Biomedical Signal Processing and Control 33 (2017) 272–280

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection

Sonali Mishra*, Banshidhar Majhi, Pankaj Kumar Sa, Lokesh Sharma

Pattern Recognition Laboratory, Department of Computer Science & Engineering, National Institute of Technology, Rourkela 769 008, India

ARTICLE INFO

Article history:
Received 30 April 2016
Received in revised form 1 November 2016
Accepted 28 November 2016
Available online 22 December 2016

Keywords:
Leukemia
Acute lymphoblastic leukemia
Gray level co-occurence matrix
Watershed segmentation

ABSTRACT

In this paper, we have proposed an acute lymphoblastic leukemia detection strategy from the microscopic images. The scheme utilizes all the steps associated with any other classification scheme, but our contribution lies on a marker-based segmentation(MBS), gray level co-occurrence matrix (GLCM) based feature extraction, and probabilistic principal component analysis(PPCA) based feature reduction. The relevant features are used in a random forest (RF) based classifier. Extensive experiments are carried out on the ALL-IDB1 dataset, and comparative analysis has been made with other existing schemes with respect to sensitivity, specificity, and classification accuracy. The proposed scheme (MBS+GLCM+PPCA+RF) achieves 96.29% segmentation accuracy and classification accuracy of 99.004% and 96% for nucleus and cytoplasm respectively.

1. Introduction

Microscopic analysis of blood is an important step in medical diagnosis to identify the blood-related diseases. The major challenge lies in recognition and classification of white blood cells from the blood smear. Leukemia [1,2] is a type of hematopoietic disease that starts with the bone marrow and results in the development of blast cells. Leukemia can be divided into two major types, namely, acute and chronic. By French-American-British (FAB) characterization plan, acute leukemia is further grouped into two sub-classes, namely, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) depending upon the hematological organ it affects [3]. In the present work, Acute Lymphoblastic Leukemia classification has been considered.

Acute lymphoblastic leukemia is a cancer of white blood cells (WBCs) also known as lymphoblasts. ALL is a collection of malignancies that starts with the overproduction of lymphoblasts in the bone marrow where WBCs continuously multiply and inhibit the production of normal blood cells, i.e., red and white blood cells, and platelets in the bone marrow. As a result, the human body loses the potential to fight with the external organisms and lead to death. This hematopoietic disorder is mainly seen in children of 2-5 years of age as well as in adults of older than 50 years old [4,5]. During diagnosis, it can be detected through identification and classification of WBCs. The uncontrolled rate of blast cells in WBCs characterizes the different phases of lymphoblastic leukemia.

The morphological discriminating proof of the existence of ALL is accomplished by the hematologists that start with a test of bone marrow collected from the spine. Wright's staining technique is used so as to make the blood cells visible throughout investigation [6]. This method includes numerous disadvantages, such as being a time-consuming procedure, having low accuracy, and necessity of a sincere hematologist. The primary objective of this work is to devise a computer-aided diagnostic (CAD) system for the detection and classification of ALL to support the decision of hematologists. The suggested strategy mainly comprises of three different steps, i.e., segmentation, feature extraction, and classification of the ailment. Recently, numerous automated procedures have been proposed for segmentation and feature extraction. However, they lack in accuracy and computational efficiency. Thus there is a scope to improve the detection and separation of grouped leukocytes more precisely. The proposed strategy suggests a marker-based watershed segmentation scheme to separate the grouped cells present in a microscopic image. A gray level co-occurrence matrix based scheme is applied for the characterization of features and assessment of the blast cells that leads to the correct diagnosis of the disease. A feature reduction algorithm based on the probabilistic principal component analysis (PPCA) is used to extract the relevant features. Subsequently, a random forest classifier is used to detect the cells as benign or malignant.

The rest of the paper is organized as follows: Section 2 deals with the related works for the recognition of leukemia from the blood smear alongside some segmentation plans. Section 3 describes the

* Corresponding author.
E-mail address: smishra.nitrkl@gmail.com (S. Mishra).

http://dx.doi.org/10.1016/j.bspc.2016.11.021
1746-8094/© 2016 Elsevier Ltd. All rights reserved.

essential steps followed for ALL detection. Section 4 gives a comparative analysis of the proposed strategy. Finally, Section 5 provides the concluding remarks.

2. Related Work

During the last two decades, researchers have been working actively in the field of medical image processing and proposed many techniques. The majority of the work is dedicated to conquer the issue in the visual assessment of blood cells by the hematologists. The major steps involved in an automated diagnosis are segmentation of white blood cells to extract the nucleus and cytoplasm followed by feature extraction and classification. The steps are interrelated, and the success of one reflects on other.

Angulo et al. [7] have proposed a two-stage segmentation scheme based on a thresholding and binary filtering. Though the scheme shows good segmentation results, it takes more computational time because of the two-way process. Ko et al. [8] have proposed a hybrid leukocyte segmentation scheme which uses a Gaussian Vector Flow (GVF) snake model to remove the boundary of an object. Two different schemes are employed independently to extract the cytoplasm and nucleus of the leukocyte. However, the segmentation accuracy for cytoplasm is poor and involves higher computation overhead. Mohapatra et al. [9] have proposed an unsupervised segmentation technique based on color based clustering for the detection of leukemia. The authors have taken care of fractal dimension, shape and texture features including contour signature. Mohapatra et al. [10] have also used functional link architecture for segmenting nucleus and cytoplasm from lymphocyte image. The proposed scheme takes the whole segmentation process as a pixel classification problem, and a neural network architecture is employed to categorize each pixel into the cytoplasm, nucleus or background. A small dataset of 96 images has been used for the segmentation process. Liu et al. [11] have suggested a clustering technique for segmenting WBCs. They have developed a combination of three methods, namely, color space transformation, and enhancement, mean shift clustering and selection of 'C' component of CMYK color space and finally nucleus mark watershed operation is used for segmenting the whole leukocyte from the blood sample. Arslan et al. [12] have proposed an algorithm that uses color and shape for segmentation of WBCs present in the bone marrow images. Further, they have focused on the color and shape characteristics of the cells by using marker-based watershed segmentation. They have got favorable results for the WBC segmentation. Madhloom et al. [13] have proposed an automatic system for segmenting the lymphoblast cells. This research presents a new method which combines color and morphological features to isolate a lymphoblast from other cells present in the microscopic image. Zhang et al. [14] have proposed a framework for the detection of Alzheimer's disease based on structural volumetric MR images using 3D-DWT. The method is employed on 178 subjects which give an overall accuracy of 81.5% using WTA-KSVM and PSOVAC method. Zhang et al. [15] have suggested an optimal multi-level thresholding technique for segmentation. They have used Tsallis entropy which gives superior results than the traditional entropy mechanism along with bee colony approach which is computationally fast. Even though the segmentation accuracy is shown to be satisfactory, the parameter needs to be tuned using trial and error.

Scotti [16] has proposed a technique for automated grouping of ALL. According to the experiments directed by them, it has been presumed that lymphoblast characterization is achievable from blood images utilizing morphological elements. Gupta et al. [17] have proposed a relevant vector machine-based system for the classification of lymphoblasts. The characterization exactness for the childhood ALL has been promising, however, it is poor in adult cases. Escalante et al. [18] have recommended an option to deal with leukemia sub-arrangement utilizing particle swarm model, where physically segregated leukemia cells are divided utilizing a Markov random field. Mohapatra et al. [19] have recommended an ensemble of classifier system in which they have tried to improve the accuracy of the diagnostic system by analyzing morphological and textural features from the peripheral blood smear. Faivdullah et al. [20] have used shape based features in a small data set of 100 images for the detection of the disease using support vector machine (SVM) as a classifier. Putzu et al. [21] have proposed an approach which isolates the whole leukocyte from a microscopic image and further separates the nucleus and cytoplasm from the leukocyte. Shape, color, and texture features are extracted to train different classification models to determine the best one for leukemia classification.

It has been observed from the literature that segmentation of WBCs from the microscopic images of blood cells plays an important role and the existing schemes lack in accuracy and separation of grouped cells. Further, the extraction of suitable features and selection of relevant features from a high dimension feature set is of utmost importance for the correct diagnosis of ALL and need further investigation. To overcome these issues, we have applied a marker-based watershed segmentation on the microscopic images to get a separation between the grouped cells. The texture features are extracted using GLCM from the nucleus and cytoplasm region. PPCA is applied to get the most relevant features.The relevant features are trained using a random forest classifier to classify the cells.

3. Proposed work

The proposed methodology has different phases like any other classification scheme and shown in Fig. 1. The pre-processing and classification phases utilize the existing methodologies. Our contribution lies in a marker-based watershed segmentation, GLCM based feature extraction, and use of probabilistic PCA (PPCA) for feature reduction. The phases are discussed below in sequel.

3.1. Pre-processing

ALL-IDB is a publicly available dataset and is a collection of microscopic images of blood samples collected from different individuals [22,23]. The specimen in ALL-IDB dataset are collected at the M. Tettamanti Research Center for childhood leukemia and hematological illnesses, Monza, Italy. The majority of the images in this dataset was caught with a research facility magnifying instrument at various amplifications, extending from 300 to 500, combined with an Olympus C2500L camera and are of 1712 × 1368 dimension. The dataset is categorized into two subsets namely, ALL-IDB1 and ALL-IDB2. All our experiments are conducted in ALL-IDB1, and it has 108 samples, out of which 49 are affected by acute lymphoblastic leukemia, where a total of 510 lymphoblasts are present. Some of the lymphoblasts are grouped together and difficult to be detected by hematologists.

Before segmentation, to find the region of interest (nucleus and cytoplasm) the images are improved by histogram equalization and Weiner filtering in succession. Subsequently, an intensity transformation is applied using thresholding to keep only WBCs on a background in a binary image. Fig. 2 shows the overall steps for detecting the WBCs from a sample microscopic image.

3.2. Marker-based watershed segmentation

Prior to feature extraction and classification, the objective is to separate the cells from the grouped ones using a suitable

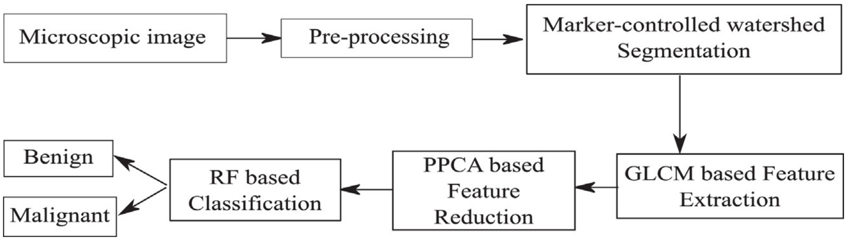
Microscopic image
Pre-processing
Marker-controlled watershed Segmentation
GLCM based Feature Extraction
PPCA based Feature Reduction
RF based Classification
Benign
Malignant

Fig. 1. Block diagram of the proposed scheme

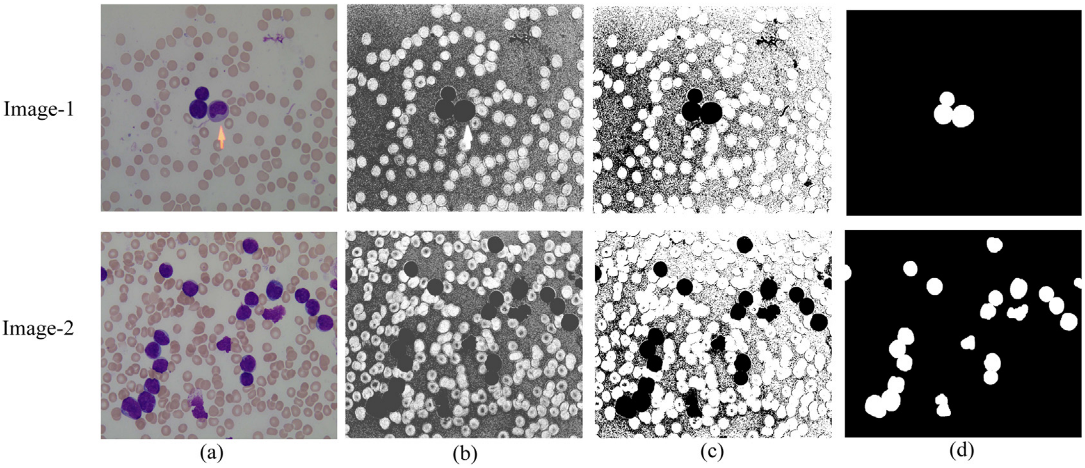

Fig. 2. Identification of grouped leukocytes: (a) input image, (b) histogram equalized image, (c) enhanced image using Wiener filtering, (d) thresholding image

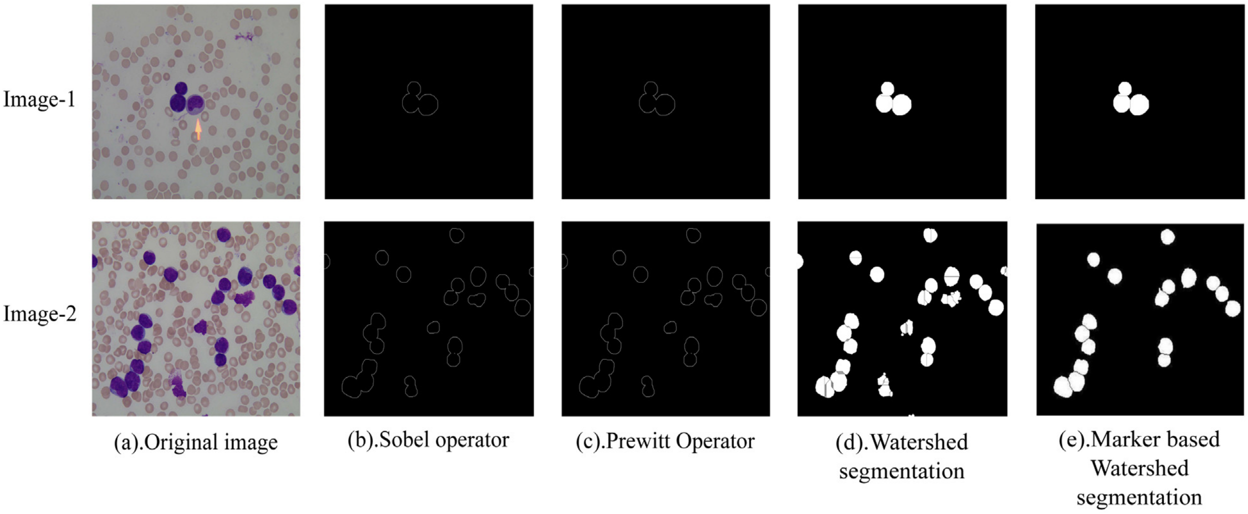

Fig. 3. Results for segmentation by applying different operators

segmentation scheme. As observed from Fig. 2 (d) for both of the image samples, most of the cells are separated, but a few of them are in touch with each other. In such cases, segmentation schemes using Prewitt, Sobel or Watershed fail to differentiate the cells accurately (Fig. 3(b)–(d)). The conventional watershed algorithm is used for the separation of grouped cells which uses distance transform. Along these lines, the separation yields wrong results in the presence of irregular shape cells whereas the utilization of separation performs well in the presence of grouped leukocytes with a perfectly round cell.

To decrease the effect of over-segmentation due to the irregular shape of the cells, marker-based watershed transformations have been suggested here. This scheme is robust and flexible to segment objects, and it consists of two steps namely, extraction of markers and imposition of markers to the gradient image. The basic steps followed are:

1 Define external and internal markers of binary microscopic image
2 Draw a circle around the external marker that generates the region of interest (ROI)
3 Determine the internal marker using Canny operator, followed by thresholding and morphological operation
4 Impose the markers into gradient image to obtain the segmented cells
5 Apply bounding box to each cell to obtain lymphocyte sub-image

The segmentation results using marker-based watershed segmentation is shown in Fig. 3(e). The next step to image segmentation is to clean the image. Image cleaning requires the evacuation of the considerable number of leukocytes situated on the edge of the image and all the irregular parts present in the image. The abnormal components present in an image can be removed by measuring the solidity value and is defined as,

$$soliditiy = \frac{area}{convex\_area} \tag{1}$$

which is used as a threshold for each leukocyte to be included for classification. In the present case, the threshold value chosen as 0.95. So all the objects having less than 0.95 threshold values are excluded from the images.

After extracting the leukocytes from the peripheral blood smear, the next level segmentation is performed, which selects the nucleus and cytoplasm. The nuclei are having a better contrast than the cytoplasm and can be extracted by combing the binary image obtained from the green component and the a* component [21]. Finally, to extract cytoplasm, a subtraction operation is performed from the whole leukocyte and the image containing the nucleus. A stepwise representation of this step is shown in Fig. 4.

The overall schematic diagram of our proposed feature extraction and classification method is shown in Fig. 5 and the steps for calculating feature matrix using GLCM is described in Algorithm 1.

3.3. Feature analysis using GLCM

Algorithm 1. Feature Extraction Algorithm

```
Require: Total number of images derived from the segmentation step.
   GLCM: Gray level co-occurrence matrix
   NGLCM: Normalized gray level co-occurrence matrix
   $\theta$: direction parameter taken as $0^\circ$, $45^\circ$, $90^\circ$, and $135^\circ$
   $p$: number of directions (4)
   $d$: distance parameter ($d$:1)
   $s$: number of feature descriptors
   $m$: total number of features
Ensure Feature_matrix: $X[1:n, 1:m]$
   graycomatrix() computes the GLCM matrix from the nucleus and
cytoplasm region of segmented lymphocytes.
1: Initialize the value of s
2: $m \leftarrow p \times s$
3: For $i \leftarrow 1$ to $n$
4: Calculate the GLCM matrix using graycomatrix() for the input image (IP)
5: For $k \leftarrow 1$ to $p$
6: $GLCM_{\theta_k} \leftarrow graycomatrix(IP, \theta_k, d)$
7: Compute NGLCM from the GLCM matrix
   $NGLCM_{\theta_k} \leftarrow GLCM_{\theta_k}$/sum(elements of $GLCM_{\theta_k}$)
8: For $a \leftarrow 1$ to $s$
   compute the $GLCM_{\theta_k}$ and $NGLCM_{\theta_k}$ and append it to $X$
9: end for
10: end for
11: end for
```

The gray level co-occurrence matrix (GLCM) is used to extract texture feature from the input image. The GLCM assigns a distribution of each gray level pairs of neighboring pixels in an image [24]. The co-occurrence matrix of the ROI is useful in the classification process. GLCM makes use of the spatial relationship between two pixels, and one is a reference pixel, and the other is the neighboring pixel. Let $y(i, j)$ be the element of GLCM of a given image f of size $X \times Y$ having gray levels L ranging from 0 to $L-1$. Then $y(i, j)$ can be defined as a matrix element and can be given by,

$$y(i,j) = \sum_{p=1}^{X}\sum_{q=1}^{Y} \begin{cases} 1, & \text{if} f(p,q) = 1 \text{and} f(p+\Delta p,\ q+\ \Delta q) = j \\ 0, & \text{otherwise} \end{cases}$$

where $(p,\ q)$ and $(p+\Delta p,\ q+\Delta q)$ denotes the reference and neighboring pixel locations respectively. $(\Delta p,\ \Delta q)$ represents the distance of an element $y(i, j)$ in GLCM having gray level values i and j respectively. It can be represented as $y(i,\ j|D,\ \theta)$, where D is the distance of separation between two pixels(reference and neighboring) and θ is the direction of neighboring pixel from the reference pixel. For the microscopic images, the direction parameter is taken as 0°, 45°, 90° and 135° and the distance parameter can be taken as the integral multiple of gray levels. For texture classification of microscopic images, each entry of the GLCM should contain a probability value. For this purpose, a normalized GLCM can be formed the original GLCM. Every component of the standardized gray level co-occurrence matrix can be characterized by,

$$x(i,j) = y(i.j) / \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} y(i,j) \tag{2}$$

The feature matrix generated is of very large dimension and will be computationally expensive if used directly. To retain the relevant features, a probabilistic principal component analysis (PPCA) has been employed [25]. PPCA has advantages over PCA [26] to examine the missing data values. When PPCA is applied to the *feature _m atrix* (X) of size $n \times m$, a reduced feature matrix of size $n \times z$ is obtained having z features, where $z < m$. The feature reduction algorithm is listed in Algorithm 2.

Algorithm 2. Feature reduction using PPCA

```
Require: $X[1:n, 1:m]$: Feature_matrix, ($n$: total number of samples, $m$:
total number of features) Ensure: $X_R[1:n, 1:z]$: Reduced feature matrix
1: State Function ppca()calculates the the principal co-efficients in the
reduced space
2: State Choose $z$
3: State Create an empty matrix $X_R[1:n, 1:z]$
4: State $X_R[1:n, 1:z] \leftarrow ppca(X, z)$
5: State Return $X_R$
```

3.4. RF based classification

Random forest (RF) is one of the most widely used and powerful machine learning techniques, which has shown a higher accuracy rate among recent machine learning algorithms. It is suitable for training large set of data with a guarantee of estimating the most appropriate features required for classification. RF is a collection of tree-structured classifiers where each tree depends on the values of a random vector sampled independently and the distribution of all trees in the forest [27]. Certain points that need to be considered for the construction of the each tree in the forest from a training dataset are shown as follows:

1 If there are N number of samples present in the dataset, then the trees are made up by drawing N samples randomly along with replacement of samples in each turn from the original dataset.
2 If there are R features present in the dataset, then $r < R$ are defined, for each node in the tree r features are selected out of R features that are required to split the node. The value of r is kept fixed through out the development of the forest.
3 Then each tree is grown to its possible extent.

After constructing a forest, an input vector is classified based on the majority of votes as an outcome of all the trees in ensemble. The detailed steps of the classification process are described in Algorithm 3.

Algorithm 3. RF Classifier

```
Require S: Training dataset with $N$ samples, $S = (x_i, y_i)$ for $i = 1, 2, \ldots, N$ and
$x_i \in X_R$ with class labels $y_i \in Y$   $X_R$ and $Y$ represents the input and output
class respectively.   $x$: testing samples   $T$: Number of trees   $r$: input
vector used for the construction of trees in the forest Ensure $D_{fin}$: Final
classifier decision
1: for $i \leftarrow 1$ to $T$
2:    $S_t \leftarrow bootstrapSamples(S)$
3:    $S_t \leftarrow buildTreeClassifier(S_t, r_t)$
4: end for
5: Return $D_{fin}$ based on a majority voting
```

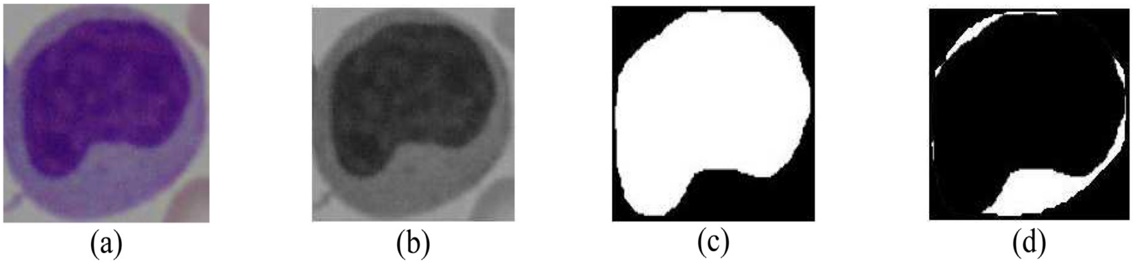

Fig. 4. Separation of nucleus and cytoplasm (a) segmented lymphocyte image, (b) corresponding gray scale image, (c) nucleus sub-image, (d) cytoplasm sub-image

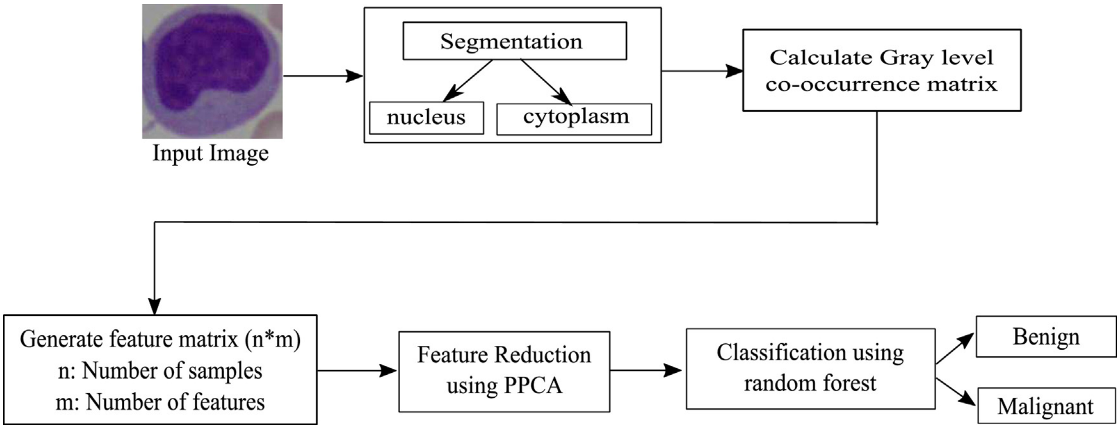
Input Image
Segmentation
nucleus
cytoplasm
Calculate Gray level co-occurrence matrix
Generate feature matrix (n*m)
n: Number of samples
m: Number of features
Feature Reduction using PPCA
Classification using random forest
Benign
Malignant

Fig. 5. Block diagram of classification of lymphoblasts

4. Experimental results and discussions

To evaluate the efficacy of our the proposed scheme (GLCM+PPCA+RF), experiments have been carried out on ALL-IDB1 dataset [22]. Each microscopic images are of size 1712 × 1368 and are pre-processed to enhance the quality of the images. The proposed scheme have been simulated using Matlab and the analysis is made on parameters as defined,

$$Sensitivity(TPR) = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity(TNR) = \frac{TN}{TN + FP} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where, *TP*(True Positive): correct classification rate for positive classes; *TN*(True Negative): correct classification rate for negative classes; *FN*(False Negative): incorrect classification rate for positive classes; *FP*(False Positive): incorrect classification rate for negative classes

To make the classifier more steady and reliable, a 5-fold cross validation (CV) is used. To train the random forest classifier, out of 799 cells in the ALL-IDB1 dataset, 421 malignant and 257 benign cases have been considered. Similarly, for testing total of 121 images (75 malignant and 46 benign) have been used. In the following, the results of different stages of the proposed method are presented. In the present work, there are two different contributions related to segmentation, feature extraction and reduction, and classification. The overall experimental study has been divided into two different experiments and is discussed below in detail.

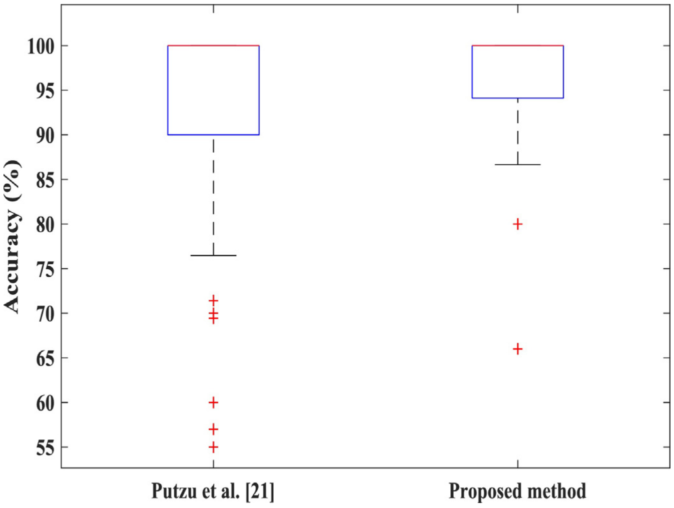

Fig. 6. Performance comparison of the two methods using box plot

4.1. Experiment 1: Performance evaluation of marker-based segmentation

The proper feature extraction depends on the extraction of leukocytes from the microscopic image. To achieve this objective, marker-based watershed segmentation is applied to alleviate the limitation of conventional watershed segmentation. The performance result of manual count by the hematologists for WBC in a set of 49 microscopic images are compared with two segmentation schemes, i.e., method by Putzu *et. al* [21] and proposed marker-based segmentation. It is observed from the experiments that, the proposed segmentation approach can detect 520 lymphocytes out of 540 cells accurately with an accuracy of 96.29% whereas the method proposed by Putzu *et al.* [21] provide an overall accuracy

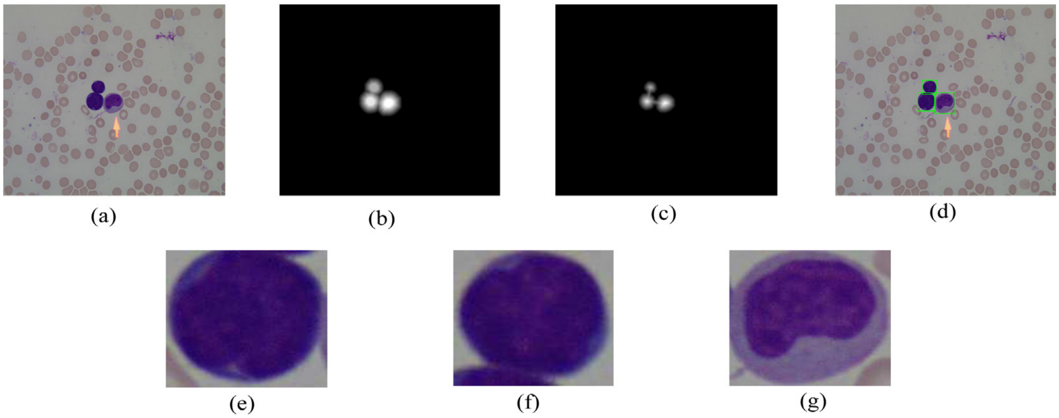

Fig. 7. Extraction of Lymphocytes: (a) input image, (b) image after external marker, (c) image after internal marker, (d) segmented lymphocyte, (e, f, g) individual lymphocyte

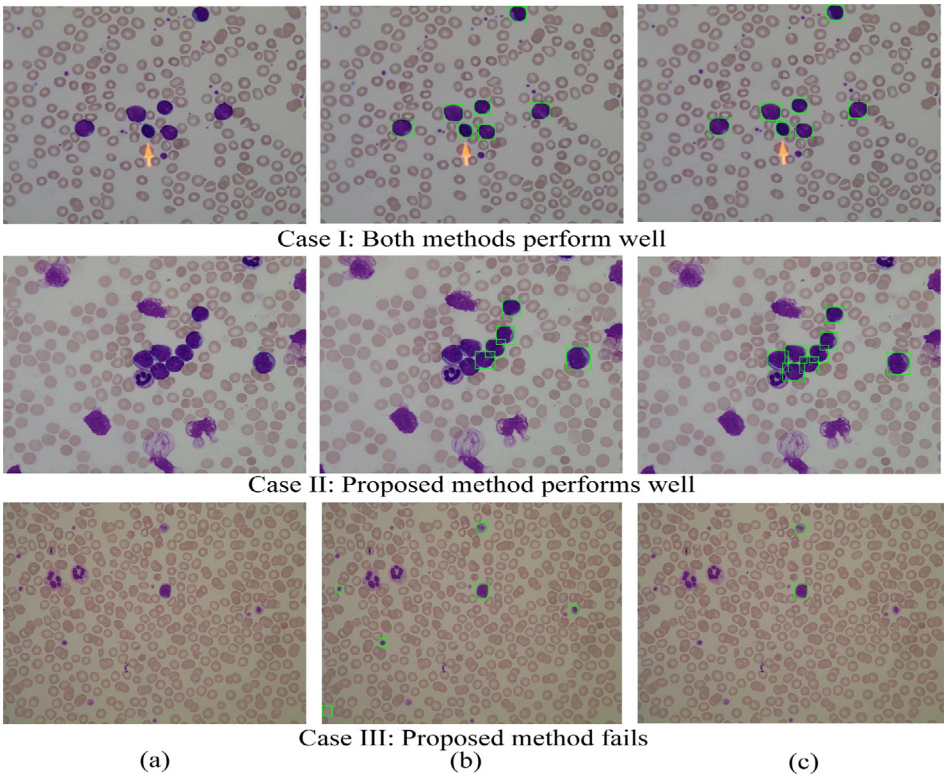
Case I: Both methods perform well
Case II: Proposed method performs well
Case III: Proposed method fails
(a)
(b)
(c)

Fig. 8. Segmentation results of two methods over 3 different cases (a) input image, (b) Putzu et al. [21], (c) Proposed method

of 88.36%. The overall comparison between these two methods is shown in Fig. 6 using a box plot. From the figure, it can be seen that the proposed scheme performs well than the existing scheme.

The extracted lymphocytes from a given image are shown in Fig. 7. Fig. 7 (b) represents the image having an external marker which helps in forming the line of separation between two cells. Fig. 7 (c) deals with identifying an internal marker that represents the minima of different cells. These two markers, i.e., internal and external assist in finding the separation between the connecting cells. Both the markers are then imposed into the gradient of the image to get the lymphocyte from the image. After identifying the lymphocytes, the bounding box is applied to the whole image to get the lymphocyte sub-image. Even though the marker-based segmentation scheme is capable of separating grouped and connected cells in general, in some cases it fails to do so. Fig. 8 depicts the different cases where both the existing and proposed schemes work well and where both fails. Hence, the proposed scheme can further be improved.

4.2. Experiment 2: Performance evaluation of proposed GLCM+PPCA+RF scheme

The proposed work employs GLCM matrix to extract texture features from the leukocyte sub-images. Therefore, 80 number of Harlick features are extracted from the nucleus and cytoplasm region. PPCA (Algorithm 2) is applied to get a reduced set of features which can classify the disease with higher accuracy. After applying the feature reduction scheme, the number of features gets reduced to 40. These are the first 40 principal components (PCs) which can preserve approximately 85% of the total variance in the dataset. The reduced features are 50% of the original features. Fig. 9 shows the plot of cumulative variance on the number

Table 1
Performance measure of different classifiers over 5-fold for nucleus feature.

Classifier	Nucleus				
	Sensitivity (%)	Specificity (%)	Precision (%)	F-Score (%)	Accuracy (%)
k-NN (*k*= 5)	100.00	96.15	97.40	98.68	97.52
BPNN	100.00	97.40	98.68	99.33	98.30
SVM-P	100.00	90.36	94.93	97.40	93.888
RF	100.00	98.08	97.80	98.9	99.004

Table 2
Performance measure of different classifiers over 5-fold for cytoplasm feature

Classifier	Cytoplasm				
	Sensitivity (%)	Specificity (%)	Precision (%)	F-Score (%)	Accuracy(%)
k-NN (*k*= 5)	94.54	98.55	98.61	95.94	96.67
BPNN	91.50	97.30	98.57	95.17	95.00
SVM	88.23	89.04	92.85	89.65	80.833
RF	86.50	99.56	99.60	96.00	96.00

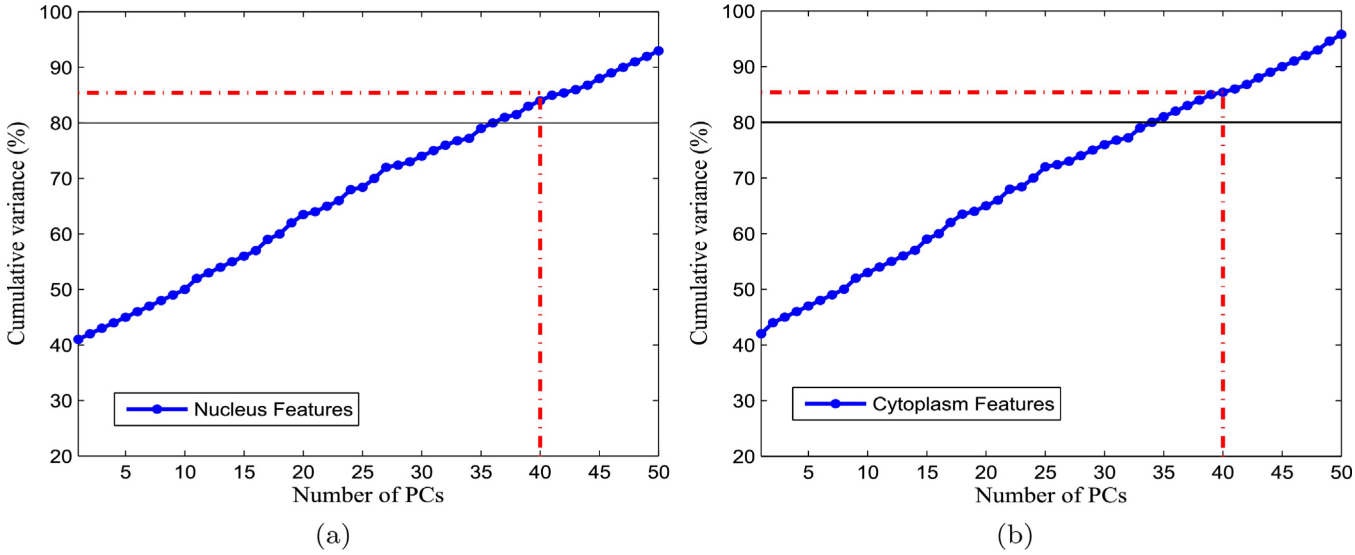

Fig. 9. Cumulative variance (%) with respect to number of PCs (a) for nucleus features, (b) for cytoplasm features

Table 3
Performance analysis of 5-fold CV Procedure with nucleus feature using RF classifier

Fold	Testing Instances	TP	FN	TN	FP	Accuracy (%)
Fold-1	121	75	0	45	1	99.17
Fold-2	121	75	0	45	1	99.17
Fold-3	121	75	0	46	0	100.00
Fold-4	121	75	0	44	2	98.34
Fold-5	121	75	0	44	2	98.34
						99.004

Table 4
Performance analysis of 5-fold CV Procedure with cytoplasm feature using RF classifier

Fold	Testing Instances	TP	FN	TN	FP	Accuracy (%)
Fold-1	121	68	7	46	0	94.17
Fold-2	121	69	6	46	0	95.00
Fold-3	121	71	4	45	1	95.83
Fold-4	121	73	2	46	0	98.33
Fold-5	121	71	4	46	0	96.67
						96.00

of PCs for nucleus and cytoplasm respectively. Now, the reduced feature set is used for the classification using classifiers like RF, *k*-NN, BPNN, and SVM. The performance of the proposed scheme for both the nucleus and cytoplasm feature is compared with other standard classifiers like *k*-NN, BPNN, and SVM-P, which are listed in Table 1 and Table 2. From the tables, it can be seen that the proposed scheme gives better results regarding sensitivity, specificity, precision, f-score, and accuracy. The performance of the proposed

Table 5
Classification performance comparison of proposed scheme with some other existing scheme

Authors	Features	Classifier	Number of features	Accuracy(%)
Mohapatra et al. [19]	Morphological+colour+textural	EOC_5	33	94.58
Putzu et al. [21]	Morphological+colour+textural	SVM-P	131	93.2
Proposed method	GLCM+PCA (nucleus)	Random forest	40	98.04
		GLCM+PCA(cytoplasm)	40	92.65
	GLCM+PPCA (nucleus)	Random forest	40	99.004
	GLCM+PPCA (cytoplasm)		40	96.00

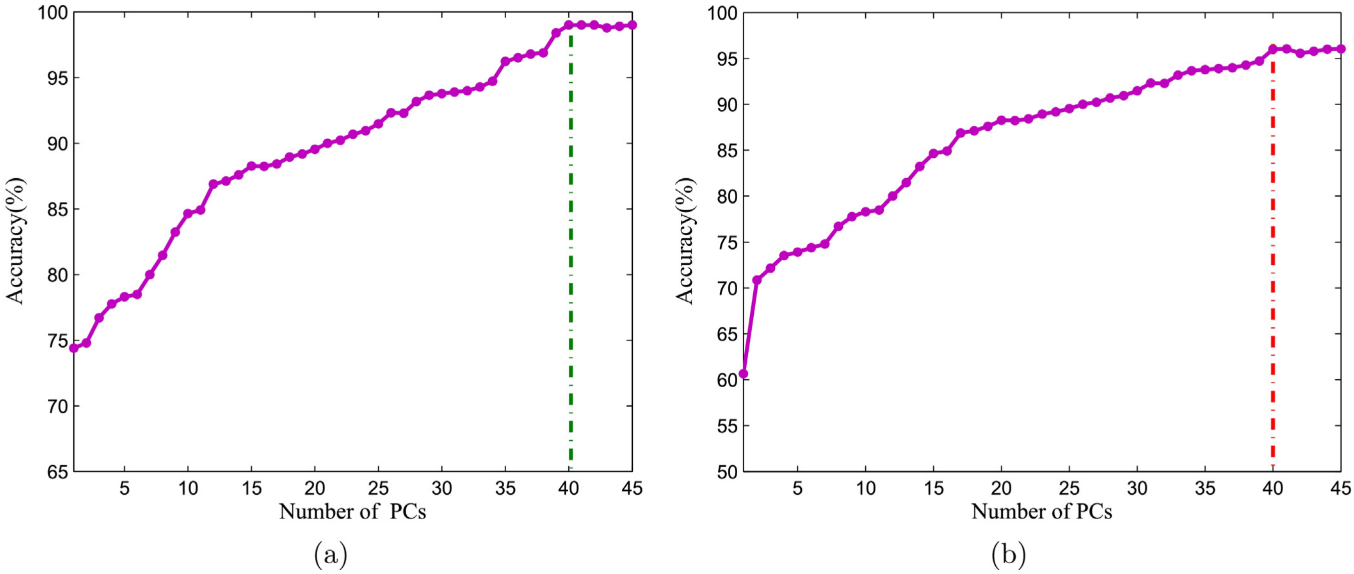

Fig. 10. Performance evaluation in terms of number of PCs (a) for nucleus features, (b) for cytoplasm features

scheme was experimented with varying the number of PCs and their respective accuracies for nucleus and cytoplasm features are shown in Fig. 10. Moreover, it is inferred that nucleus features are more suitable for the correct classification than the cytoplasm features. Table 3 and Table 4 show the 5-fold CV procedure for the test instances of nucleus and cytoplasm features using RF classifier respectively. Further, PCA has been applied separately which results in 98.04% and 92.65% of accuracy for nucleus and cytoplasm respectively. These results are determined using the same number of PCs.

Finally, a comparative analysis has been made on the number of features and accuracy obtained by various schemes and given in Table 5. It is inferred that the method proposed in [21]give 93.2% of accuracy with a large feature set. The number of features can be significantly reduced by the method proposed by [19] with a greater accuracy. However, due to the ensembling technique used for classification, the time complexity is very high. Our proposed scheme uses 40 features to represent each image while producing the highest classification accuracy. The classification accuracy obtained for ALL-IDB1 Dataset is 99.004% and 96.00% respectively.

Further, the computation time required for segmentation, feature extraction, feature reduction, and classification for a lymphocyte are 0.23s, 0.098s, 0.016s, and 0.002s respectively. The total computation time taken for each lymphocyte is found to be 0.346s. Similarly, while comparing the classification performance for the ALL-IDB1 dataset (Table 6), it is observed that RF classifier has lesser time requirement as compared to BPNN and SVM-P.

Table 6
Computational time for different classifier for the detection of lymphoblast

Classifier	Computation time (s)
k-NN	0.260
BPNN	3.67
SVM-P	55.814
RF	1.851

5. Conclusion

In this paper, we have proposed a novel scheme for the automatic identification and classification of lymphoblasts from the microscopic image to support the decision of hematologists. An efficient segmentation scheme is employed in the identification of lymphoblasts by separating it from the microscopic image. A new method of extracting the textural feature from the nucleus and cytoplasm region of the cropped image is described using gray level co-occurrence matrix. To validate the efficacy of the proposed scheme, it is being tested on publicly available dataset ALL-IDB1. The overall accuracy of the proposed segmentation scheme is found to be 96.29%. The experimental outcome of the proposed method shows that, for texture feature analysis the nucleus region gives a better accuracy than the cytoplasm region. The achieved accuracy for the nucleus feature is found to be 99.004% whereas the same for the cytoplasm is 96 %. Further color and morphological features can be used for performance improvement, and the validation needs to be done on a larger dataset. Also, the work can be extended towards the sub-classification of acute lymphoblastic leukemia. Another area of research can be taken into account where the proposed scheme can be extended to acute myeloid leukemia (AML).

References

[1] R. Siegel, D. Naishadham, A. Jemal, Cancer statistics, 2013, CA Cancer J. Clin. 63 (1) (2013) 11–30.

[2] S. Pelengaris, M. Khan, The Molecular Biology of Cancer: A Bridge From Bench to Bedside, John Wiley & Sons, 2013.

[3] J.M. Bennett, D. Catovsky, M.-T. Daniel, G. Flandrin, D. Galton, H.R. Gralnick, C. Sultan, Proposals for the classification of the acute leukaemias French-American-British (FAB) co-operative group, Br. J. Haematol. 33 (4) (1976) 451–458.

[4] K.P. Kulkarni, R.S. Arora, R.K. Marwaha, Survival outcome of childhood acute lymphoblastic leukemia in India: a resource-limited perspective of more than 40 years, J. Pediatric Hematol./Oncol. 33 (6) (2011) 475–479.

[5] R. Arora, T. Eden, G. Kapoor, et al., Epidemiology of childhood cancer in India, Indian J. Cancer 46 (4) (2009) 264–273.

[6] J.H. Wright, The Histogenesis of the Blood Platelets, vol. 3, 1910.

[7] J. Angulo, G. Flandrin, Microscopic image analysis using mathematical morphology: application to haematological cytology, Sci. Technol. Educ. Microsc.: Overview 1 (2003) 304–312.

[8] B.C. Ko, J.-W. Gim, J.-Y. Nam, Automatic white blood cell segmentation using stepwise merging rules and gradient vector flow snake, Micron 42 (7) (2011) 695–705.

[9] S. Mohapatra, D. Patra, S. Satpathy, Unsupervised blood microscopic image segmentation and leukemia detection using color based clustering, Int. J. Comput. Inf. Syst. Ind. Manage. Appl. 4 (2012) 477–485.

[10] S. Mohapatra, D. Patra, S. Kumar, S. Satpathy, Lymphocyte image segmentation using functional link neural architecture for acute leukemia detection, Biomed. Eng. Lett. 2 (2) (2012) 100–110.

[11] Z. Liu, J. Liu, X. Xiao, H. Yuan, X. Li, J. Chang, C. Zheng, Segmentation of white blood cells through nucleus mark watershed operations and mean shift clustering, Sensors 15 (9) (2015) 22561–22586.

[12] S. Arslan, E. Ozyurek, C. Gunduz-Demir, A color and shape based algorithm for segmentation of white blood cells in peripheral blood and bone marrow images, Cytometry A 85 (6) (2014) 480–490.

[13] H.T. Madhloom, S.A. Kareem, H. Ariffin, An image processing application for the localization and segmentation of lymphoblast cell using peripheral blood images, J. Med. Syst. 36 (4) (2012) 2149–2158.

[14] Y. Zhang, S. Wang, P. Phillips, Z. Dong, G. Ji, J. Yang, Detection of Alzheimer's disease and mild cognitive impairment based on structural volumetric MR images using 3D-DWT and WTA-KSVM trained by PSOTVAC, Biomed. Signal Process. Control 21 (2015) 58–73.

[15] Y. Zhang, L. Wu, Optimal multi-level thresholding based on maximum Tsallis entropy via an artificial bee colony approach, Entropy 13 (4) (2011) 841–859.

[16] F. Scotti, Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images, in: IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, 2005, pp. 96–101.

[17] L. Gupta, S. Jayavanth, A. Ramaiah, Identification of different types of lymphoblasts in acute lymphoblastic leukemia using relevance vector machines, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2009, 2008, pp. 6675–6678.

[18] H.J. Escalante, M. Montes-y Gómez, J.A. González, P. Gómez-Gil, L. Altamirano, C.A. Reyes, C. Reta, A. Rosales, Acute leukemia classification by ensemble particle swarm model selection, Artif. Intell. Med. 55 (3) (2012) 163–175.

[19] S. Mohapatra, D. Patra, S. Satpathy, An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images, Neural Comput. Appl. 24 (7-8) (2014) 1887–1904.

[20] L. Faivdullah, F. Azahar, Z.Z. Htike, W.Y.N. Naing, Leukemia detection from blood smears, J. Med. Bioeng. 4 (6) (2015).

[21] L. Putzu, G. Caocci, C. Di Ruberto, Leucocyte classification for leukaemia detection using image processing techniques, Artif. Intell. Med. 62 (3) (2014) 179–191.

[22] ALL-IDB dataset for ALL classification, http://crema.di.unimi.it/fscotti/all/.

[23] R.D. Labati, V. Piuri, F. Scotti, All-idb: the acute lymphoblastic leukemia image database for image processing, in: 2011 18th IEEE International Conference on Image Processing, IEEE, 2011, pp. 2045–2048.

[24] R.M. Haralick, K. Shanmugam, I.H. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 3 (6) (1973) 610–621.

[25] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 61 (3) (1999) 611–622.

[26] Y. Zhang, L. Wu, An MR brain images classifier via principal component analysis and kernel support vector machine, Prog. Electromagnet. Res. 130 (2012) 369–388.

[27] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.