Descriptive Characteristics Generation and Selection for Acute Leukemia Subtype Classification from Bone Marrow Digital Images

M. Coral Galindo Domínguez[1], Jesús A. González[1] ,Leopoldo Altamirano Robles[1] and Ivan Olmos Pineda[2]

[1] Instituto Nacional de Astrofísica Óptica y Electrónica, Coordinación de Ciencias Computacionales
Calle Luis Enrique Erro No. 1 Sta. Ma. Tonanzintla Puebla México.
{cgalindo, jagonzalez, robles}@inaoep.mx
[2] Universidad Politécnica de Puebla
3er Carril del Ejido Serrano San Mateo Cuanala, Puebla, México.
ivanop_rkl@yahoo.com.mx

Abstract. In this paper we present a method for descriptive characteristics generation and selection applied to acute leukemia classification from bone marrow cells digital images. We obtain texture, geometric, statistical, and eigenvalues (PCA) as descriptive characteristics from our regions of interest (segmented cells). These characteristics are then used as input attributes to the data mining step (using different classifiers) to be able to classify acute leukemia families and then acute leukemia subtypes according to the FAB [1] classification. Our leukemia database also presented the class imbalance problem because of the proportion of cases of each subtype of leukemia, for which we applied over-sampling techniques. We made two types of evaluation, one with our domain expert and other using the cross validation technique with statistical significance. Results for each leukemia subtype were superior to 85% of accuracy.

Keywords: image processing, machine learning, acute leukemia.

1 Introduction

According to statistics shown by INEGI in 2005 [2], different types of cancer are becoming a large public health threat in Mexico. This is why physicians are working to improve its opportune detection, accurate diagnosis, and effective treatment. Leukemia is one of these public health enemies with a hospitalization rate of 49.7% for men and 18.4% for women and with a mortality rate of 6.3% for men and 5.5% for women.

Leukemia is a blood illness that produces an uncontrolled growth of white globules (a hematological cancer). Leukemia cells live longer than normal cells; they are different and also behave in a different way because they are unable to realize their main function (they should be antibodies to defend the organism). Leukemia is

Advances in Artificial Intelligence and Applications
Research in Computer Science 32, 2007, pp. 460–470

Received 30/06/07
Accepted 31/08/07
Final version 11/09/07

classified as myeloblastic or lymphoblastic, depending on the type of white globules affected and can be presented in acute or chronic ways. Acute leukemia grows faster while chronic leukemia grows gradually. If leukemia is detected in its first stages, its treatment might be more effective. This is where we focus our research, in early leukemia detection through morphological characteristics. Although other detection methods exist, morphological studies continue being important because they can be done at low cost and without expensive specialized equipment. That is, in spite of the recent advances in hematological skills such as flux cytometry (with immunophenotype), and DNA analysis; morphological analysis of bone marrow smears (even peripheral blood) continues being the starting point to detect patients that suffer of blood disorders. A morphological blood study consists of a cells study where cells are counted to verify if its amount is abnormal (according to specified limits) and then the cells morphology is examined to detect abnormal cells, this is done by a chemist or hematologist with experience in the detection of the disease (the morphological analysis is difficult and this is why it requires of experienced people to perform it). The problem with this type of analysis is that it is slow and it is not as precise as required because of its subjectivity (there are differences among physician's diagnosis or a physician's accuracy to detect the disease may vary depending on his state of tiredness). On the other hand, an automated morphological analysis to detect leukemia only requires images taken from bone marrow smears which are analyzed without the subjectivity related to tiredness or experience and could also reduce costs and improve accuracy; it can even be used remotely for places of low resources and could also be used to train chemists and hematologists. A tool like this is necessary for the diagnosis and treatment prescription of many diseases such as leukemia.

The aim of this work consists on a process to generate a set of geometric, texture, and statistical descriptive characteristics of acute leukemia subtypes as well as eigenvalues (obtained from principal component analysis or PCA's) from the segmentation (using color and boundaries detection) of regions of interest (ROI) or white globules cells. We use the obtained ROI's characteristics to feed classification algorithms such as neural networks, decision trees, and support vector machines among others to create discriminating patterns to classify leukemia subtypes.

This paper is structured as follows; we first describe our database in section 2. We then show our methodology to classify acute leukemia subtypes in section 3. In section 4 we present our experimental results and in section 5 we show our conclusions and future work.

2 Acute Leukemia Subtypes Classification Methodology

The classification of acute leukemia subtypes is a difficult task because affected cells are very similar. For this problem, we present the methodology shown in figure 1 to be able to classify acute leukemia subtypes. According to our methodology (see figure 1), we first select the images of patients for which an acute leukemia subtype has been detected. We then start the pre-processing phase, which consists of the segmentation of ROI's using first a color segmentation algorithm (to eliminate the background) and

then with a boundaries segmentation algorithm to get the ROI's. We then eliminate red globules to only keep white globules, from which the domain expert identifies lymphoblasts or myeloblasts that we will use for training with our learning algorithms. We then extract descriptive features from each ROI, which will be used to find patterns in the following step. Finally, the classification module is the one that we use to classify acute leukemia subtypes.

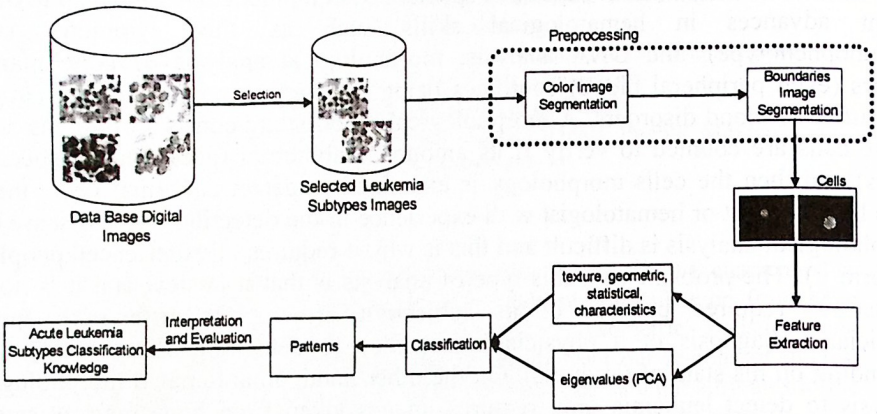

Fig. 1. Acute Leukemia Subtypes Classification Methodology

Our images database was created with the help of hematologists to select and classify acute leukemia cells in subtypes. In section 2.1 we describe our acute leukemia database.

2.1 Leukemia Data Base Description

Our database contains 1028 images from 74 patients. We have 415 images classified as Acute Lymphoblastic Leukemia (ALL) and 613 as Acute Myeloblastic Leukemia (AML) that were selected by domain experts. These images have a resolution of 800 x 600 pixels with a depth of 24 bits. We have only 11 cases of patients with ALL-L1, 8 patients with ALL-L2, 3 patients with AML-M2, 3 patients with AML-M3, and 1 patient with AML-M5. The number of cases presented before is related to the incidence of the disease subtypes. The number of images obtained of each leukemia subtype from the cases mentioned before is shown in table 1 as well as the number of ROI's (lymphoblast and myeloblast), which were labeled by the domain expert.

Table 1. Regions of Interest for every Subtype of Leukemia.

Acute Leukemia	Subtypes	Images	Regions of Interest
Lymphoblastic	L1	103	65
	L2	73	30
Myeloblastic	M2	57	38
	M3	48	26
	M5	11	10

2.2 Bone Marrow Smears Images Preprocessing

One of the most important phases of our methodology is the pre-processing of the images from which we obtain our ROI's through their segmentation so that we can obtain descriptive characteristics for each of them. Our pre-processing phase consists of four steps as shown in figure 2.

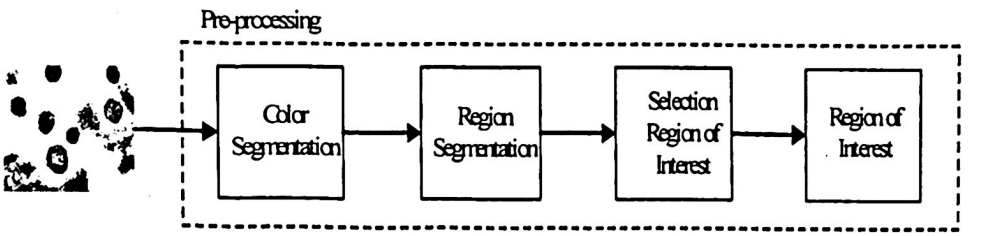

Fig. 2. Bone Marrow Images Pre-processing.

The image segmentation process consists of the division of an image in different ROI's, where each ROI has characteristics that distinguish it from others and each region is formed by a set of pixels. In this work we use a segmentation technique based on color and another to detect boundaries to separate leukemia cells (white globules) from the rest of the image. This is a very difficult task because in many cases the ROI's are overlaped and some others the boundaries of the cells are scarcely visible; furthermore, the colors of different types of ROI's (i.e. white and red globules) are very similar.

Color is an important descriptor in images analysis, it frequently helps to identify and extract regions of interest [3]. There exist different types of color formats such as RGB, CMYK, and Lab among others [4].

The RGB format works with three data channels where each them takes values from 0 to 255. An image in RGB is then composed of three bands with the primary colors of light Red (R), Green (V) and Blue (B) constructed with 8 bits per channel [4][5].

The Lab format consists of three channels, a channel L for Luminosity and two chromatic channels. The first chromatic channel, named A, ranges between green and red while the second, B ranges between blue and yellow. The luminosity component L ranges from 0 (black) to 100 (white). The chromatic components A and B range from +120 to -120 [4][5].

The original acute leukemia digital images are in RGB format, nevertheless; in order to improve the preprocessing efficiency we transformed them into the Lab format. That is, the RGB format is composed of three matrices, one for each channel, while in the Lab format we only work with a single matrix containing the information of channels A and B without considering the brightness channel (stored in a third matrix). Then, working with a single matrix reduces the work load. In figure 3 we show an image in the RGB and Lab formats. As we can see in figure 3, some ROI's (white globules) are very close to red globules (some of them even overlap), we can also see that the boundaries that delimit the cells are not very well defined and these are some of the problems that make the segmentation process difficult.

a b

Fig.3. Acute Lymphocytic Leukemia Subtype L1 Image. a) RGB Format. b) Lab Format.

2.1.1 Color Segmentation

As we mentioned before, for our color segmentation process we worked with the Lab format, omitting the luminosity channel. Figure 4 shows our general process for color segmentation, for which we used the K-means algorithm to group different colors of pixels in clusters to be used to find ROI's

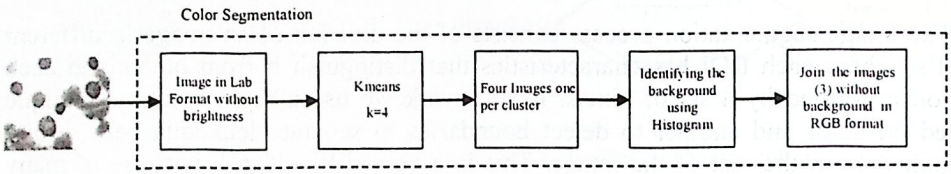

Fig. 4. Color Segmentation Process.

The K- means algorithm finds k groups, where each object belongs to one of the k groups. Each of these k groups is represented by a vector created as the average of the objects that belong to it [6].

In used the K-means algorithm [6] with the Euclidean distance and k = 4 to generate color groups. The chose k = 4 because the digital cell images of acute leukemia have 4 main tones, one for white globules, another for cytoplasm, one more for the background and finally one for red globules. We generated four images, one for each group as it is shown in Figures 5 a, b, c, and d respectively.

a b c d

Fig. 5. Acute Lymphocytic Leukemia Subtype L1 separated by color, with k = 4. a) White Globules. b) Cytoplasm. c) Background. d) Red Globules and Platelets.

For each of the images generated, we obtain the histogram, this will allow us to determine which images contain the ROI's (white globules), and which of them contains the background. In the last step we join three of the resultant images to create

an image without background in RGB format (we transform our Lab images to RGB format before mixing them) to get an image as shown in Figure 6.

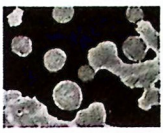

Fig. 6. Acute Lymphocytic Leukemia Subtype L1 separated by color, without background.

2.1.2 Regions Segmentation

After we eliminated the image background through color segmentation we perform a ROI's segmentation. In figure 7 we show our ROI's segmentation process.

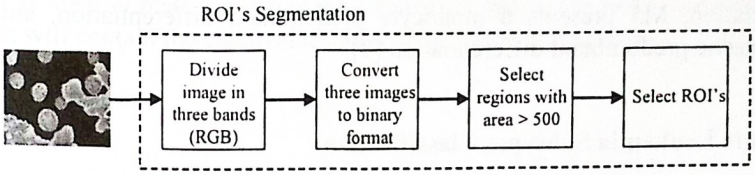

Fig.7. ROI's Segmentation Process.

Since we need to obtain descriptive characteristics from our ROI's, we use the images in their original format (RGB) and we separate them in their three channels to obtain their eigenvalues for each channel in gray format.

In this process we take the image segmented by color (without the background) and separate it in its three bands (RGB). Each image (3) is then converted into binary format; we then find the boundaries of each region and assign an index to each of them as shown in Figure 8 where the different tonalities indicate the different identified regions, we only keep regions with an area larger than 500 pixels because smaller areas correspond to noise. Finally, with the help of the domain expert we decided which of the identified regions corresponded to our ROI's (white globules) were.

Fig. 8. ROI's Segmentation of an Acute Lymphocytic Leukemia Subtype L1 Image identified by color.

ROI's in our case correspond to white globules from which we can identify acute leukemia (myeloblastic or lymphoblastic). The previous process is repeated for each

image in our database or for each new image to be classified. Once we have segmented our ROI's, we need to obtain their descriptive characteristics to be used as input to data mining algorithms and find patterns that distinguish among the different subtypes of leukemia.

For our classification process we use the French-American-British classification (FAB) based on morphologic and cytochemical characteristics that considers the maturity level of the cells (white globules) and the cellular line to which the white globule belongs (i.e. blasts, erythroblasts among others) [1]. According to the FAB classification [7], acute lymphocytic leukemia and acute myelogenous leukemia are divided in three and six groups respectively.

Acute lymphocytic leukemia splits into three types: L1, L2 and L3 according with the occurrence of cytological features and to the degree of heterogeneity in population of leukemia cells. Acute myelogenous leukemia splits into six principal types according with the direction of the differentiation along one or more cellular lines and their maturity degree. AML subtypes M1, M2, and M3 can exhibit a differentiation with granulocytic predominance. M4 presents a granulocytic and monocytic differentiation, M5 presents a monocyte predominant differentiation, and M6 an erythroblastic predominant differentiation [7].

2.3 Acute Leukemia Subtypes Classification

In order to recognize the subtype of each ROI (white globules in our case), we obtain descriptive characteristics that help to discriminate among them. In this work we use three types of characteristics: texture, geometric and statistical (we also obtain eigenvalues from PCA's but we do not consider them as descriptive in the same sense of these). A characteristic of a ROI is a primitive distinctive attribute. Some characteristics are defined by the visual appearance of an image, whereas others result from some specific manipulation.

Our process to obtain descriptive characteristics from our ROI's begins with the conversion of the original region of interest to tones of gray, from which we obtained characteristics (see Table 2) such as Perimeter, Area, Major and Minor Axes, Gray Threshold, Solidity, Total Number of Pixels, and Average among others [12]. Some of these characteristics were obtained using procedures of the Matlab [13] for images library, while others were programmed (also in Matlab). As we mentioned before, we obtained texture characteristics, which are defined according to the interrelation between arrangements of pixels in the image. These are seen in the image as changes in intensity or in tones of gray. We also generated statistical characteristics from the image pixels distribution. Finally we obtained geometric information from the ROI's shape.

Table 2. Some Descriptive Characteristics Obtained from ROI's

Texture Characteristics	Geometric Characteristics	Statistical Characteristics
Threshold of Gray, Maximum value of Gray, Minimal value of Gray ...,	Perimeter, Area, Center, ...	Average, Standard Deviation, Variance, ...

We obtained 30 descriptive characteristics for each region of interest. We now need to find those characteristics that help us to discriminate among the different acute leukemia subtypes in order to achieve good classification accuracy [11]. Looking for other types of characteristics that help us to improve our classification accuracy, we explored the use of principal components analysis to find the eigenvalues to describe our ROI's as described in section 3.2.

3.2 PCA

The aim of Principal Components Analysis is to reduce the dimension of a set of p variables to a set of m of minor size.

The set of Eigen values took as attributes of entry for the classification. The process that we follow to obtain the eigenvalues is described in [8] in contrast that we use the roots of the typical equation (eigenvalues) as attributes. To obtain the Eigen values and eigenvectors solve the characteristic equation [8][9]showed in the equation 2, where I is identity matrix and ρ is the correlation matrix (3) and λ is the questions matrix that will contain the eigenvalues on having solved the equation.

$$| \rho - \lambda I | = 0 \tag{2}$$

$$\rho = \frac{\mathrm{Cov}(x)_{i,j}}{\sqrt{\mathrm{Var}(x)_{i,i}}\ \sqrt{\mathrm{Var}(x)_{j,j}}} \tag{3}$$

We calculate three matrix of correlation, for each image of three channels (RGB), and we obtained eigenvalues and eigenvectors resolving the characteristic equation (2). As descriptive attributes or characteristics we use the eigenvalues, only we select eigenvalues that accumulate the 80% of changeability, in our case were 10 eigenvalues for each band (30 eigenvalues for each ROI) to comply with this condition [10]. Once we obtained Eigen values we did experiments with the different characteristics obtained of ROI's (geometric, texture, statistical, and eigenvalues) to obtain the best possible accuracy to classify the different sub–types of acute leukemia myeloblastic or lymphoblastic.

3 Experimental Results

In this section we describe our classification experiments using the descriptive characteristics obtained from the bone marrow smears using different machine learning algorithms implemented in the WEKA [14] data mining tool. We used a binary classes technique for It is important to mention that our databases presents the class imbalance problem (i.e. we have 38 examples of myeloblastic leukemia subtype M2 and only 10 examples of myeloblastic leukemia subtype M5) between leukemia subtypes and then we included oversampling techniques only for texture, geometric,

and statistical characteristics for the L2 (100% of oversampling to get a total of 60 examples), M3 (100% to get 52 examples), and M5 (300% to get 40 examples).

In our oversampling proces each subtype (only for the L2, M3, and M5 subtypes) we divide the original set in 10 groups and we keep one of those subgroups to validate (or test) the model obtained with the rest of the groups (9 groups), we applied the oversamplig process only for the training phase (with SMOTE we generate new examples).

We performed two classification experiments for the classification of acute leukemia subtypes. In the first experiment we only considered texture, geometric, and statistical characteristics (30 attributes). In the second experiment we only used PCA eigenvalues (also 30 attributes). For our results evaluation we performed a 10-fold cross validation.

3.1 Geometric, Texture, and Statistical Based Classification

For our first experiment, the task consists on classifying acute leukemia subtypes, we have 5 classes; two of them correspond to the lymphoblastic acute leukemia family (L1 and L2) and three to the myeloblastic acute leukemia family (M2, M3, and M5). It is important to mention that in order to do the acute leukemia subtypes classification, we first need to perform a family classification (between lymphoblastic and myeloblastic) and then we do a classification among the leukemia subtypes from each family. Then we created two data groups, the first one containing the L1 and L2 subtypes and the second one with the M2, M3, and M5 subtypes. Finally, for the M1, M2, and M3 subtypes classification we used a two class scheme (we created three data subgroups with a positive and a negative class where the positive class was M2, M3, or M5 respectively for each group). This was the scheme that gave us the best results for the AML subtypes.

The best results were obtained with the IBk, Random Forest Ensemble, AdaBostM IBk ensemble, and the AdaBoostM Bayes Net TAN ensemble algorithms as shown in table 3. For the IBk algorithm we always set k to 3 (we tested with k from 1 to 6 but k = 3 led us to the best results). For all the experiments we ordered the 30 attributes according to their correlation with the class, we did this using a ranker filter of Weka. For the IBk experiment for lymphoblastic leukemia we additionally performed an attributes selection considering attributes: perimeter, distance, area, bounding box, mayor Axis, minor Axis, convex Area, diameter of a circle, entropy, gray threshold, total histogram, max value of histogram, max value of histogram, average, standard deviation, variance. As we can see in table 3, we obtained better accuracy for the myeloblastic leukemia subtypes and this can be a consequence of the high similarity between the L1 and L2 subtypes. We can also notice the use of ensembles improves the classification accuracy. We also found that using oversampling techniques improve the classification accuracy for the M5 leukemia subtype up to 95.38%. The classification accuracy average by family was of 85.2% for acute lymphoblastic leukemia (L1 and L2) and 91.7% for acute myeloblastic leukemia (M2, M3, and M5).

Table 3. Classification Results using Geometric, Texture, and Statistical Characteristics

Acute Leukemia Subtypes	Classifier	Accuracy
L1, L2	RandomCommittee.RandomForest	86.40
	IBk	84.00
M2 and the others M's	AdaBoostM.IBk	91.53
M3 and the others M's	IBk	88.46
M5 and the others M's	AdaBoostM.BayesNet.TAN	95.38

3.2 PCA's Classification

For these experiments we used eigenvalues as attributes, which are already ordered according to the variability that they represent, then it was not necessary to apply an attribute selection algorithm; we only used 10 eigenvalues for each band of the RGB format.

Table 4. Eigenvalues Results Classification

Acute Leukemia Subtypes	Classifier	Accuracy
L1, L2	trees.ADTree	87.20
M2 and the others M's	AdaBoost.IB1	90.76
M3 and the others M's	AdaBoost.KStar	90.00
M5 and the others M's	AdaBoost.SMO	98.46

In this case, the best results were obtained with the Adtree, AdaBoost.IB1, AdaBoost.Ktar and AdaBoost.SMO and Decision Table algorithms. We tried other algorithms but we only show the best results in table 4. As we can see we obtained better accuracy results for the acute myeloblastic leukemia subtypes (as in the previous test) and this suggest that it was not possible to completely differentiate the lymphoblastic leukemia subtypes neither with the geometric-texture-statistical characteristics nor with eigenvalues because of the high similarity between them. The classification accuracy average by family was of 87.2% for acute lymphoblastic leukemia (L1 and L2) and 93.40% for acute myeloblastic leukemia (M2, M3, and M5).

4 Conclusions and Future Work

In this paper we presented a methodology to classify acute leukemia subtypes using different types of characteristics: geometric, texture, statistical, and eigenvalues. This method is based on a morphological and statistical analysis of white globules from bone marrow smears. The implementation of this process can be used as a diagnosis support tool for the identification of acute leukemia subtypes and to provide more accurate treatments to patients.

We would also like to compare our system with others but we have not found any to do it yet. We will continue looking for other systems to do it. What we will do is to compare our results with data from our domain experts. What we are now comparing

is how the use of two different types of characteristics performs in the process of acute leukemia subtypes classification.

We could see that both descriptive characteristics (geometric, texture, and statistical) and eigenvalues (from PCA) produce high accuracy results around 85% for the L1 and L2 subtypes and around 91% for the M2, M3, and M5 subtypes.

The accuracy achieved up to now is enough for a diagnosis support tool but we think it can be improved through the use of new characteristics, enhancing the use of oversampling techniques, and the combination of both types of attributes (geometric, texture, and statistical combined with eigenvalues).

Acknowledgments

I wish to thank Dr. José E. Alonso and Chávez, Dr. Rubén Lobato Tolama, Chemist Laura O. Olvera Oropeza for helping us with the classification of the images of the leukemia database and the Mexican Institute Social Security (IMSS) San Jose in Puebla for providing the data and Blanca Aurora Morales Gonzalez for contributing to the creation of the database. Finally, I especially thank CONACYT for my scholarship Number 202001.

References

1. Bennett J. M. et al. Proposals for the classification of the acute leukemias. French-American-Brithis (FAB) co-operative group", Br J Haematol.; Vol 4, No. 33(1976), 451-458
2. http://www.inegi.gob.mx/inegi/default.aspx
3. Pajares G., M. de la Cruz J., Vision by Computer, AlfaOmega 1st ed, 123
4. Gonzalez R., Woods R., Eddins S., "Digital Image Processing Using MATLAB", Pearson-Prentice Hall,1st ed, 94, 194, 195, 206, 463, 478, 485
5. http://www.desarrolloweb.com/articulos/s1778.php
6. J. B. MacQueen (1967): 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, 1:281-297
7. McDonald G, Paul J. Cruickshank B., "Hematology Atlas, Pan-American", 5th ed, 85
8. Jackson J., "A User's Guide to Principal Components", Wiley-Interscience, 1st ed, 7, 64
9. Jollife I.T., "Principal Component Analysis", Springer, 2nd ed 1, 5, 21
10. Sanei S., Lee T., Cell Recognition Based on PCA and Bayesian Classification, 4th International Symposium on Independent Component Analysis and Blind Signals Separation(ICA2003), Nara Japan (2003)
11. Scotti F., Automatic Morphological Analysis for Acute Leukemia Identification in Peripheral Blood Microscope Images, IEEE International Conference on Computational Intelligence for Measurement Systems And Applications, Italy(2005)
12. Osowski S., Markiewicz T., Marianska B., Moszczyński L., Feature Generation for the Cell Image Recognition of Myelogenous Leukemia, 12th European Signal Processing Conference (EUSIPCO). Vienna(2004)
13. Matlab user manual – Image processing toolbox, MathWorks, Natick, 1999
14. http://www.cs.waikato.ac.nz/ml/weka/