Original Article

Enhanced Recognition of Acute Lymphoblastic Leukemia Cells in Microscopic Images based on Feature Reduction using Principle Component Analysis

Morteza MoradiAmin[1], Nasser Samadzadehaghdam[1,*], Saeed Kermani[2], and Ardeshir Talebi[3]

1- Department of Medical Physics and Biomedical Engineering, School of Medicine, Tehran University of Medical Sciences (TUMS), Tehran, Iran.

2- Department of Biomedical Engineering, Faculty of Advanced Medical Technology, Isfahan University of Medical Sciences, Isfahan, Iran.

3- Department of Pathology, Faculty of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran.

Received: 28 April 2015

Accepted: 1 September 2015

ABSTRACT

Purpose- Acute lymphoblastic leukemia (ALL) is the most common form of pediatric cancer of white blood cells which is categorized into three types of L1, L2, and L3. It is usually detected through screening of blood and bone marrow smears by pathologists. Since manual detection is time-consuming and boring, computer-based systems are preferred for convenient detection. The rigorous similarity between morphology of ALL types and that of normal, reactive and atypical lymphocytes, makes the automatic recognition a challenging problem. In this paper, we tried to improve the sensitivity of detection based on principle component analysis (PCA).

Methods- After segmenting cell nucleus using fuzzy c-means clustering algorithm, several geometric and statistical features are extracted. Then the feature space dimensionality is reduced based on (PCA). The first 8 components of the feature space are applied to support vector machine (SVM) classifier. Then the cancerous and lymphocyte cells are classified into their subtypes.

Results- For evaluating the proposed method, we used an expert pathologist's classification as a reference. Classification was evaluated by three parameters: sensitivity, specificity and accuracy. A comparison with our previous work showed that using dimensionality reduced feature space based on PCA, instead of using individually selected features, improved the average sensitivity and precision of classification more than 10%.

Conclusion- The results show that proposed algorithm performs better than our previous work. Its acceptable performance for the diagnosis of ALL and its sub-types as well as other lymphocyte types makes it an assistant diagnostic tool for pathologists.

Key words:
Acute lymphoblastic leukemia,
Segmentation,
Fuzzy c-means,
PCA,
SVM.

1. Introduction

Leukemia is one of the most common cancers in the world, which involves bone marrow and blood, and effects within the body with duplication of a large number of abnormal white blood cells. The French-American-British (FAB) classification categorizes ALL into three L1, L2 and L3 morphological subtypes. The first step to identify this type of leukemia is the observation of blast cells in the peripheral blood smear or increase in the smear of bone marrow by pathologists. The most

**** Corresponding Author:***
Nasser Samadzadehaghdam, MSc
Department of Medical Physics and Biomedical Engineering, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran.
Tel: (+98) 9141102154
E-mail: nsamadzadeh_a@yahoo.com

challenging parts of this procedure are being time-consuming, tedious and laborious for pathologists and the diagnosis is dependent on the pathologist's experience. In order to overcome these problems, many researchers have noticed automatic systems for ALL detection from microscopic images. However, high similarity between morphology of ALL (L1, L2, L3) and lymphocyte subtypes (normal, reactive and atypical) is a big challenge in designing automatic systems. Figure 1 shows a sample of ALL and lymphocyte subtypes.

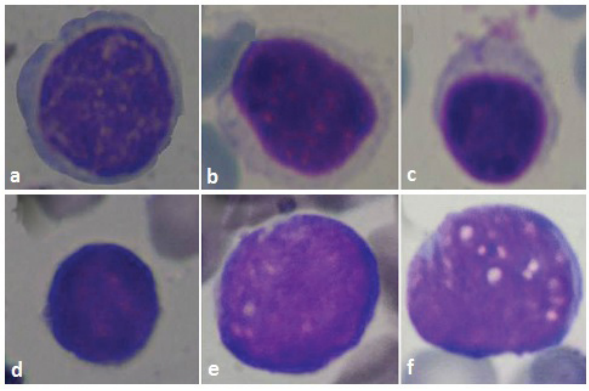

Figure 1. Sample cell images. a) atypical; b) reactive; c) normal; d) L1; e) L2; f) L3.

Sinha *et al*. [1] for WBC segmentation in color images of stained peripheral blood smears, proposed a two-step process in HSV color space using fuzzy k-means clustering followed by the EM-algorithm. After extracting features from the segmented cytoplasm and nucleus, various classifiers have been explored on different combinations of feature sets. Theera-Umpon *et al*. [2] proposed an automatic algorithm for segmentation of nucleus and cytoplasm of bone marrow WBC based on the fuzzy c-means clustering algorithm and basic image processing morphology operations. For the evaluation of their proposed technique, they compared their results with the cell images manually segmented by an expert. Scotti *et al*. [3] proposed a method to enhance the microscope images by removing the undesired background, a method for the robust estimation of the mean cell diameter and an adaptive segmentation strategy to robustly identify white cells permitting. Finally, the features of white cells were extracted for subsequent automatic diagnosis of blood diseases. Wang *et al*. [4] proposed a cell detection method that utilizes both intensity and shape information of cell to improve the segmentation. Meanwhile, they proposed an online support vector classifier (OSVC), which features the removal of support vectors from the old model and assigning the new training examples with different weights according to their importance. Theera-Umpon *et al*. [5] investigated whether information about the nucleus alone is adequate to classify white blood cells. Features based on the morphological granulometries were extracted from each segmented blood cell's nucleus. Bayes' classifiers and artificial neural networks were applied as classifiers. The results show that the features using nucleus alone can be utilized to achieve a classification rate of 77% on the test sets. Madhloom *et al*. [6] focused on white blood cell segmentation using a combination of automatic contrast stretching supported by image arithmetic operation, minimum filter and global threshold techniques. They achieved an accuracy between 85-98%. Halim *et al*. [7] used segmentation on HSV color space in order to eliminate the white blood cells (WBC) from the background. In order to handle the overlapping cells, they used the erosion morphological operator. Their method provided the highest average accuracy of 97.8% for counting both ALL and AML cases. For segmentation of Acute Myeloid Leukemia cells, Lim *et al*. [8] proposed a method consisted of gradient magnitude, thresholding, morphological operations and watershed transform. They reported a segmentation accuracy of 94.5% for 50 tested images while the average accuracies for M_2, M_5 and M_6 subtypes were 94.58%, 95.06% and 95.65% respectively. Mohapatra *et al*. [9] proposed a quantitative microscopic approach toward the discrimination of lymphoblasts (malignant) from lymphocytes (normal) in stained blood smear and bone marrow samples. The identification and segmentation of WBCs realized through image clustering followed by the extraction of different types of features, such as shape, contour, fractal, texture, color and Fourier descriptors. Finally, an ensemble of classifiers is trained to recognize ALL. The results of this method were good, but they were obtained by using a proprietary dataset, so the reproducibility of the experiment and comparisons with other methods are not possible. Abbas *et al*. [10] segmented the Nuclei of Leukocytes by image processing methods such as OTSU

global thresholding and morphological operation dilation. They worked on 380 microscopic images and obtained a segmentation accuracy of 96.5%. In our previous work [11], after applying image preprocessing step, cells nuclei were segmented by k-means algorithm. From the segmented nuclei 77 statistical and geometric features were extracted. Features which resulted in high sensitivity, specificity and accuracy in the classification step were selected as the inputs of SVM classifier with 10-fold cross validation to classify the cells as cancerous and non-cancerous. These cells were also classified into their sub-types by Multi-SVM classifier.

The steps of the proposed algorithm in this paper are like [11] except that we replaced the nuclei segmentation algorithm, i.e. k-means, with fuzzy c-means because the former sometimes produces empty cluster while running. Another modification is related to the feature selection step. Here we used the first 8 principle components of the feature space which were obtained by PCA dimensionality reduction algorithm.

The rest of the paper is organized as follows: in section 2 a brief description of the database is provided. Section 3 covers the proposed method and finally the results and discussion are provided in section 4.

2. Materials and Methods

2.1. Database

Database of this study included 21 peripheral blood smear and bone marrow slides of 14 patients with ALL and 7 normal persons. The slides were acquired at Isfahan Al-Zahra and Omid hospital pathology laboratories and prepared and stained using giemsa staining for visualization of cell components. The acquired images were digitized by a Nikon1 V1, high resolution digital camera coupled to Nikon Eclipse 50i light microscope under 100X power objective oil immersed setting and with an effective magnification of 1000. The format of the images were JPEG at the maximum resolution of the camera i.e. 2592 × 3872 pixels in RGB color space. The captured images were revised by the hematologist to determine the true type of the blood cell. In this research, 312 digital images have been acquired including 146 images of ALL sub-types (L1, L2, L3) and 166 images of lymphocyte cells (normal, reactive and atypical). A total number of 958 cells were obtained. The data set consisted of six classes of white blood cells—L1, L2, L3, normal, reactive and atypical—with the numbers of 277, 215, 151, 50, 94 and 171cells, respectively. It should be noted that to speed up the performance of the proposed algorithm, the resolution of the images was reduced by a factor of five i.e. 519 × 775.

2.2. Proposed Algorithm

The overall steps of the proposed algorithm are shown in the block diagram of Figure 2.

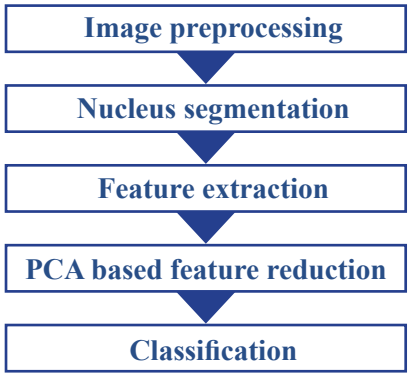

Figure 2. Block diagram of the proposed method.

2.2.1. Image Preprocessing

There is blurriness and the effects of unwanted noise on blood leukemia images, e.g. exposure of the microscope which influences the quality of the captured images, may result in false diagnosis [12]. Therefore, an image pre-processing such as image enhancement techniques are needed to improve this situation.

First, the image is converted from RGB color space to HSV. This reduces correlation between the color channels (compared to RGB) and enables dealing with three H, S and V channels separately. In HSV color space, color information is embedded in H and S channels while V component corresponds directly to the concept of intensity and matches the human perception of lightness. Then the popular histogram equalization technique is applied on V band for equalizing the gray level of image intensities. Histogram equalization reduces the

effects of different lighting conditions in different image acquisition sessions [13] so all the images will have approximately the same brightness. In Figure 3 a sample of histogram equalization is given.

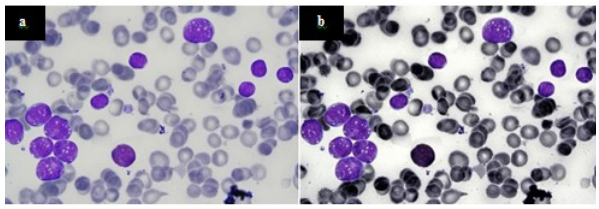

Figure 3. Result of preprocessing step. (a) Original image. (b) Enhanced image.

2.2.2. Nucleus Segmentation

Segmentation plays a key role since it will directly affect subsequent processing that is namely feature extraction and classification. In our previous work, for the segmentation of nuclei we used k-means algorithm, but this algorithm sometimes creates empty cluster [11]. To prevent this, we preferred here fuzzy c-means clustering method.

Fuzzy c-means was first proposed by Bezdek *et al*. [14]. The algorithm returns values between 0 and 1 called the partition matrix, which represent the degree of membership between each data and centers of clusters. It is based on minimization of the objective function J_m:

$$J_m(U,C) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \| X_k - C_i \|^2 \quad (1)$$

where m is a real number greater than one. Sets X_1, X_2, ..., X_n and C_1, C_2,..., C_c are data sample vectors and cluster centers, respectively. U is the partition matrix and u_{ik} is the ith cluster membership value of kth input sample X_k. Optimization of the objective function shown above is done with the updating of the membership u_{ik}, and cluster center C_i

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left\{\frac{\| X_k - C_i \|}{\| X_k - C_j \|}\right\}^{\frac{2}{m-1}}} \quad (2)$$

$$C_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} X_k}{\sum_{k=1}^{n} u_{ik}^{m}} \quad (3)$$

The fuzzy c-means algorithm with 4 clusters was applied on the 3D image in HSV color space. The clusters correspond to nucleus, background, and other cell parts (e.g., erythrocytes, plackets and WBC cytoplasm). Figure 4 shows four clusters of a sample image after applying fuzzy c-means.

It was observed that the cluster with minimum red color is the cluster related to nuclei. So, the mean value of R channel was calculated for each cluster and the cluster with minimum value was considered as a cluster of nuclei i.e. the rightmost image in Figure 4.

We had to do a two-step post processing on the nuclei cluster image: first in order to remove stain artifacts from the nuclei cluster and fill some small holes in the nuclei, we performed binary morphological opening and closing operation on the image. The size of the structuring element for morphological operators must be smaller than the minimum size of a nucleus that will be determined and big enough to eliminate the stain artifacts areas. Second, to separate the connected nuclei we applied watershed algorithm. This algorithm is able to detect the boundary lines between the connected cells [15] and successfully separated all connected nuclei in the image into its individual nuclei. Figure 5 shows a sample nuclei cluster and its post-processed version.

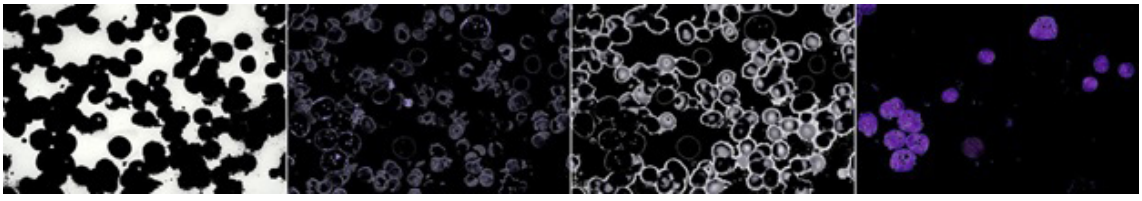

Figure 4. Result of fuzzy c-means clustering.

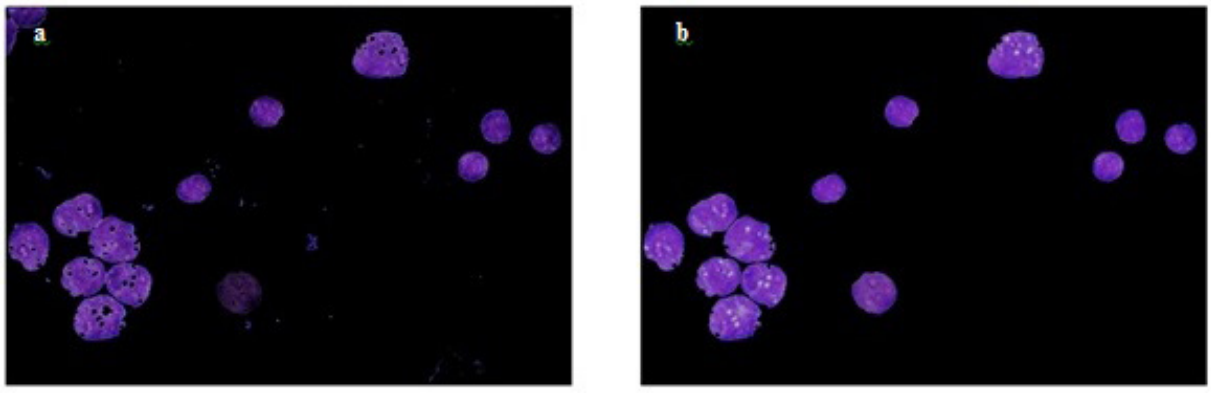

Figure 5. Left: cluster of nuclei, right: extracted nuclei.

2.2.3. Feature Extraction

In order to classify the cells as cancerous or non-cancerous and determine their sub-type i.e. L1, L2, L3, atypical, reactive and normal, several features must be extracted from the nuclei cluster. Here we used both geometric and statistical features. Geometric features provide information about the size and shape of a nucleus while statistical features give information about gray scale image histogram of the pixels located in a nucleus. According to hematologists, the geometric of the nucleus is one of the essential features which can be used for characterization of the cells. The used geometrical features include: area, perimeter, solidity, eccentricity and extent of nucleus from the binary image of nucleus. Statistical features give information about the distribution of intensities in an image. These features are generated from the gray scale image histograms of the red, green and blue, as well as the hue, saturation and value channels from original and enhanced image of nucleus and include measures such as: mean, standard deviation, energy, entropy, skewness and kurtosis. 72 statistical features have been created by this way.

2.2.4. PCA-based Feature Reduction

The important idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set [16]. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first *few* retain most of the variation present in *all* of the original variables [16].

6 general steps for performing a principal component analysis include [17]:

1. Put all the *d*-dimensional features in a matrix. The dimension of features in this study is 77 and total number of data, i.e. cells, equals 958. So we will have a feature matrix of size 77 × 958.
2. For PCA to work properly, we have to subtract the mean from each of the features. To do so we compute the mean of each row of the feature matrix and then subtract it from every element of that row.
3. Compute the scatter matrix (alternatively, the covariance matrix) of the whole data set.
4. Calculate the eigenvectors and eigenvalues of the covariance matrix.
5. Sort the eigenvectors by decreasing eigenvalues and choose *k* eigenvectors with the largest eigenvalues to form a *d* x *k* dimensional matrix *W* (where every column represents an eigenvector).
6. Use this *d* x *k* eigenvector matrix to transform the samples onto the new subspace.

After applying the above steps, we chose the first 8 components as the classifiers inputs.

2.2.5. SVM Classifier

Since the patterns are very close in the feature space, support vector machine (SVM) is a suitable choice for classification [18]. It is a powerful tool for data classification based on hyper plane

classifier. To classify cells as cancerous or non-cancerous we used traditional SVM which is in fact a binary classifier and for detection of cell subtype we used a multiclass SVM classifier. This classification is achieved by a separating surface in the input space of the data set using by different kernel functions as linear or non-linear such as quadratic, polynomials and radial basis functions (RBF) [19]. Based on optimum accuracy of separation RBF kernel with sigma 3 was used in this paper. For the evaluation of the classifier, the k-fold cross validation method with k=10 was applied.

3. Results

Results of classification in three images are shown in Figure 6 and 7.

Confusion matrices of the binary SVM and Multi-SVM were obtained. They are provided in Table 1 and 2, respectively.

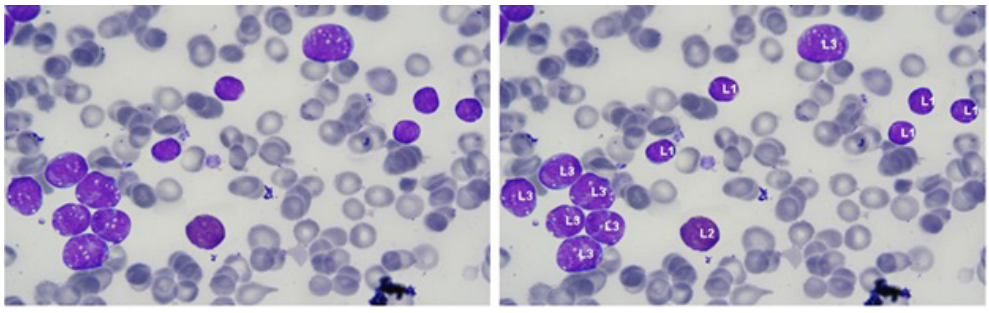

Figure 6. Result of classification. Left: original image containing cancerous cells, Right: labeled cancerous cells (L1, L2 and L3).

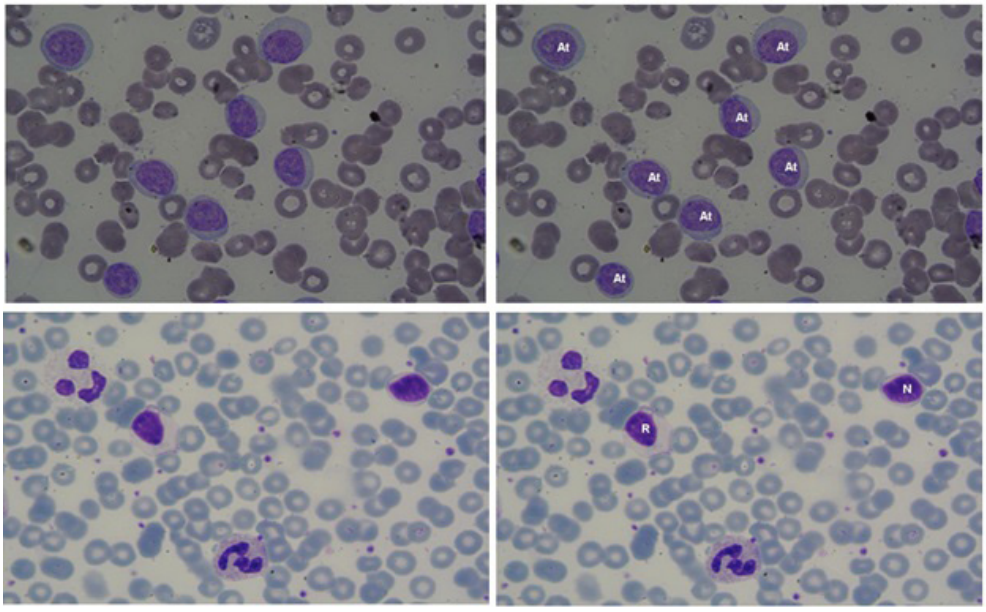

Figure 7. Results of classification. Left: original images containing non-cancerous cells, Right: labeled non-cancerous cells (At: atypical, N: normal and R: reactive).

Table 1. Cancerous and noncancerous cells versus result of binary SVM classifier.

		Binary SVM output	
		Cancerous	Non- Cancerous
Binary SVM input	Cancerous	630	13
	Non- Cancerous	11	304

Table 2. L1, L2, L3, atypical, normal, and reactive cells versus result of multi-SVM classifier.

		Multi- SVM output					
		L1	L2	L3	Atypical	Normal	Reactive
Multi-SVM input	L1	253	18	2	1	2	1
	L2	17	185	3	10	0	1
	L3	1	4	146	0	0	0
	Atypical	0	7	2	162	0	0
	Normal	2	0	0	0	39	9
	Reactive	1	2	0	1	10	80

From these matrices the performance of the classifiers was evaluated by the statistical parameters including sensitivity, specificity, accuracy, precision and false negative rate.

As mentioned above, the input of the classifiers are dimension-reduced feature space. We compared the classification results with our previous method of using individually selected features with high performance as inputs of the classifiers [11]. The comparative results are provided in table 3 through 5. According to the values of sensitivity, specificity, accuracy, precision and false negative rate in table 3, there is little improvement in the performance of the binary classifier using PCA-based dimension reduced features. According to table 4 a similar situation also exists for multiclass SVM in classifying ALL subtypes. But as table 5 shows, sensitivities of recognition of lymphocytes, namely normal and reactive cells, have improved significantly. In comparison to our previous method [11], the average sensitivity of recognition of both normal and reactive lymphocytes has increased 12%. The average precision of recognition of normal and reactive lymphocytes has increased 15% and 9%, respectively.

Table 3. Performance of the binary classifiers using selected features vs. dimension reduced features.

Statistical parameters	Selected features	Dimension reduced features
Sensitivity	98%	98%
Specificity	95%	97%
Accuracy	97%	98%
Precision	98%	98%
False negative rate	2%	2%

Table 4. Performance of the Multi-SVM classifiers using selected features vs. dimension reduced features.

	L1		L2		L3	
Statistical parameters	**selected features**	**dimension reduced features**	**selected features**	**dimension reduced features**	**selected features**	**dimension reduced features**
Sensitivity	91%	92%	84%	86%	97%	97%
Specificity	97%	97%	95%	96%	99%	99%
Accuracy	95%	96%	92%	94%	99%	99%
Precision	92%	92%	84%	86%	95%	95%
False negative rate	9%	8%	16%	14%	3%	3%

Table 5. Performance of the Multi-SVM classifiers using selected features vs. dimension reduced features.

	Atypical		Normal		Reactive	
Statistical parameters	**selected features**	**dimension reduced features**	**selected features**	**dimension reduced features**	**selected features**	**dimension reduced features**
Sensitivity	95%	95%	***66%***	***78%***	***73%***	***85%***
Specificity	98%	99%	97%	99%	98%	99%
Accuracy	97%	98%	96%	98%	95%	98%
Precision	93%	93%	***61%***	***76%***	***79%***	***88%***
False negative rate	5%	5%	34%	22%	27%	15%

4. Discussion

In this paper, an enhanced computer-based method is proposed for the classification of cancerous and non-cancerous cells as well as their subtypes, just by using PCA based features extracted from the image of their nucleus. As results show, the proposed algorithm can be used as an assistant diagnostic tool for pathologists. Besides, the clinical impact of this research is that it will provide the ability to pathologists to examine a blood smear for finding cancerous cells. Apart from detecting cancerous cells subtype the algorithm can also differentiate non-cancerous cells subtype with improved sensitivity.

The main contribution of this study is improving our previously proposed method of feature classification. This was achieved by implementing an additional step of dimension reduction of feature space using PCA.

It can be considered that one problem we encountered while testing of our method was the absence of publicly available datasets. In fact, many authors tested their system with only a few sample images, or with their own datasets, which are not publicly available. Thus, we could not directly compare our findings with the results obtained by various proposed systems, limiting the reproducibility of the innovations proposed by similar systems. Moreover, as there is no work on sub-types ALL detection yet, this aim can be considered as the contribution of this study.

For future works, authors believe that in addition to nucleus, the segmentation of cytoplasm and extraction its features can also improve the performance of this computer-based system.

Conflict of Interest

The authors confirm that there is no conflict of interest.

References

1- N. Sinha and A. Ramakrishnan, "Automation of differential blood count," in *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*, pp. 547-551, 2003.

2- N. Theera-Umpon, "White blood cell segmentation and classification in microscopic bone marrow images," in *Fuzzy systems and knowledge discovery*, ed: Springer, pp. 787-796, 2005.

3- F. Scotti, "Robust segmentation and measurements techniques of white cells in blood microscope images," in *Instrumentation and Measurement Technology Conference, 2006. IMTC 2006. Proceedings of the IEEE*, pp. 43-48, 2006.

4- M. Wang, X. Zhou, F. Li, J. Huckins, R. W. King, and S. T. Wong, "Novel cell segmentation and online learning algorithms for cell phase identification in automated time-lapse microscopy," in *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pp. 65-68, 2007.

5- N. Theera-Umpon and S. Dhompongsa, "Morphological granulometric features of nucleus in automatic bone marrow white blood cell classification," *Information Technology in Biomedicine, IEEE Transactions on,* vol. 11, pp. 353-359, 2007.

6- H. Madhloom, S. Kareem, H. Ariffin, A. Zaidan, H. Alanazi, and B. Zaidan, "An automated white blood cell nucleus localization and segmentation using image arithmetic and automatic threshold," 2010.

7- N. H. A. Halim, M. Y. Mashor, and R. Hassan,

"Automatic Blasts Counting for Acute Leukemia Based on Blood Samples," *International Journal of Research & Reviews in Computer Science,* vol. 2, 2011.

8- H. N. Lim, M. Y. Mashor, and R. Hassan, "White blood cell segmentation for acute leukemia bone marrow images," in *Biomedical Engineering (ICoBE), 2012 International Conference on*, pp. 357-361, 2012.

9- S. Mohapatra, D. Patra, and S. Satpathy, "An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images," *Neural Computing and Applications,* vol. 24, pp. 1887-1904, 2014.

10- N. Abbas and D. Mohamad, "Automatic Color Nuclei Segmentation of Leukocytes for Acute Leukemia," *Research Journal of Applied Sciences, Engineering and Technology,* vol. 7, pp. 2987-2993, 2014.

11- M. M. Amin, S. Kermani, A. Talebi, and M. G. Oghli, "Recognition of Acute Lymphoblastic Leukemia Cells in Microscopic Images Using K-Means Clustering and Support Vector Machine Classifier," *Journal of medical signals and sensors,* vol. 5, p. 49, 2015.

12- N. Mokhtar, H. Nor Hazlyna, M. Yusoff, H. Roseline, M. Nazahah, R. Adollah*, et al.*, "Image enhancement techniques using local, global, bright, dark and partial contrast stretching for acute leukemia images," 2009.

13- K. Rodenacker and E. Bengtsson, "A feature set for cytometry on digitized microscopic images," *Analytical Cellular Pathology,* vol. 25, pp. 1-36, 2003.

14- J. Keller, R. Krisnapuram, and N. R. Pal, Fuzzy models and algorithms for pattern recognition and image processing: *Springer Science & Business Media,* vol. 4, 2005.

15- H. Khajehpour, A. M. Dehnavi, H. Taghizad, E. Khajehpour, and M. Naeemabadi, "Detection and segmentation of erythrocytes in blood smear images using a line operator and watershed algorithm," *Journal of medical signals and sensors,* vol. 3, p. 164, 2013.

16- I. Jolliffe, *Principal component analysis*: Wiley Online Library, 2002.

17- L. I. Smith, "A tutorial on principal components analysis," *Cornell University, USA,* vol. 51, p. 52, 2002.

18- S. Mohapatra and D. Patra, "Automated cell nucleus segmentation and acute leukemia detection in blood microscopic images," in *Systems in Medicine and Biology (ICSMB), 2010 International Conference on*, pp. 49-54, 2010.

19- A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing,* vol. 14, pp. 199-222, 2004.