mathbf{AP}(i, 1) = \mathbf{A}(i)$ and $\mathbf{AP}(i, 2) = \mathbf{P}(i)$;
the computed circularity coefficients $\mathbf{C}(i)$, $i = \overline{1, N}$, using (2);

1. Apply 3-means procedure to classify data according to the circularity coefficients \mathbf{C} ; get the clusters \mathcal{N} , \mathcal{NC} and \mathcal{CC} consisting of noise, non-circular corpuscles and circular corpuscles respectively.
2. Apply 2-means procedure to \mathcal{CC} according to the corpuscle areas \mathbf{A} ; get \mathcal{CC}_1 , \mathcal{CC}_2 the clusters containing "large" circular corpuscles and "small" circular corpuscles respectively.
3. Compute μ_1 , μ_2 , η according to (3), (4).
4. If $\eta \geq 4$ then answer = yes;
If $2 \leq \eta < 4$ then answer = suspect;
If $\eta < 2$ then answer = no;

Output: Diagnostic decision answer.

	Noise	Non-circular corpuscles	Circular corpuscles
<i>average area</i>	5	107.7500	81.3118
<i>standard deviation of area</i>	0	111.4389	25.1499
<i>minimum area</i>	5	13	17
<i>maximum area</i>	5	518	131
<i>average circularity</i>	0	0.7916	0.9446
<i>standard deviation of circularity</i>	0	0.0773	0.0312
<i>minimum circularity</i>	0	0.5094	0.8730
<i>maximum circularity</i>	0	0.8644	1.0705

Table 1: Statistics for circularity/area of clusters obtained by initial 3-means. (see figure 4 (b), (c), (d))

	\mathcal{CC}_1	\mathcal{CC}_2
<i>average area</i>	53.5588	97.3051
<i>standard deviation of area</i>	15.2281	12.7960
<i>minimum area</i>	17	76
<i>maximum area</i>	72	131
<i>average circularity</i>	0.9362	0.9494
<i>standard deviation of circularity</i>	0.0322	0.0295
<i>minimum circularity</i>	0.8807	0.8730
<i>maximum circularity</i>	1.0027	1.0705

Table 2: *Statistics for circularity/area of clusters obtained by 2-means applied to the cluster of detected circular corpuscles. (see figure 4 (e), (f))*

The resulted value $\eta = 1.8728$ entails negative diagnostic.

The computed value of η is 4.2157, hence the output of the algorithm is *answer = yes*, entailing positive diagnostic.

	Noise	Non-circular corpuscles	Circular corpuscles
<i>average area</i>	283.8333	72.7381	124.8118
<i>standard deviation of area</i>	154.7949	57.6872	97.9888
<i>minimum area</i>	16	13	17
<i>maximum area</i>	518	251	369
<i>average circularity</i>	0.6386	0.8313	0.9618
<i>standard deviation</i>			
<i>standard deviation of circularity</i>	0.0784	0.0395	0.0360
<i>minimum circularity</i>	0.5094	0.7508	0.8985
<i>maximum circularity</i>	0.7307	0.8932	1.0705

Table 3: *Statistics for circularity/area of clusters obtained by initial 3-means. (see figure 5 (b), (c), (d))*

Several tests have been performed on both, positive and negative blood samples and the accuracy of k -MBADA proved very high. However one major disadvantage resides from the fact that it involves the application of the k -means algorithm twice. Obviously, we can not give

up with the application of the 3-means algorithm to isolate the circular corpuscles present in the image. We can replace the application of the 2-means algorithm to the cluster of detected circular corpuscles by a heuristical rule of thumb as follows.

	\mathcal{CC}_1	\mathcal{CC}_2
<i>average area</i>	79.9130	318.4375
<i>standard deviation of area</i>	25.5636	44.7981
<i>minimum area</i>	17	258
<i>maximum area</i>	131	369
<i>average circularity</i>	0.9502	1.0117
<i>standard deviation of circularity</i>	0.0293	0.0099
<i>minimum circularity</i>	0.8985	1.0000
<i>maximum circularity</i>	1.0705	1.0232

Table 4: *Statistics for circularity/area of clusters obtained by 2-means applied to the cluster of detected circular corpuscles. (see figure 5 (e), (f))*

Let \mathbf{A}_{mean} be the arithmetic mean of circular corpuscles areas belonging to \mathcal{C} the cluster of circular corpuscles. We split this cluster into three sub-clusters according to the thresholds \mathbf{A}_{mean} and $2\mathbf{A}_{mean}$,

$$\begin{aligned}
 \mathcal{CP} &= \{ \mathbf{X}_i \in \mathcal{C} \mid \mathbf{A}_i \geq 2\mathbf{A}_{mean} \} , \\
 \mathcal{CN} &= \{ \mathbf{X}_i \in \mathcal{C} \mid \mathbf{A}_i < 2\mathbf{A}_{mean} \} , \\
 \mathcal{CS} &= \{ \mathbf{X}_i \in \mathcal{C} \mid \mathbf{A}_{mean} \leq \mathbf{A}_i < 2\mathbf{A}_{mean} \} .
 \end{aligned} \tag{5}$$

In this case the computation of the diagnostic variable answer can be carried out as:

- if $|\mathcal{CP}| \neq 0$ then *answer* = *yes*;
- if $|\mathcal{CP}| = 0$ and $|\mathcal{CS}| > |\mathcal{CN}|$ then *answer* = *suspect* otherwise *answer* = *no*.

The application of this variant to the data presented in figure 6 corresponds to the analysis of a blood sample coming from a negative diagnostic and produces the sub-clusters of volumes $|\mathcal{CP}| = 0$, $|\mathcal{CS}| = 40$, $|\mathcal{CN}| = 41$ therefore the negative diagnostic is slightly entailed. In case

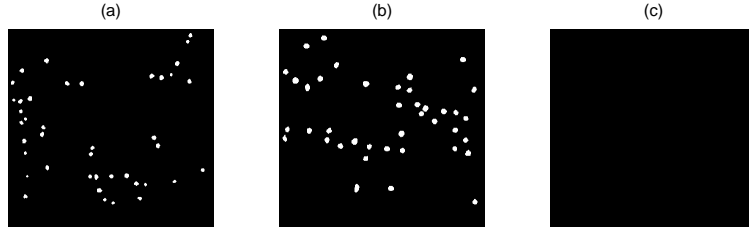


Figure 6: *The sub-clusters \mathcal{CN} (a), \mathcal{CS} (b), \mathcal{CP} (c) produced by the heuristic rule in case of a blood sample coming from a negative diagnostic.*

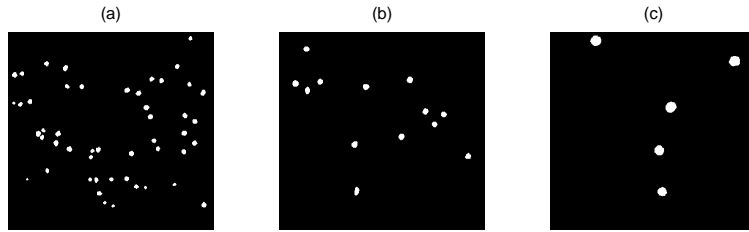


Figure 7: *The sub-clusters \mathcal{CN} (a), \mathcal{CS} (b), \mathcal{CP} (c) produced by the heuristic rule in case of a blood sample coming from a positive diagnostic.*

of a blood sample coming from a positive diagnostic the volumes of the resulted sub-clusters are $|\mathcal{CP}| = 5$, $|\mathcal{CS}| = 13$, $|\mathcal{CN}| = 49$ therefore the positive diagnostic is strongly entailed. In general, the performed tests proved, as it is expected, less accuracy as compared to the k -MBADA performance in discriminating between negative and not sure cases. Hopefully, in case of true positive diagnostic this variant is almost as good as k -MBADA.

The current method combines the deterministic identification and segmentation of the corpuscles with the heuristic classification of their shape based on the k -means clustering ([1], [7]) with respect to the circularity coefficients calculated as in formula (2).

5 Conclusive remarks

Several conclusions can be formulated on the experimental basis.

1. The circularity coefficient given by (2) proves good results in estimating the corpuscle circularity.

2. The values of the circularity coefficient for almost circular corpuscles lie between 0.9389 and 1.0116 while in case of non-circular corpuscles the values are significantly smaller (less than 0.82).

3. The heuristic thresholds used in the simplified variant of k -MBADA prove acceptable good in splitting the cluster of almost circular corpuscles into the group of corpuscles entailing positive diagnostic and the complementary set of corpuscles less significant in discriminating between yes/no answers.

Current efforts are focused on refining the proposed schemes and comparing the performance against other similar technics and the results are going to be published in a future report.

References

- [1] J. Abonyi, B. Feil, *Cluster Analysis for Data Mining and System Identification*, Birkh user Verlag AG Basel-Boston-Berlin, 2007.
- [2] L. Georgescu, N. Tudose, E. Potencz, *Morfopatologie*, Editura Didactica si Pedagogica, Bucuresti, 1982.
- [3] R.C. Gonzalez, R.E. Woods, *Digital Image Processing*, Addison Wesley Publ.Co., 2002.
- [4] A.K. Jain, M.N. Murty, Flynn P.J., *Data Clustering: A Review*, ACM Computing Surveys, Vol. 31, September 1991.
- [5] A.K. Jain, R.C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall advanced reference series, Prentice-Hall Inc., 1988.
- [6] B. Javidi (Ed.), *Image Recognition and Classification Algorithms, Systems, and Applications*, Marcel Dekker, Inc., 2002.
- [7] J.B. MacQueen, *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, (1967) 281–297.
- [8] A. Meyer-B se, *Pattern Recognition for Medical Imaging*, Elsevier Academic Press, 2004.

- [9] N. Popescu-Bodorin, L. State, *Optimal Luminance-Chrominance Down-sampling through Fast K-Means Quantization*, (The 3rd) Annual South-East European Doctoral Student Conference, vol. **2**, pp. 111-125, ISBN 978-960-89629-7-2, ISSN 1791-3578, South-East European Research Centre (SEERC, www.serc.org), June 2008.
- [10] W.K. Pratt, *Digital Image Processing*, Wiley & Sons, Inc., 2007.
- [11] J.C. Russ, *The image processing handbook*, 5th edition, CRC Press, 2007.
- [12] D. Salomon, *Data Compression. The Complete Reference.*, Fourth Edition, Springer-Verlag, 2007.
- [13] Karno Zbigniew, *The Lattice of Domains of an Extremally Disconnected Space*, Formalized Mathematics, **3**, Number 2, 1992.
- [14] T.Y. Kong, A. Rosenfeld, *Topological Algorithms for Digital Image Processing (Machine Intelligence and Pattern Recognition)*, Elsevier Science Inc., New York, NY, 1996.
- [15] T.Y. Kong, A. Rosenfeld, *Digital topology: introduction and survey*, Computer Vision, Graphics, and Image Processing, **48**, 3, p.357-393, 1989.