Automatic recognition of acute lymphoblastic leukemia using multi-SVM classifier

Pouria Mirmohammadi¹, Amirhossien Rasooli², Meghdad Ashtiyani¹, Morteza Moradi Amin² and Mohammad Reza Deevband¹,*

¹Department of Biomedical Engineering and Medical Physics, Faculty of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
²Department 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 smears by 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 multi-SVM classifier was trained to identify cancerous cells. Though the outcomes of this techniques were good, the repeatability of the trial 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 study⁹, nuclei of cells were segmented by k-means process using image preprocessing phase. Seventy seven geometric

Keywords: ALL, fuzzy c-means method, PCA analysis, SVM classifier.
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. 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, 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. To avoid this, fuzzy c-means clustering technique was used in this study. For the first time, fuzzy c-means was presented by Bezek et al.. 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 $J_m$:

$$J_m(U, C) = \sum_{i=1}^{c} \sum_{k=1}^{n} \mu_{ki}^m \| X_k - C_i \|^2,$$  (1)
where $m$ is a real number larger than one. Sets $C_1$, $C_2$, ..., $C_c$ and $X_1$, $X_2$, ..., $X_n$ are cluster centres and data sample vectors. And $u_{ik}$ is the $i$th cluster belonging to the $k$th input sample $X_k$. Minimization of the cost function shown above is applied by updating the belonging $u_{ik}$ and cluster centre $C_i$

\[
  u_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \| x_k - C_j \| \right)^{2/(m-1)}}
\]  

\[
  C_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}
\]

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 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.

**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. 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.

The six general steps for applying PCA are:

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 $77 \times 958$.
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.png)
Table 1. ALL and lymphocytes cells versus result of binary SVM classifier

<table>
<thead>
<tr>
<th>Binary SVM input</th>
<th>ALL</th>
<th>Lymphocytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>634</td>
<td>9</td>
</tr>
<tr>
<td>Lymphocytes</td>
<td>9</td>
<td>306</td>
</tr>
</tbody>
</table>

(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. 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

<table>
<thead>
<tr>
<th>Multi-SVM input</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>Atypical</th>
<th>Normal</th>
<th>Reactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>252</td>
<td>19</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L2</td>
<td>15</td>
<td>184</td>
<td>4</td>
<td>11</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>L3</td>
<td>1</td>
<td>3</td>
<td>147</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Atypical</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>161</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Normal</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>8</td>
</tr>
<tr>
<td>Reactive</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>82</td>
</tr>
</tbody>
</table>
Table 3. Performance of the binary classifiers using selected features versus dimension reduced features

<table>
<thead>
<tr>
<th>Statistical parameters</th>
<th>Selected features (%)</th>
<th>Dimension reduced features (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>Specificity</td>
<td>95</td>
<td>97</td>
</tr>
<tr>
<td>Accuracy</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>Precision</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>False negative rate</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Statistical parameters</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected features (%)</td>
<td>Dimension reduced features (%)</td>
<td>Selected features (%)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>91</td>
<td>91</td>
<td>84</td>
</tr>
<tr>
<td>Specificity</td>
<td>97</td>
<td>97</td>
<td>95</td>
</tr>
<tr>
<td>Accuracy</td>
<td>95</td>
<td>96</td>
<td>92</td>
</tr>
<tr>
<td>Precision</td>
<td>92</td>
<td>93</td>
<td>84</td>
</tr>
<tr>
<td>False negative rate</td>
<td>9</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Statistical parameters</th>
<th>Atypical</th>
<th>Normal</th>
<th>Reactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected features (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced features (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>95</td>
<td>66</td>
<td>73</td>
</tr>
<tr>
<td>Specificity</td>
<td>98</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>Accuracy</td>
<td>97</td>
<td>96</td>
<td>95</td>
</tr>
<tr>
<td>Precision</td>
<td>93</td>
<td>61</td>
<td>79</td>
</tr>
<tr>
<td>False negative rate</td>
<td>5</td>
<td>34</td>
<td>27</td>
</tr>
</tbody>
</table>

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.


Received 28 October 2017; revised accepted 23 July 2018
doi: 10.18520/cs/v115/i8/1512-1518