Automatic recognition of acute lymphoblastic leukemia using multi-SVM classifier

Pouria Mirmohammadi[1], Amirhossien Rasooli[2], Meghdad Ashtiyani[1], Morteza Moradi Amin[2] and Mohammad Reza Deevband[1,*]

[1]Department of Biomedical Engineering and Medical Physics, Faculty of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Department of Medical Physics and Biomedical Engineering, Tehran University of Medical Sciences, Tehran, Iran

Acute lymphoblastic leukemia (ALL) is the most popular form of white blood cells cancer in children. It is classified into three forms of L1, L2 and L3. Typically, it is identified through screening of blood smearsvia pathologists. Since this is laborious and tedious, automatic systems are desired for suitable detection; but the high similarity between morphology of ALL forms and that of normal, reactive and a typical lymphocytes, makes the automatic detection a challenging problem. This study tried to improve the accuracy of detection based on principle component analysis (PCA). After segmenting nuclei of cells, numerous features were extracted. The first six components of this feature space were used for the binary and multiclass support vector machine classifiers. An expert pathologist was used to appraise this method as a gold standard. A collation with similar work indicated that using PCA instead of using exclusively selected features enhanced the average sensitivity and specificity of classification up to 10%. The results demonstrate that this algorithm performs better than similar studies. Its permissible efficiency for identifying ALL and its sub-types as well as other lymphocyte forms makes it an associate diagnostic device for pathologists.

Keywords: ALL, fuzzy c-means method, PCA analysis, SVM classifier.

LEUKEMIA is one of the most fatal cancer diseases in the world. It affects both the blood and bone marrow, with repetition of an abundant number of eccentric white blood cells (WBC). Acute lymphoblastic leukemia (ALL) has been categorized into three morphological subtypes by the French–American–British (FAB) classification: L1, L2 and L3. Seeing blast cells in the peripheral blood slide or increase in the slide of bone marrow is the first phase of detecting this type of leukemia. Boring and grinding are the most challenging sections of this method for pathologists because it is time-consuming and the diagnosis is dependent on the skill of the pathologist.

To solve these problems, some researchers noticed automatic methods for ALL identification from microscopic images. However, several similarities between the morphology of ALL (L1, L2, L3) and lymphocyte forms (normal, reactive and atypical) have remained a high challenge in implementing automatic systems. Figure 1 demonstrates two instances of ALL and lymphocyte forms.

Some researchers have tried to segment blood cells using fuzzy c-means (FCM) clustering and HSV (hue, saturation, value) space features[1]. These approaches have been used to reach better classification methods for performing the segmentation automatically in some studies[2]. Furthermore, the features of WBCs were extracted for automatic system diagnosis of blood disorders. Theera-Umpon *et al.*[3] posed a question whether information from the nuclei alone is sufficient to categorize WBCs. Morphological features (granulometries) were extracted from any segmented nuclei of cells. Naive Bayes classifier and neural networks (NN) were also employed as classifiers. The outcomes showed that the features extracted from nuclei led to an accuracy of 77% on the test classes. Hayan *et al.*[4] focused their study on WBCs segmentation via a fusion of automatic contrast stretching supported by applying image arithmetic process, global threshold methods and minimum filter. Halim *et al.*[5] used segmentation on HSI colour space for removing the WBCs in the background. They applied the erosion morphological operator to separate the overlapping cells. Their method showed the highest average accuracy of 97.8% for counting both AML and ALL cells. Lim *et al.*[6] presented a system consisting of gradient amount, thresholding, morphological processes and watershed algorithm for segmentation of AML cells. They achieved a separation correctness of 94.5% for 50 microscopic images but the mean accuracy for M2, M5 and M6 forms were 94%, 95% and 95% respectively. Furthermore, Subrajeet *et al.*[7] provided a numerical microscopic method for the differentiation of cancerous from noncancerous cells in microscopic images. Using image clustering and extraction of various forms of features, white blood cells were identified and segmented. Lastly, a group of classifiers were trained to identify cancerous cells. Though the outcomes of this techniques were good, the repeatability of the trial and evaluation with other techniques were not feasible because they were acquired by means of a dedicated dataset. Abbas *et al.*[8] segmented the nuclei of cells by image processing techniques such as OTSU global thresholding and morphological operation dilation. In a similar study[9], nuclei of cells were segmented by k-means process using image preprocessing phase. Seventy seven geometric

*For correspondence. (e-mail: mdeevband@sbmu.ac.ir)

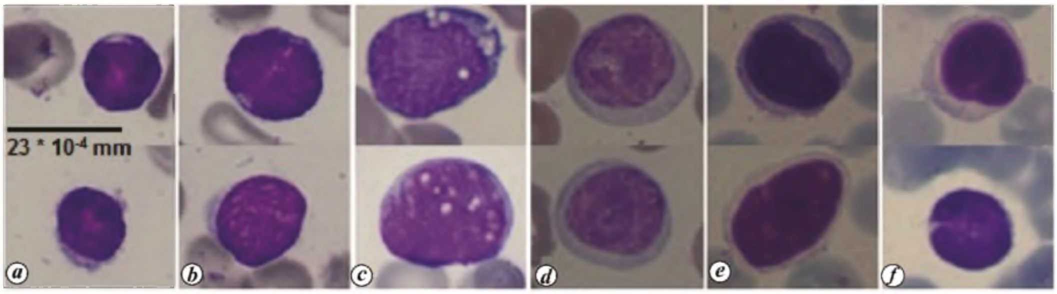

Figure 1. Sample cell images. ***a***, L1; ***b***, L2; ***c***, L3; ***d***, Atypical; ***e***, Reactive; ***f***, Normal.

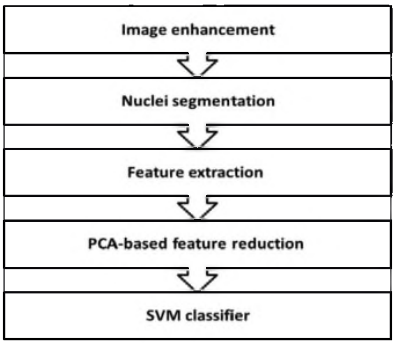

Figure 2. Block diagram of the presented algorithm.

and statistical features were extracted from these nuclei. The features that led to high accuracy in the classification phase were chosen as the inputs of SVM classifier with ten-fold cross validation to categorize the cells as ALL and lymphocytes. These cells were categorized into their forms using multi-SVM classifier. The steps of the proposed algorithm in this work are similar to those adopted by Amin *et al.*[9] except that we improved the nucleus segmentation procedure, i.e. k-means, with fuzzy c-means as the former rarely leads to an empty cluster when running. Also for feature selection, we considered the first 6 principal components of the feature space, which were taken via principal component analysis (PCA) method.

Materials and methods

Dataset

The dataset of this work contained 21 blood and bone marrow smears of 7 normal people and 14 individuals with ALL. The smears were stained by giemsa for visualization of nuclei ingredients. The images obtained were digitalized by digital camera (Nikon1 V1), connected to a light microscope (Nikon Eclipse 50i) with an effectual magnification of 1000×. The template of the images was Joint Photographic Experts Group (JPEG) at the utmost resolution of the digital camera, i.e. 2592 × 3872 pixels in red–green–blue (RGB) colour space. The images obtained were reviewed via the pathologist to decide the true form of the blood smears. In this study, 312 images have been obtained containing 146 images of ALL forms (L1, L2 and L3) and 166 images of lymphocyte cells (normal, reactive and atypical). A whole number of 958 cells were acquired. The dataset included six classes of WBCs – L1, L2, L3, normal, reactive and atypical – with 277, 215, 151, 50, 94 and 171 cells respectively.

Presented algorithm

The total steps of this algorithm are presented in the flow chart of Figure 2.

Image enhancement: The illumination of the microscope while acquiring the images is set manually by the user. Therefore, the brightness of the images taken at different times will be different which affects the textural features and result in wrong diagnosis. To reduce this error, histogram equalization on V channel of HSV colour space is proposed.

The image was first transformed from RGB to HSI colour space. This decreases relevance among the colour components (compared with RGB) and deals with H, S and I components singly. In HSI colour space, colour information was inserted in H and S components, whereas the I channel related directly to the intensity and matched human perception of light. The common histogram equalization method was then used on I channel for equalizing the grey level of image lightness. Histogram equalization decreases the effects of various luminosity situations in various image acquisition conditions[11], thus all the images had almost the same lightness. In Figure 3, four examples of histogram equalization are shown.

Nuclei segmentation: Nuclei segmentation plays a vital role as it affects subsequent steps such as feature classification and extraction. In a similar work, *k*-means algorithm was used for segmentation of nuclei, but this algorithm yields an empty cluster at times[9]. To avoid this, fuzzy *c*-means clustering technique was used in this study. For the first time, fuzzy *c*-means was presented by Bezdek *et al.*[12–15]. This algorithm assigns a value between 0 and 1 to every pixel of the image that expresses the degree of membership to each cluster centre. It is according to optimization of the cost function J_m

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

where m is a real number larger than one. Sets $C_1, C_2, \ldots, C_c$ and $X_1, X_2, \ldots, X_n$ are cluster centres and data sample vectors. And u_{ik} is the ith cluster belonging to the kth input sample X_k. Minimization of the cost function shown above is applied by updating the belonging u_{ik} and cluster centre C_i

$$u_{ik} = \frac{1}{\sum_{j=1}^{C} \left\{ \frac{\| x_k - C_i \|}{\| x_k - C_j \|} \right\}^{2/(m-1)}}, \tag{2}$$

$$C_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} X_k}{\sum_{k=1}^{n} u_{ik}^{m}}. \tag{3}$$

The fuzzy c-means method with four clusters was used on the 3D image in HSI colour space. The clusters relate to nuclei, background and other cell components. Figure 4 displays four clusters of four sample images after using FCM.

According to experience, the cluster with the minimum red component is the cluster corresponding to the nuclei. Thus, the average amount of red component was considered for all clusters, and the cluster with minimum amount was considered as cluster of nuclei.

In addition, a two-step post-processing on the nuclei cluster image was performed. First, to fill some minor holes and remove smear noises from the nucleus cluster, opening and closing process in morphological operation binary was performed on the image. Second, to make the linked nuclei discrete, watershed transform was applied. Watershed transform can identify the border lines among the linked nuclei[16] and effectively divide all linked cells into their individual nuclei. In Figure 5, four clusters of nuclei and their post-processed style are shown.

Feature extraction: Some features (geometric and statistical) should be extracted from the nuclei to classify the cells as ALL or lymphocytes and determine their subtype, i.e., L1, L2, L3, atypical, reactive and normal. Geometric features give information on the scale and figure of a cell whereas statistical features provide information on grey level of image histogram. According to pathologists, the geometry of the nuclei is one of the important features that can be applied for its specifications. The geometrical features include: area, perimeter, solidity, eccentricity and extent of nucleus from the binary image of the nucleus. Statistical features give information on the repartition of intensities in one image. These features are produced from the grey level image histograms of the blue, red and green, plus the intensity (value), saturation and hue components from the original and enhanced image of the nuclei. It contains measures such as standard deviation, mean, energy, entropy, kurtosis and skewness. Seventy two statistical features have been generated by this method[17].

PCA-based feature reduction: The main purpose of PCA is dimensionality reduction of data including a number of interdependent variables while preserving the variation present in the data as much as possible[18–23]. This is achieved by converting to a new set of variables, the principle components (PCs), that are uncorrelated, and are arranged so that the first few maintain maximum variation present in all of the primary variables[24].

The six general steps for applying PCA are[25]:

(1) Place all the d-dimensional features in one matrix. In this study, the dimension of features is 77 and whole number of data, i.e. nuclei, is 958. Thus we have a matrix of features with dimension of 77 × 958.

(2) For PCA to work correctly, the mean must be deducted from each matrix element. For this part, the average of each row of the matrix is calculated and deducted from every row of the matrix.

(3) Calculate the scatter matrix (or the covariance matrix) of the total data set.

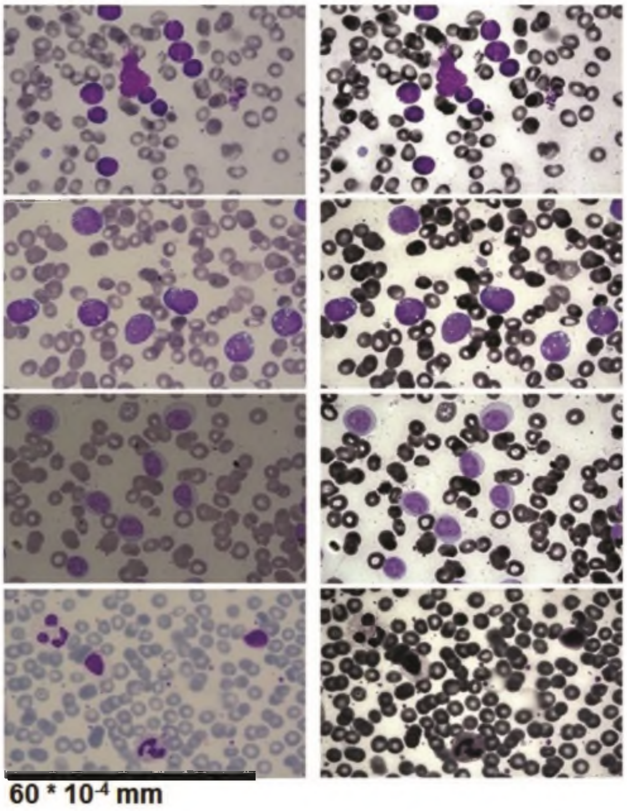

Figure 3. Result of preprocessing step. Left column: Original images. Right column: Enhanced images.

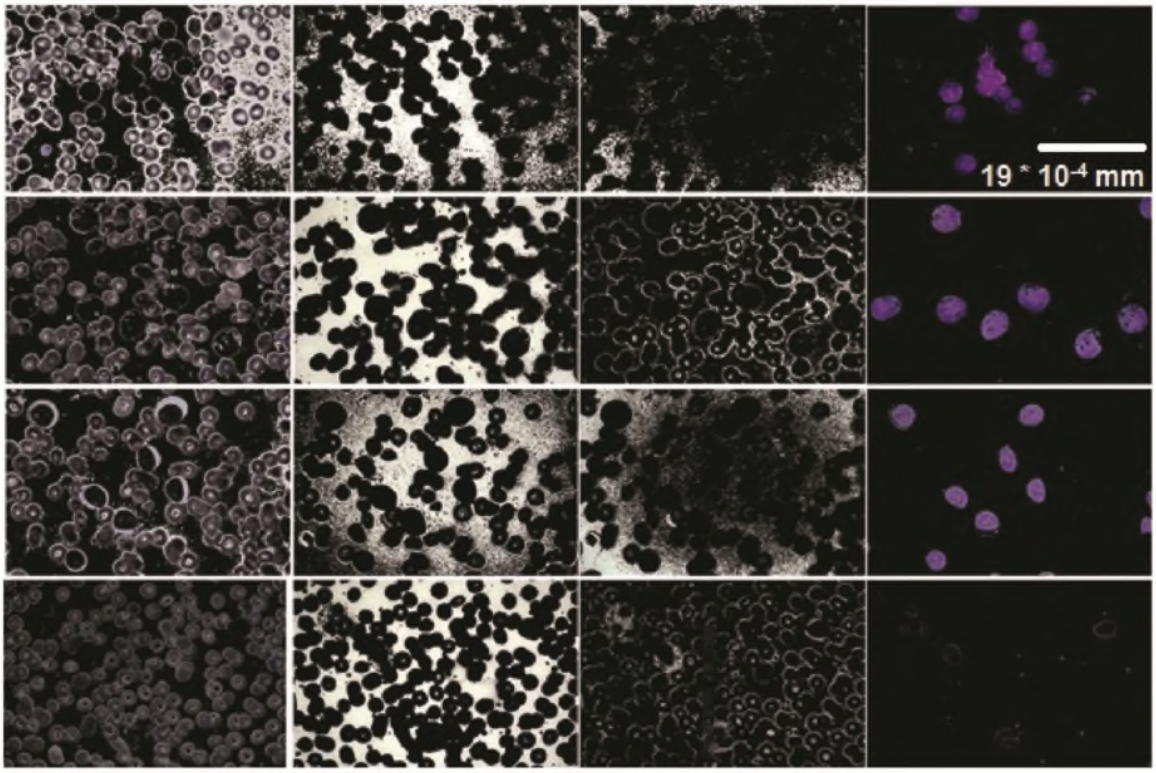

Figure 4. Result of fuzzy *c*-means clustering.

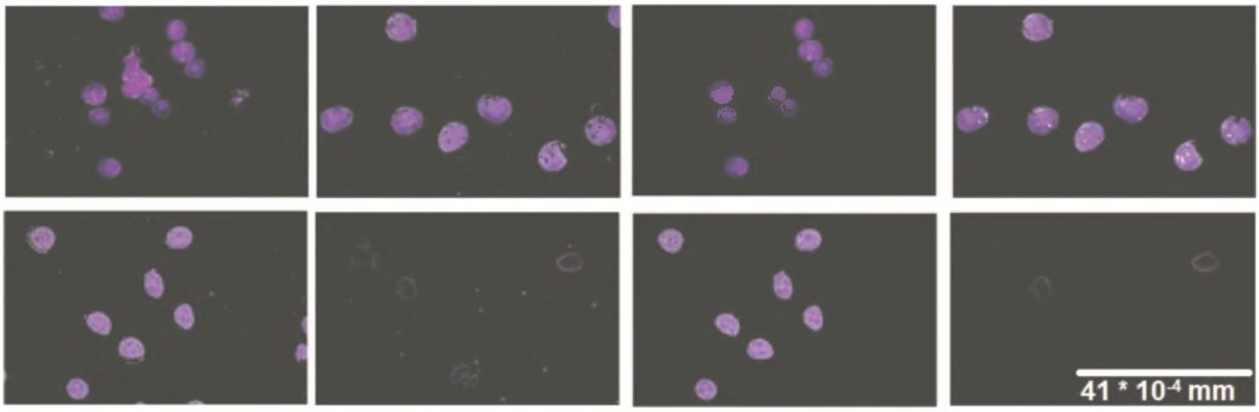

Figure 5. Left column: clusters of nuclei. Right column: extracted nuclei.

Table 1. ALL and lymphocytes cells versus result of binary SVM classifier

		Binary SVM output	
		ALL	Lymphocytes
Binary SVM input	ALL	634	9
	Lymphocytes	9	306

(4) Compute the eigenvalues and eigenvectors from the scatter matrix.

(5) Sort the eigenvectors by reducing eigenvalues and select k eigenvectors with the greatest eigenvalues to form a $d \times k$ dimensional matrix W (each column represents an eigenvector).

(6) Apply this $d \times k$ eigenvector matrix to convert the samples into the new subspace.

After above the steps were performed, the first six PCs were chosen as the classifier inputs.

SVM classifier: The SVM classifier is a suitable selection for classification especially for the patterns that are very close to each other in the feature space[26]. To classify cells as ALL or lymphocytes, common SVM was applied, which is in fact binary classification, and multiclass SVM classifier was applied for detection of cell subtypes. This matter is attained via a separating surface in the input space of the data set, by applying several

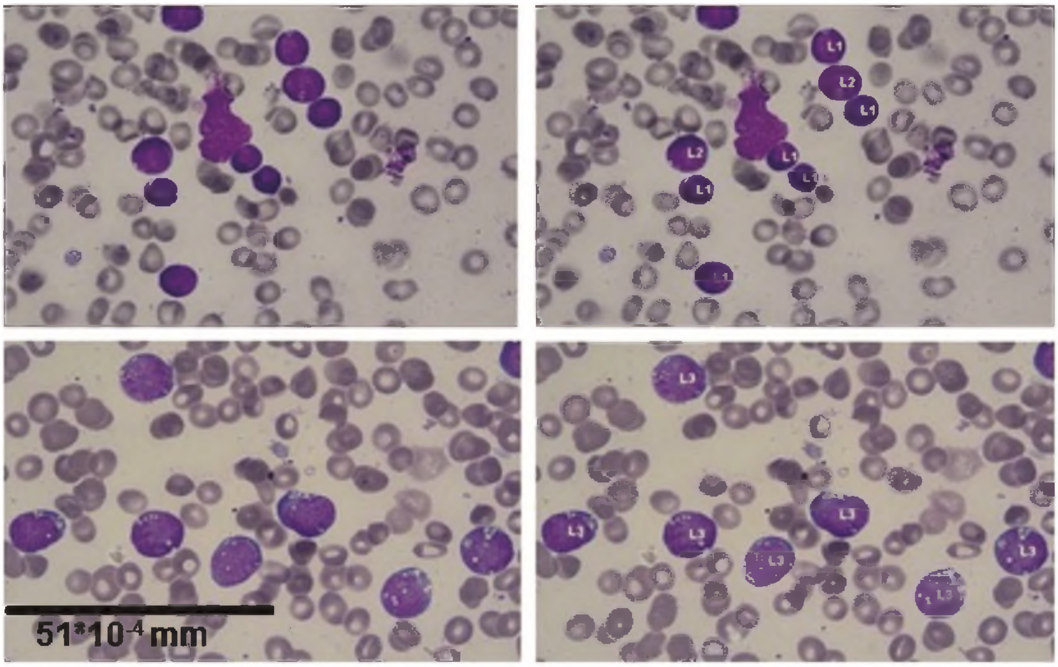

Figure 6. Results of classification. Left column: original image containing ALL cells. Right column: labelled ALL cells (L1, L2 and L3).

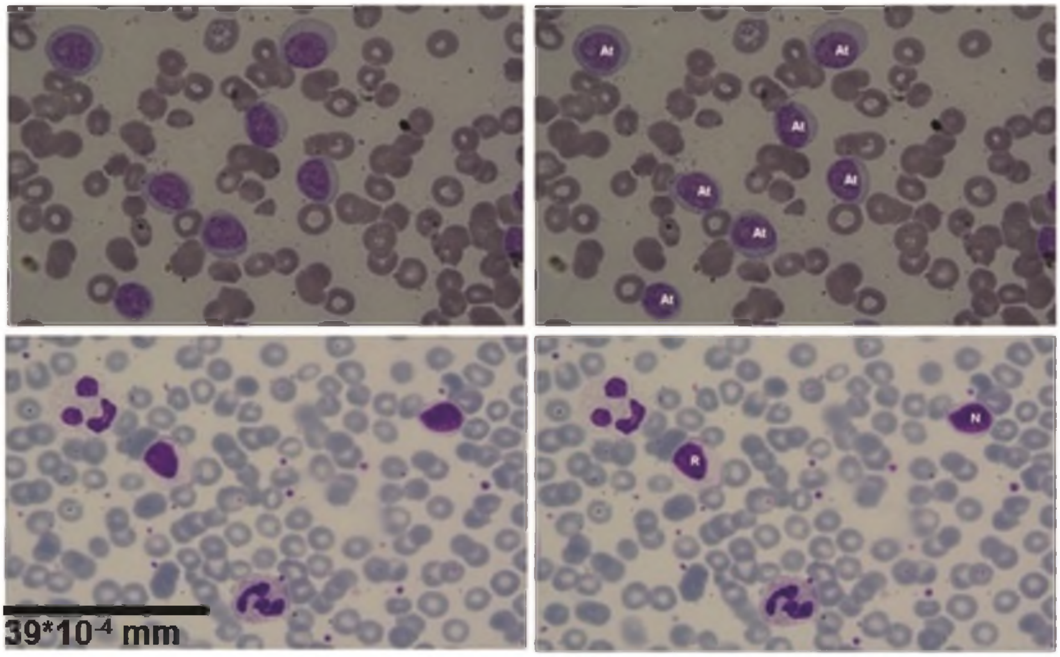

Figure 7. Results of classification. Left column: original image containing lymphocytes cells. Right column: labelled lymphocytes cells (At: atypical, N: normal and R: reactive).

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	252	19	2	2	1	1
	L2	15	184	4	11	0	1
	L3	1	3	147	0	0	0
	Atypical	0	8	2	161	0	0
	Normal	1	0	0	0	41	8
	Reactive	1	1	0	1	9	82

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

Statistical parameters	Selected features (%)	Dimension reduced features (%)
Sensitivity	98	99
Specificity	95	97
Accuracy	97	98
Precision	98	99
False negative rate	2	1

Table 4. Performance of the multi-SVM classifiers using selected features versus dimension reduced features for ALL

	L1		L2		L3	
Statistical parameters	Selected features (%)	Dimension reduced features (%)	Selected features (%)	Dimension reduced features (%)	Selected features (%)	Dimension reduced features (%)
Sensitivity	91	91	84	86	97	97
Specificity	97	97	95	96	99	99
Accuracy	95	96	92	94	99	99
Precision	92	93	84	86	95	95
False negative rate	9	9	16	14	3	3

kernel functions as linear or nonlinear such as quadratic, polynomials and radial basis functions (RBF)[27,28]. Based on optimum accuracy of differentiation, RBF kernel with sigma 3 was used in this work. To assess the classifier, the *k*-fold cross validation technique with $k = 10$ was used.

Results

After applying the proposed methods on four microscopic images, outcomes of classification were presented (Figures 6 and 7).

Confusion matrices of the binary SVM and Multi-SVM were achieved (Tables 1 and 2).

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

As mentioned above, the inputs of the classifiers are the first six components of PCA. The classification results were compared with similar studies using individually chosen features with high efficiency as inputs of the classifiers[9]. The comparative outcomes are presented in Tables 3–5. Using PCA-based dimension reduced features, there was little improvement in the performance of the binary classifier according to the values of sensitivity, specificity, accuracy, precision and false negative rate in Table 3. From Table 4, it can be seen that a similar condition also occurs for multiclass SVM in classifying ALL subtypes. However, it can be observed from Table 5 that sensitivities of detection of lymphocytes, namely normal and reactive cells, have enhanced meaningfully. Compared to a similar study[9], the average sensitivity of detection of both normal and reactive lymphocytes has improved by 15%. The average precision of detection of normal and reactive lymphocytes has improved by 19% and 10% respectively.

Discussion

An enhanced automatic system is presented for classification of ALL and lymphocytes cells as well as their subtypes in this study. The basic contribution of this paper is to enhance similar studies that used individually chosen features with high efficiency as inputs of the classifiers. This was reached by applying an extra step of dimension reduction of feature space by using PCA.

Comparison with a similar study indicated that using PCA instead of exclusively selected features enhanced the average sensitivity and precision of classification up to 10%. The results determine that this method is superior to the one used earlier.

By referring to the classification results obtained in this study, it is clear that although our proposed system is relatively simple, this method has an acceptable efficiency for the identification of ALL and lymphocytes as well as categorizing ALL into L1, L2 and L3 subsets, and the lymphocytes into atypical, reactive and normal cells. Hence, its permissible efficiency for identifying the ALL and lymphocyte cells makes it an associate diagnostic device for pathologists.

It can be considered that an issue we faced when testing our method was the lack of an existing database. Actually, numerous researchers examined their methods with a few images only, that are not widely existing. Therefore, we cannot straightaway compare our results with the outcomes achieved via numerous presented systems.

Table 5. Performance of the multi-SVM classifiers using selected features versus dimension reduced features for lymphocytes

	Atypical		Normal		Reactive	
Statistical parameters	Selected features (%)	Dimension reduced features (%)	Selected features (%)	Dimension reduced features (%)	Selected features (%)	Dimension reduced features (%)
Sensitivity	95	94	66	82	73	87
Specificity	98	98	97	99	98	99
Accuracy	97	97	96	98	95	98
Precision	93	92	61	80	79	89
False negative rate	5	6	34	18	27	13

For future studies, we believe that besides nuclei, segmentation of cytoplasm of cells and feature extraction from it, can enhance efficiency of our automatic system. We also believe that, manually fixing blood slides can be a great challenge in microscopic image studies. If there is a structure that can fix blood slides in a typical and automatic process in addition to capturing the digital images automatically, it will result in more similar images and this leads to more accurate response on a variety of data sets. Finally, applying such a system can result in an online recognition method.

1. Sinha, N. and Ramakrishnan, A., Automation of differential blood count. In TENCON 2003, Conference on Convergent Technologies for the Asia-Pacific Region, IEEE, 2003.
2. Theera-Umpon, N., White Blood Cell Segmentation and Classification in Microscopic Bone Marrow Images, Fuzzy systems and knowledge discovery, 2005, p. 485.
3. Theera-Umpon, N. and Dhompongsa, S., Morphological granulometric features of nucleus in automatic bone marrow white blood cell classification. In *IEEE Transactions on Information Technology in Biomedicine*, 2007, vol. 11, issue 3.
4. Madhloom, H. *et al.*, An automated white blood cell nucleus localization and segmentation using image arithmetic and automatic threshold. *J. Appl. Sci.*, 2010, **10**, 959–966.
5. Halim, N. H. A., Mashor, M. Y. and Hassan, R., Automatic blasts counting for acute leukemia based on blood samples. *Int. J. Res. Rev. Comput. Sci.*, 2011, **2**(4), 278–284.
6. Nee, L. H., Mashor, M. Y. and Hassan, R., White blood cell segmentation for acute leukemia bone marrow images. *J. Med. Imaging Health Inform.*, 2012, **2**(3), 278–284.
7. Mohapatra, S., Patra, D. and Satpathy, S., An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Comput. Appl.*, 2014, **24**(7–8), 1887–1904.
8. Abbas, N. and Mohamad, D., Automatic color nuclei segmentation of leukocytes for acute leukemia. *Res. J. Appl. Sci., Eng. Technol.*, 2014, **7**(14), 2987–2993.
9. Amin, M. M. *et al.*, Recognition of acute lymphoblastic leukemia cells in microscopic images using *k*-means clustering and support vector machine classifier. *J. Med. Signals Sensing*, 2015, **5**(1), 49.
10. Mokhtar, N. *et al.*, Image enhancement techniques using local, global, bright, dark and partial contrast stretching for acute leukemia images. *Anal. Cell. Pathol.*, 2003, **25**, 1–36.
11. Rodenacker, K. and Bengtsson, E., A feature set for cytometry on digitized microscopic images. *Anal. Cell. Pathol.*, 2003, **25**(1), 1–36.
12. Keller, J., Krisnapuram, R. and Pal, N. R., *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Springer Science & Business Media, 2005, vol. 4.
13. Ashtiyani, M., Asadi, S. and Birgani, P. M., ICA-based EEG classification using fuzzy C-mean algorithm. In 3rd International Conference on IEEE, Information and Communication Technologies: From Theory to Applications, 2008.
14. Ashtiani, M., Asadi, S. and Goudarzi, P. H., A new method in transmitting encrypted data by FCM algorithm. In Information and Communication Technologies, ICTTA'06, IEEE, 2006.
15. Birgani, P. M., Ashtiyani, M. and Asadi, S., MRI segmentation using fuzzy c-means clustering algorithm basis neural network. In 3rd International Conference on Information and Communication Technologies: From Theory to Applications, ICTTA, IEEE, 2008.
16. Khajehpour, H. *et al.*, Detection and segmentation of erythrocytes in blood smear images using a line operator and watershed algorithm. *J. Med. Signals Sens.*, 2013, **3**(3), 164.
17. Ashtiyani, M. *et al.*, Transmitting encrypted data by wavelet transform and neural network. In IEEE International Symposium on Signal Processing and Information Technology, 2007.
18. Alizadeh, A., Fatemizadeh, E. and Deevband, M. R., Investigation of Brain Default Network's activation in autism spectrum disorders using group independent component analysis. In 21st Iranian Conference on Biomedical Engineering, IEEE, 2014.
19. Abdi, H. and Williams, L. J., *Principal Component Analysis*, Wiley interdisciplinary reviews: computational statistics, 2010, **2**(4), 433–459.
20. Mansoory, M. S., Ashtiyani, M. and Tajik, H., Cardiac motion evaluation for disease diagnosis using ICA basis neural network. In IACSITSC'09, International Association, IEEE Computer Science and Information Technology-Spring Conference, 2009.
21. Osman, N. A. A. *et al.*, 4th Kuala Lumpur International Conference on Biomedical Engineering 2008, BIOMED 2008, 25–28 June 2008, Kuala Lumpur, Malaysia, Springer Science & Business Media, 2008, vol. 21.
22. Mirmohammadi, P., Taghavi, A. and Ameri, A., Automatic Recognition of Acute Lymphoblastic Leukemia Cells from Microscopic Images.
23. Ashtiyani, M. *et al.*, EEG classification using neural networks and independent component analysis. In 4th Kuala Lumpur International Conference on Biomedical Engineering, Springer, 2008, pp. 179–182; https://link.springer.com/chapter/10.1007/978-3-540-69139-6_48
24. Jolliffe, I., *Principal Component Analysis*, Wiley Online Library, 2002.
25. Smith, L. I., *A Tutorial on Principal Components Analysis*, Cornell University, USA, 2002, vol. 51, p. 52.
26. Mohapatra, S. and Patra, D., Automated cell nucleus segmentation and acute leukemia detection in blood microscopic images. In International Conference on IEEE Systems in Medicine and Biology (ICSMB), 2010.
27. Smola, A. J. and Schölkopf, B., A tutorial on support vector regression. *Stat. Comput.*, 2004, **14**(3), 199–222.
28. Alvar, A. A., Deevband, M. R. and Ashtiyani, M., Neutron spectrum unfolding using radial basis function neural networks. *Appl. Radiat. Isot.*, 2017, **129**, 35–41.

Received 28 October 2017; revised accepted 23 July 2018

doi: 10.18520/cs/v115/i8/1512-1518