Segmentation and Classification of Leucocyte Images for Detection of Acute Lymphoblastic Leukemia

Ivana Pešić

Abstract—**Acute lymphoblastic leukemia (ALL) is the most common type of leukemia among children. It is characterized by excessive occurrence of immature lymphocytes – lymphoblasts, especially in bone marrow and lymphoid organs. Diagnosing ALL based on blood samples is routinely done by hematologists, often using cytogenetic analysis or immunophenotyping. However, these visual examinations are often slow and limited by subjective interpretation; therefore standardized inexpensive automated systems for detecting lymphoblasts in blood images are needed. Systems for ALL detection are still in the development stage; improving techniques of lymphocyte segmentation from blood images and classification techniques is a challenge. The aim of this paper is to perform a segmentation of lymphocytes on images from the publicly available ALL_IDB2 database and to compare different classification algorithms for ALL detection. Morphological, color and texture feature extraction from segmented lymphocyte images was performed as well as Principal Component Analysis (PCA) for dimensionality reduction. Classification into two groups (healthy lymphocytes and lymphoblasts) was performed using three different algorithms: k-Nearest Neighbours (kNN), Support Vector Machine (SVM) and feedforward neural network (NN).**

Index Terms—**acute lymphoblastic leukemia; lymphocyte segmentation; feature extraction; lymphoblast classification**

I. INTRODUCTION

Leukemia is a blood cancer which is characterized by uncontrolled proliferation of white blood cells in bone marrow. Leukemias can be divided into acute and chronic; according to the type of white blood cells that are affected by malignant transformation, they are divided into myeloblastic, lymphoblastic and monoblastic. Acute lymphoblastic leukemia (ALL) [1] is a disease caused by the proliferation of immature lymphoid cells (lymphoblasts), mostly in bone marrow and lymphoid organs. According to the French-American-British (FAB) classification of ALL [2], ALL is divided into three groups: L1, L2 and L3. L1 is the most common form of ALL among children, L2 is much more common among adults, while L3 is very rare. This disease mostly occurs in children (about 80%), and in adults it occurs usually around the age of 50. It can be fatal if not treated adequately, hence, early diagnosis is of the utmost importance.

Ivana Pešić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: ivanapesic207@ gmail.com).

Routinely, diagnosing of ALL is done manually by a medical expert - hematologist. Disadvantages of such an analysis are slowness, as well as the fact that the results are under subjective influence of the doctor. For these reasons, there was a need to create a fully automated computer-aided system that would perform the detection of lymphoblastic cells from blood images. The advantages of such systems would be speed, affordability and procedure standardization.

The most important step in automated ALL systems is the high accuracy recognition of lymphoblasts in microscopic images of blood. Lymphocytes are cells with regular circular shape and with a compact cytoplasm. In contrast, lymphoblasts have more irregular shape and may contain small cavities in the cytoplasm. Each system for detecting lymphoblasts using blood images should contain 3 main parts: segmentation of lymphocytes from blood images, feature extraction from segmented lymphocytes (from whole lymphocytes or their nuclei) and classification into two groups: lymphoblastic and healthy cells.

Various segmentation and classification methods can be found in literature. There are several attempts to design semi-automated system for detecting ALL. The image database ALL_IDB [3], which consists of two sets of images: ALL_IDB1 (blood images with more than one blood cell) and ALL_IDB2 (cropped areas of individual normal and blast cells that belong to ALL_IDB1) , was created to develop and test such systems by Department of Information Technology - *Università degli Studi di Milano*. F. Scotti et al. [4] proposed a method for lymphocyte segmentation and classification based on the feedforward neural network (NN) with the accuracy of 98.67% on the ALL_IDB1 image database. M. Madhukar et al. [5] used c-mean clustering for leukocyte segmentation from the ALL_IDB1 image database, morphological and texture features were extracted and the Support Vector Machine (SVM) classificatory reached the accuracy of 93.5%. S. Mishra et al. [6] performed image segmentation using watersheds and classification using Random Forest algorithm with the classification accuracy of 99% on the ALL_IDB1. Putz et al. [7] performed a transformation to CMYK color space, determined a threshold value for segmentation using Zack's algorithm, morphological, texture and color features were extracted, and classification was performed using Support Vector Machine with the accuracy of 93% on the ALL_IDB1. V. Singhal and P. Singh [8] used the transformation of images into HSV space, threshold segmentation and classification by Support Vector Machine and the classification accuracy was 89.72% on the ALL_IDB2. Abdeldaim, Ahmed and Talaat et al. [9]

performed a transformation into CMYK color space, segmentation was done using Zack's algorithm, texture, color, and shape features were extracted, and the highest accuracy of 96.01% on the ALL_IDB2 was obtained using k-Nearest Neighbours (kNN). Siew Chin Neoh et al. [10] performed segmentation using clustering, and classification accuracy of 96.67% was obtained using 10-fold cross-validation and Support Vector Machine method on the ALL_IDB2.

In this paper, following steps were performed: segmentation of leukocyte images from the ALL_IDB2 database (because it contains only one blood cell per image) using Otsu's threshold, extraction of 3 types of features (morphological, color and texture), dimensionality reduction using Principal Component Analysis (PCA) and classification using 3 different algorithms: k-Nearest Neighbours, Support Vector Machine and feedforward neural network. The aim of this paper was to segment both whole leukocytes and their nuclei from a publicly available blood image database, to extract different features from segmented cells/nuclei, and finally, to test different classification methods, in order to conclude which method has the greatest potential for further development, as well as which features are the most distinctive.

II. THE METHOD

A. Dataset

Segmentation and classification methods were tested on ALL_IDB2. All images in this database were taken with an optical laboratory microscope coupled with a Canon Power Shot G5 camera. ALL lymphoblast classification was provided by expert oncologists. ALL_IDB2 dataset consists of 260 cropped regions of one healthy leucocyte or lymhoblast; resolution of each image is 257 x 257 pixels. There are 130 lymphoblast images (class 1) and 130 images of normal cells (class 0). Fig. 1 shows some example images from this database.
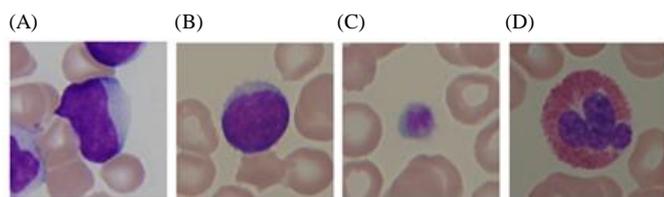
(A)   (B)   (C)   (D)

Fig. 1. Images from ALL_IDB2 database: (A) lymphoblast, (B) healthy lymphocyte, (C) platelet, (D) eosinophil

B. Segmentation

Segmentation method consisted of three steps: preprocessing, threshold-based segmentaton and postprocessing. It is necessary to segment the whole cell (both nucleus and cytoplasm), as well as nucleus only, because parameters of both elements are important in order to properly distinguish lymphoblast cell from healthy one. At the end, only leucocyte images should be selected for further classification.

1. Preprocessing

In the preprocessing step, two images based on the original one were prepared for further processing [13]. For the first one, the original image was converted to a grayscale image and global contrast correction method was applied using following formula:

$$I_p(x,y) = 255 \cdot \frac{(I_o(x,y) - min)}{(max - min)}, \qquad (1)$$

where (x, y) are pixel coordinates, I_p is in output grayscale image, I_o is input grayscale image, min and max are the minimum and maximum gray values of an input image. This image will be used to segment only nucleus from lymphocyte. The second image for further processing is an image of H channel from HSV color space: this image will be used to segment the whole cell, because clear difference can be seen between leucocytes, red blood cells and background.

2. Treshold-based segmentation

In histograms of grayscale image with contrast stretching and H channel image (Fig. 2), three sections can be observed. The part with low levels of gray belongs to the background, the middle part belongs to the remains of cytoplasm in the grayscale image and to the remains of red blood cells in the H channel image and the part with the highest brightness levels belongs to the nucleus in grayscale image and the whole cell in the H channel image. Therefore, segmenting the part of the image with the highest brightness levels should be done. Both in grayscale and H channel image, Otsu's threshold method was used twice in a row in order to achieve this [11]. In the first use of Otsu's threshold method background is eliminated; the second one was performed to remove the remains of the cytoplasm/red blood cells.
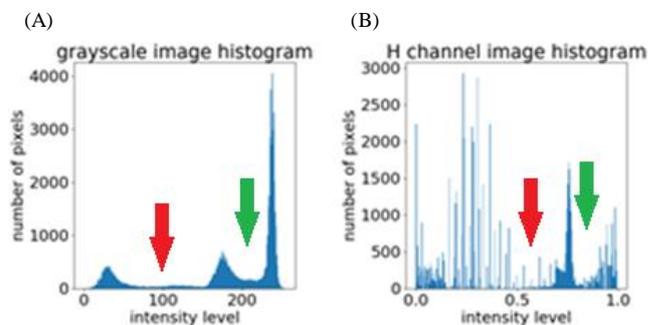
(A)   (B)

Fig. 2. (A) Grayscale image histogram, (B) H channel image histogram; red arrow - first Otsu's threshold, green arrow – second Otsu's threshold

3. Postprocessing

After the segmentation step, there will still be remains of the noise in the segmented images. The binarization of these two segmented images was done to perform morphological opening [11], using a 5 x 5 kernel of circular shape (due to the circular shape of the lymphocytes). Extraction of the largest connected region was performed and then median filtering with a 15 x 15 kernel and, lastly, morphological closing to fill the holes and remove dots.

4. Leucocyte selection

Since the ALL_IDB2 dataset contains images of all types of blood cells, only leucocyte images should be selected to be used in further classification. The following three parameters

[7] were used to determine whether a cell would participate in the classification:

1) Roundness

$$roundness = \frac{4\pi \cdot P}{O^2},\qquad(2)$$

where P is the area and O is the perimeter of the object. If roundness is 1, then the shape is circle. This parameter is relatively insensitive to irregular edges of the object, and experimentally the threshold was set to 0.58.

2) Solidity

$$solidity = \frac{P}{P_c},\qquad(3)$$

where P is the area and P_c is the area of the convex shell of the object. The closer solidity is to 1, the more regular the edge of the object is and it does not have many holes. The threshold is set to 0.899.

3) Area

The area of the segmented object should be at least one-twelfth of the whole image area, in order to avoid selection of platelets which are circular, but very small.

If the value of at least one of these parameters was less than the specified threshold, the image would not participate in further classification. Out of all 260 leucocyte images, 130 lymphoblasts and 76 healthy cells were selected.

C. Feature extraction

After segmenting whole lymphocyte cells and their nucleuses, the next step was extracting features from these segmented areas. Three types of features were extracted: 26 morphological, 10 color and 24 texture features [7].

1. Morphological features: area, perimeter, major axis, minor axis, convex area, convex perimeter, orientation, roundness, solidity, elongation, eccentricity, rectangularity and convexity are calculated for both nucleus and whole segmented cell.
2. Color features: mean value, standard deviation, skewness, kurtosis and entropy are calculated for both nucles and whole cell.
3. Texture features: contrast, dissimilarity, homogeneity, ASM (Angular Second Moment - Energy), root value of ASM and correlation are calculated from GLCM - *Gray Level Co-occurence Matrix* [11]. These features were calculated only for the segmented whole cell for 4 different angles: – 0°, 45°, 90° and 135°.

D. Principal Component Analysis

Dimensionality reduction is performed to eliminate redundant features. One of the most widely used method for this is the Principal Component Analysis method [12] and it was perfomed in this paper. The process for reducing the feature vector X from n to m dimensions is:

1. Estimate the covariance matrix from vector X.

2. Determine the eigenvectors and eigenvalues of the covariance matrix and sort eigenvalues in descending order.
3. Form a transformation matrix A that contains m eigenvectors that correspond to the largest eigenvalues.
4. Calculate transformation

$$Y = A^T X,\qquad(4)$$

where Y is the new feature vector with m dimensions.

E. Classification

Class 0 represents the class of healthy cells, while class 1 represents lymphoblast cells and data is imbalanced (there are 63.1% lymphoblast images and 36.9% healthy cell images); 80% of the images from both classes were used for the traning set and 20% for the test set. The following three algorithms were used and compared:

1. K-Nearest Neighbours (kNN), where 5 neighbours were taken.
2. Support Vector Machine (SVM) with Gaussian kernel
3. Feedforward neural network with hyperaramter optimization – the goal was to find hyperparameters that maximize a certain criterion, in this case F1 score. Hyperparameters for optimization were: number of nodes in hidden layers, activation functions and class weights, where 20% of the training set was used as a set for validation.

III. RESULTS AND DISCUSSION

A. Segmentation results

Fig. 3 shows a characteristic example of the original image of a single lymphoblast from ALL_IDB2. Fig. 4 shows the H-channel of the original image used for segmenting the whole cell, the results after segmentation based on the first Otsu threshold, the binary image after segmentation based on the second Otsu threshold and finally, the segmented whole cell after morphological processing. Fig. 5 shows the grayscale image of the original image, grayscale image with corrected contrast, the results of its two segmentations and the final segmented nucleus. Two additional examples of segmentation results from this dataset can be seen in Fig. 6.
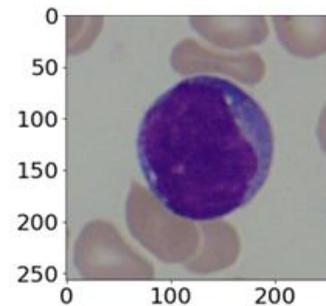

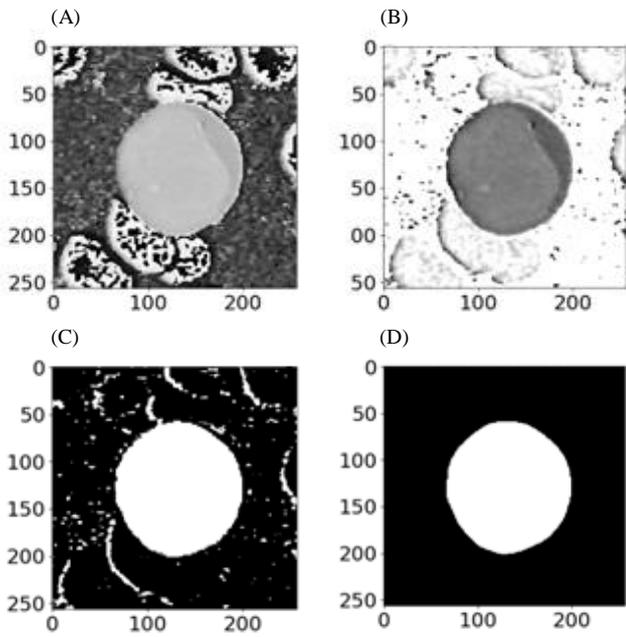Fig. 3. Original image of one cell from ALL_IDB2

Fig. 4. Steps for segmenting the whole cell: (A) H-channel of the original image, (B) result after first Otsu thresholding, (C) result after second Otsu thresholding (D) segmented whole cell
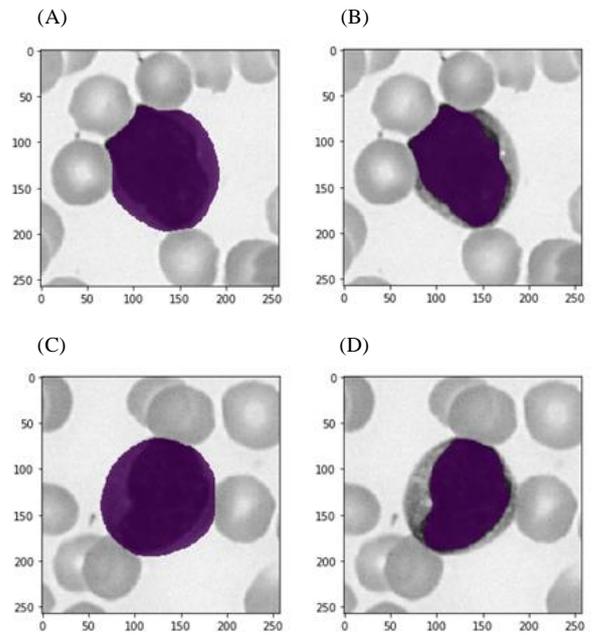

Fig. 6. Two examples of (A), (C) segmented cells and (B), (D) segmented nuclei

From each of the selected 206 images 60 features were extracted, and then dimensionality reduction was performed using the PCA method. It was decided to reduce the dimensions from 60 to 15; the contribution of each eigenvalue can be seen in Fig. 7. Also, dimensionality reduction was performed again for all features except morphological, because morphological features are the most sensitive to segmentation. In this case the dimensions were reduced from 34 to 12 (Fig. 8).
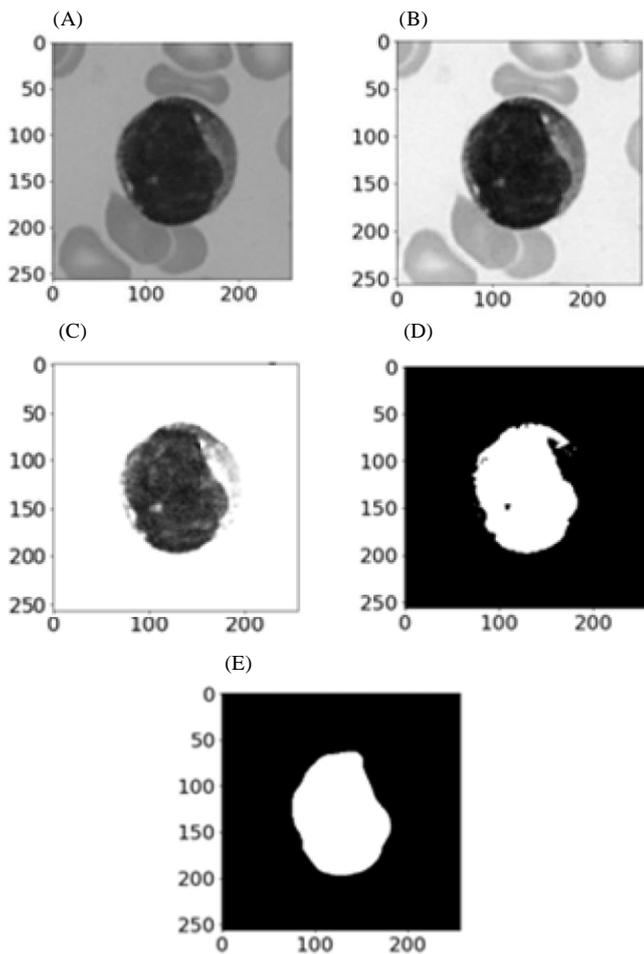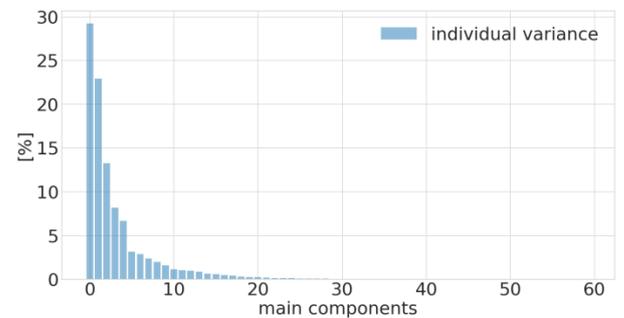

Fig. 7. Eigenvalues for all features

Fig. 5. Steps for segmenting the nucleus: (A) grayscale image of the original image, (B) grayscale image with contrast stretching, (C) result after first Otsu thresholding, (D) result after second Otsu thresholding (E) segmented nucleus
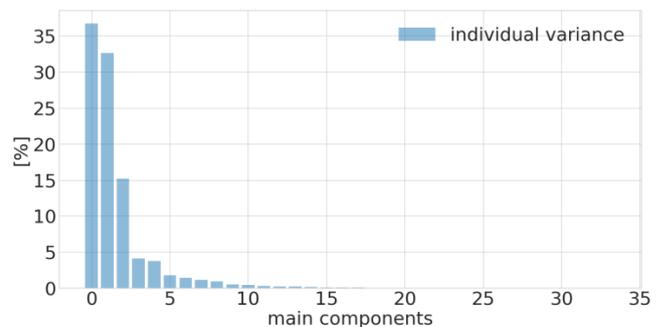

Fig. 8. Eigenvalues for texture and color features

B. Classification results

Based on 20 consecutive classification results, mean values and standard deviations for precision, sensitivity and classification accuracy were calculated for random train and test splits. Four combinations of features were used: 1) all 60 features, 2) 15 features when PCA was applied to all 60 features, 3) 34 color and texture features and 4) 12 features when these features were reduced. The following table shows the results for 3 different classification algorithms – k-Nearest Neighbours, Support Vector Machine, and feedforward neural network.

TABLE I
CLASSIFICATION RESULTS

		Precision [%]		Sensitivity [%]		Accuracy [%]
		Class 0	Class 1	Class 0	Class 1	
kNN	all feat.	60.45±10.72	95.52±3.90	84.28±10.85	74.45±3.86	**75.85±4.78**
	all feat. + PCA	75.53±10.44	92.37±4.92	85.24±8.50	87.41±4.79	**86.35±4.82**
	color and texture feat.	67.76±10.76	93.56±4.57	85.89±9.03	83.79±4.89	**83.99±4.94**
	color, texture + PCA	84.40±8.73	95.36±3.86	91.42±6.74	91.85±4.24	**91.45±4.00**
SVM	all feat.	72.49±10.15	96.40±3.97	92.58±7.40	86.66±5.12	**87.86±4.44**
	all feat. + PCA	81.33±9.71	94.65±4.22	89.96±7.26	90.36±4.57	**89.89±4.20**
	color and texture feat.	81.28±9.11	94.78±4.26	90.25±7.24	90.38±4.70	**89.96±4.18**
	color, texture + PCA	87.00±7.74	95.30±3.88	92.31±6.50	93.12±3.87	**92.53±3.60**
NN	all feat.	77.67±4.71	95.04±4.14	83.59±5.97	91.50±3.95	**88.76±4.66**
	all feat. + PCA	85.83±9.52	93.33±5.24	81.75±7.63	95.40±5.64	**90.65±5.76**
	color and texture feat.	91.67±7.66	90.26±4.44	86.90±3.98	95.25±4.6	**91.90±3.19**
	color, texture + PCA	95.00±6.45	94.02±1.79	90.80±2.56	97.28±3.38	**95.33±2.04**

C. Discussion

It can be seen from the Table I that the highest classification accuracy is achieved using a feedforward neural network when only color and texture features are used with PCA - from 34 dimensions to 12. The best accuracy is 95.33%. A significant disadvantage of this method is the long execution time (about 240 s). This method, on the other hand, has the most opportunities for further improvement out of all 3 used algorithms. When using color and texture features with PCA, the results of classification using kNN and SVM algorithms are also satisfactory - 91.45% and 92.63%, respectively. It can be concluded that this selection of features used with PCA shows the best results when using images from the ALL_IDB2 database. Compared to the reference papers, accuracy of 95.33% was higher than in [8], where it is achieved 89.72%, and slightly lower than in [9] and [10], where it is 96.01% and 96.67%, respectively.

IV. CONCLUSION

This paper presents steps for ALL detection from blood images: segmentation, feature extraction, dimensionality reduction and classification. Different classification algorithms were tried out to determine which one should be developed further. Future work could focus on improving automatic segmentation, developing one of the classification methods and achieving better accuracy. Less sensitivity to initial data partition into train and test sets could be achieved using k-fold crossvalidation when classification using neural networks is performed. Finally, testing on larger image datasets is necessary to decide which features are the most distinctive in general case and so that progress in future research work could be made.

REFERENCES

[1] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, C. Sultan, "Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group", *Br. J. Haematol.*, Vol. 4, No. 33, pp. 451-458, August, 1976.

[2] T. Terwilliger, M. Abdul-Hay, "Acute lymphoblastic leukemia: a comprehensive review and 2017 update", *Blood Cancer J.*, Vol. 7, Jun, 2017.

[3] R. D. Labati, V. Piuri, and F. Scotti, "All-IDB: the acute lymphoblastic leukemia image database for image processing" , Proc. of the 18th IEEE International Conference on Image Processing, Brussels, Belgium, pp. 2045–2048, September 11-14, 2011.

[4] F. Scotti, "Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images", Proc. of the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, pp. 96–101, Messian, Italy, July 20-22, 2005.

[5] M. Madhukar, S. Agaian, A. T. Chronopoulos, "New decision support tool for acute lymphoblastic leukemia classification", Proc. Of SPIE – The International Society for Optical Engineering, Burlingame, California, USA, Vol. 8295, February, 2012

[6] S. Mishra, B. Majhi, P. K. Sa, L. Sharma, "Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection", *Biomed. Signal Process. Control.*, Vol. 33, pp. 272–280, March, 2017.

[7]   L. Putzu, G. Caocci, C. Di Ruberto, "Leucocyte classification for leukaemia detection using image processing techniques", in *Artif. Intell. Med.,* Vol. 62, pp 179-191, September, 2014.

[8]   V. Singhal, P. Singh, "Local Binary Pattern for automatic detection of Acute Lymphoblastic Leukemia*",* Twentieth National Conference on Communications, Kanpur, India, pp. 1-5, February, 2014.

[9]   A. Abdeldaim, A. Talaat, M. Elhoseny, A. E. Hassanien, "Computer-Aided Acute Lymphoblastic Leukemia Diagnosis System Based on Image Analysis", *Studies Comp. Intell.*, pp. 131-147, January, 2018.

[10] S. Neoh, W. Srisukkham, L. Zhang, S. Todryk, B. Greystoke, C. P. Lim, M. A. Hossain, N. Aslam,"An Intelligent Decision Support System for Leukaemia Diagnosis using Microscopic Blood Images**",** *Scientific Reports*, Vol. 5, October, 2015.

[11] R. C. Gonzalez, R. E. Woods, *Digital Image Processing,* 3rd ed., Prentice Hall, Upper Saddle River, NJ, 2008.

[12] Ž. Đurović, Pattern Recognition OS4PO, University of Belgrade - School of Electrical Engineering, 2019. http://automatika.etf.bg.ac.rs/sr/prepoznavanje-oblika-os4po/217-serbian/sadr%C5%BEaj/predmeti/izborni-kursevi-za-os/obrada-signala/prepoznavanje-oblika-os4po/427-os4po-materijali

[13] Y. Li, R. Zhu, L. Mi, Y. Cao, D. Yao, "Segmentation of White Blood Cells from Acute Lymphoblastic Leukemia Images Using Dual-Threshold Method", *Comput. Math. Method. M.*, Vol. 2016, pp. 1-12, January, 2016.