Biomedical Signal Processing and Control 47 (2019) 303–311

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection

Sonali Mishra*, Banshidhar Majhi, Pankaj Kumar Sa

Pattern Recognition Laboratory, Department of Computer Science and Engineering, National Institute of Technology, Rourkela 769008, India

ARTICLE INFO

Article history:
Received 8 June 2017
Received in revised form 21 March 2018
Accepted 8 August 2018
Available online 7 September 2018

Keywords:
Acute lymphoblastic leukemia (ALL)
Discrete orthonormal S-transform (DOST)
AdaBoost with RF (ADBRF)
Computer-aided diagnosis (CAD)

ABSTRACT

This paper presents an effective scheme for classification of the normal white blood cells from the affected cells in a microscopic image. The proposed method initially pre-processes the input images using Y component of the CMYK image and a triangle method of thresholding. Subsequently, it utilizes discrete orthonormal *S*-transform (DOST) to extract the texture features, and its dimensionality is reduced using linear discriminant analysis. The reduced features are then supplied to the proposed Adaboost algorithm with RF (ADBRF) classifier where the random forest is used as the base classifier. A publicly available dataset, ALL-IDB1 is used to validate the proposed scheme. The simulation results based on the five runs of *k*-fold stratified cross-validation indicate that the proposed method yields superior accuracy (99.66%) as compared to existing schemes.

1. Introduction

The visual examination of blood samples is a major criterion for the analysis of leukemia [1,2]. Acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) are the two different types of Leukemia which can lead to death if not treated at the right time. AML affects the myeloid organs, whereas ALL is seen in the bone marrow. ALL is a significant hematopoietic disease which is generated by the abnormal collection of white blood cells (WBCs). With the increase in the number of malignant WBCs, the fighting capability of the body with the foreign material gets diminished. The early detection of ALL can considerably improve the probability of recovery, especially in the case of children [3,4]. The recognition of blast(unhealthy WBCs) cell in the bone marrow is also an important step for the detection of ALL. The percentage of blasts is a major concern for detecting the proper stage of the ALL and is also helpful in the proper treatment of the patients. According to French-American-British standard (FAB) [5], three distinct types of ALL are characterized based on the morphological differences among the lymphoblast [6].

So far, the detection of the disease highly depends upon the perfection of the hematologists and pathologists. To support the hematologists, a computer-aided diagnosis (CAD) is a basic necessity for accurate classification and early detection of ALL. The main step in a CAD system is to generate features of WBCs which will classify the cells as healthy or affected. The most identifiable feature of a normal blood cell can be categorized as morphological, statistical, and textural features. Further, it is also necessary to classify different blast types. A typical blood smear having a lymphocyte (healthy) and a lymphoblast (affected) are given in Figure 1. In this paper, we have proposed a texture based feature in microscopic images using Discrete Orthonormal *S*-Transform (DOST). The suggested feature defines the characteristic of an image as rough, smooth, silky, or bumpy as a function of the spatial variation in pixel intensities.

Rest of the article is organized as follows. Section 2 describes the related work. Section 3 outlines the proposed method for the detection of ALL. Section 4 presents the details about the data source and gives a relative comparison of the proposed methodology. Finally, the conclusion is given in Section 5.

2. Related work

In recent years, many researchers have been working on the development of CAD systems. Investigations have been made for the detection of lymphoblasts cells in microscopic images. They have taken into consideration for various morphological, textural, and color features for the detection of the disease. Those features are then classified using different classifiers. The computer-assisted discovery and analysis techniques can be broadly divided into two categories. The first category applies the genetic information, while

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

https://doi.org/10.1016/j.bspc.2018.08.012
1746-8094/

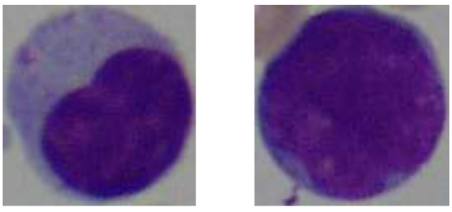

Fig. 1. Representation of a peripheral blood smear having a lymphocyte and a lymphoblast.

the second one uses the information present in the image modeled by different machine learning techniques.

Lin et al. [7] have suggested an approach for prediction of tumor in the microscopic blood images. They have used genetic algorithms for feature selection and silhouette statistics to differentiate between six subtypes of ALL. They have shown a better accuracy rate by using microarray data and gene expression. The testing accuracy using 23 genes are found to be 100%, and CFS/SVM only performed an accuracy of 96% with the help of more than 20 predictive genes. Zong et al. [8] have proposed a technique for the detection of lymphoblasts based on flow cytometer data and reported a classification accuracy of 96.67%. But it is very difficult to extract the gene expression from the bone marrow samples, and it requires very sophisticated equipment. Ross et al. [9] have suggested an scheme to differentiate different types of pediatric ALL. They have analyzed blasts from 132 samples using higher density nucleotide arrays. A set of newly selected genes is incorporated into class predicting algorithms to get an overall accuracy of 97%.

Escalante et al. [10] have suggested an application to the problem of acute leukemia detection. They have classified the subtypes of acute leukemia using an ensemble particle swarm optimization selection technique (EPSMS). The classification of acute leukemia based on morphological features can be done by building ensembles. For a 2 class classification, the authors have got an accuracy of 97.68% whereas the accuracy got decreased to 94.21% for multiclass classification. Foran et al. [11] have proposed a framework for differentiating lymphoma and leukemia. They have found an accuracy of 83% on 19 different cases of leukemia and lymphoma. Though they have differentiated leukemia and lymphoma, the suggested method has not been validated in ALL and on a larger dataset. Scotti et al. [12] have suggested distinguishing different WBCs or leukocyte by examining the morphological properties of a color image. The proposed system first distinguishes between the WBC and other components of blood cells. This procedure provides satisfying results in finding different components which show a way for identification of tumor deformation in the cell morphology. The authors have taken a dataset of 134 images containing leukocytes. For differentiating different types of WBCs, they have used parallel FF-NN to get an accuracy of 92% with a feature size of 23. In another work, Scotti [13] suggested a scheme for detection of ALL from the microscopic images. The experiments are being conducted on 150 images and shown that morphological features are more feasible for lymphoblast recognition for the detection of ALL with a classification error 0f 0.0133 using feed-forward neural network.

Halim et al. [14] have presented an approach to count the number of blasts in the case of ALL. They have taken histogram based thresholding technique succeeded by S-component on the HSV space. Subsequently, morphological erosion is performed for counting the blast cells. The overall accuracy of the proposed system is found to be 97.8% with a very small dataset consisting of 50 images. Also, the authors have not specified the threshold value used for separating nucleus and cytoplasm. Mohapatra et al. [15] have recommended an ensemble of classifier system in which accuracy has been improved by analyzing morphological and textural features from the peripheral blood smear having an accuracy of 99% with ALL-IDB1 dataset. Putzu et al. [16] have proposed an approach which isolates the whole leukocyte from a microscopic image and subsequently separates the nucleus and cytoplasm. For every cell, distinct features like shape, color, and texture are extracted and are used to train different classification models to determine the best one for leukemia classification. The authors have found the accuracy of 93.2% with the help of 131 features. Angulo et al. [17] utilized watershed segmentation for lymphocyte identification, where morphological properties are extracted for characterizing lymphocytes in light of cell typology. Though this technique shows exact results for segmentation, it has not been used for classification.

It has been observed that morphological, textural, and color-based features are predominant while classifying lymphoblasts. Among classifiers, ANN and SVM have been widely used. In this paper, we have proposed a lymphoblast classification scheme using features extracted from discrete orthonormal S-transform (DOST) followed by feature reduction using a hybrid approach which includes PCA+LDA. Finally, the discriminant features so obtained are passed to Adaboost Random Forest classifier. The proposed model uses 2D discrete orthonormal S-transform (2D-DOST) as the feature extractor. The 2D-DOST is based on a set of orthonormal basis function that preserves both time-frequency and phase information of a signal. Thus, DOST gives features with zero information redundancy.

3. Proposed work

The proposed methodology has different phases like any other classification scheme which include the preprocessing, sub imaging, feature extraction, feature reduction, and classification. We have utilized the standard preprocessing techniques like noise reduction and smoothing of background. The main contribution lies on DOST based feature extraction and PCA+LDA based feature reduction. The relevant features are subsequently used for classification on an AdaBoost based random forest (ADBRF) classifier. The block diagram of the proposed scheme is depicted in Figure 2. The phases are discussed below in sequel.

3.1. Pre-processing, sub-imaging and PBS segmentation

The images from the ALL-IDB1 dataset have been collected under different magnification and as a result, comprised of noise and background effects. The RBCs and platelets present in the smear are unwanted for the detection of disease. To extract the WBCs from the blood smear, background subtraction is performed. Since the *Y* component of the image contains maximum information regarding the WBC, the original RGB image is regenerated to CMYK color space. Triangle method [18] for thresholding has been used for extracting the WBCs from the background. The overall steps followed in pre-processing are described in Figure 3.

Microscopic blood images are relatively larger in size and consists of more than one WBCs per image. However, the region of interest must contain only one WBC for the detection of ALL. Marker-based watershed segmentation [19] is used to separate the grouped cells. The use of marker-based watershed segmentation results in separating the grouped leukocytes by imposing the markers on the blood smear. The result for marker-based watershed segmentation is depicted in Figure 4. The images taken for this experiment is a combination of normal and abnormal images present in the ALL-IDB1 dataset. A more detailed explanation of the proposed marker-based watershed segmentation scheme can be found in our previous work [19].Though all the WBCs require

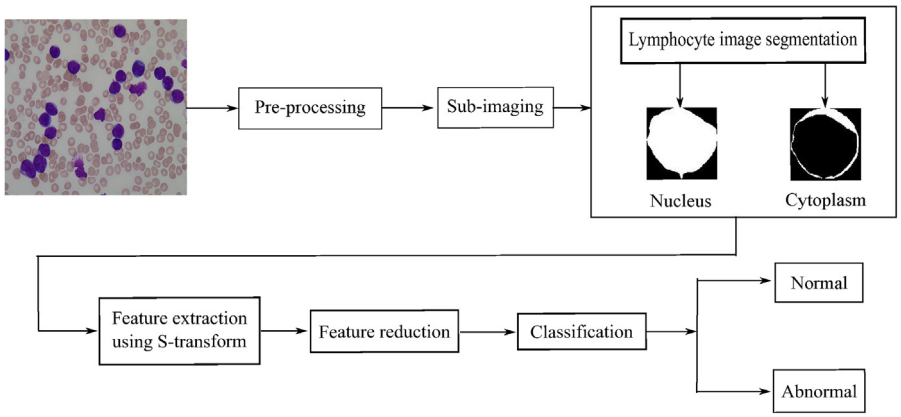

Fig. 2. Block diagram of an automated diagnosis of ALL.

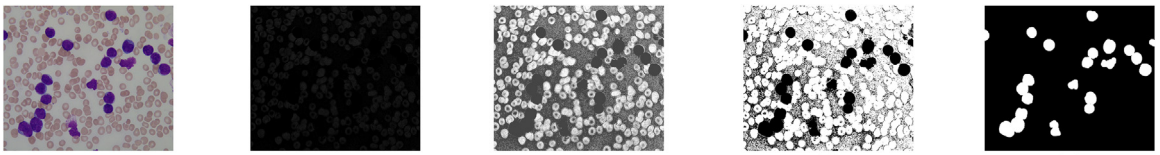

Fig. 3. Steps for pre-processing: (a) input image, (b) corresponding Y component, (c) histogram equalized image, (d) thresholding image, (e) image after background removal operation.

being examined for the detection of disease, bounding box method is applied to crop the image around each WBC [20]. The next process is to clean the image that helps in removing some WBCs present on the edge of the image as well as the number of abnormal components. It can be done by evaluating the solidity of an object. Solidity measures the density of the object. A solidity value of 1 signifies a solid object whereas the solidity value less than 1 signifies the presence of holes in the object. The solidity value is calculated for each WBC (leukocyte) sub-images obtained from the marker-based watershed segmentation. It can be defined as,

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

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. To separate the nucleus and cytoplasm Shadowed C-means (SCM) clustering-based approach is used which can easily separate the nucleus and cytoplasm. SCM clustering is used to classify each pixel into one of the three regions, i.e., cytoplasm, nucleus, and background [15].

3.2. Feature extraction

Discriminative features from the segmented nucleus and cytoplasm region can be found out and given to classifier for differentiation between lymphocyte (healthy) and lymphoblasts (unhealthy). The criteria for diagnosis depends on the number of lymphoblasts reside in the blood smear. More than 20% lymphoblasts in an image show the presence of disease in the blood. A lymphoblast is morphologically defined by a large shape nucleus including an irregular shape and size, the nucleoli are present and prominent, and moreover, the cytoplasm is intensely colored. The changes in the nucleus and cytoplasm region play an important role in finding out the malignancies in blood. Table 1 shows a brief description of the different types of features used to differentiate between a normal and malignant cell.

Table 1
Description of different types of features of lymphocyte and lymphoblast.

Types of features	Computed
Morphological	Nucleus area, cytoplasm area, lymphocyte area, nucleus perimeter, nucleus shape (roundness, form factor, elongation) fractal dimension, contour signature
Textural	Wavelet co-efficients, DCT, GLCM(Gray Level Co-occurrence matrix), GLRLM (Gray Level Run Length Matrix), Fourier transform, S-transform
Colour	region color(Can be computed from the intensity value of different color component in RGB and HSI color space)

3.2.1. Discrete orthonormal S-transform

A two-dimensional S-transform is a multiresolution technique which was developed by Stockwell [21]. DOST is a modified version of S-transform [22] which is used to characterize the texture of an image. A texture feature signifies the intensity variation in an image. Many techniques are there for evaluating the texture feature using co-occurrence matrix, wavelet feature analysis using DWT, etc. The S-transform is nearly related to the continuous wavelet transform with the help of Morlet mother wavelet that gives a computational complexity of $O(N^4 + N^4 log(N))$ and storage compleexity of $O(N^4)$, where $N \times N$ is the image size. The significant limitation of S-transform is its higher complexity (time and space) due to its redundant nature which is eliminated by DOST. DOST utilizes an orthonormal set of basis function to have less computational complexity ($O(N^2 + N^2 log(N))$) and storage complexity ($O(N^2)$).

A 2D-DOST scheme can provide the pixel description of texture by giving the horizontal and vertical frequency spectra. The DOST of a 2D-image in the frequency domain can be calculated by using a dyadic sampling scheme. The steps for calculating the DOST coefficients of a lymphocyte $f(a, b)$ of size $M \times N$ are described as follows:

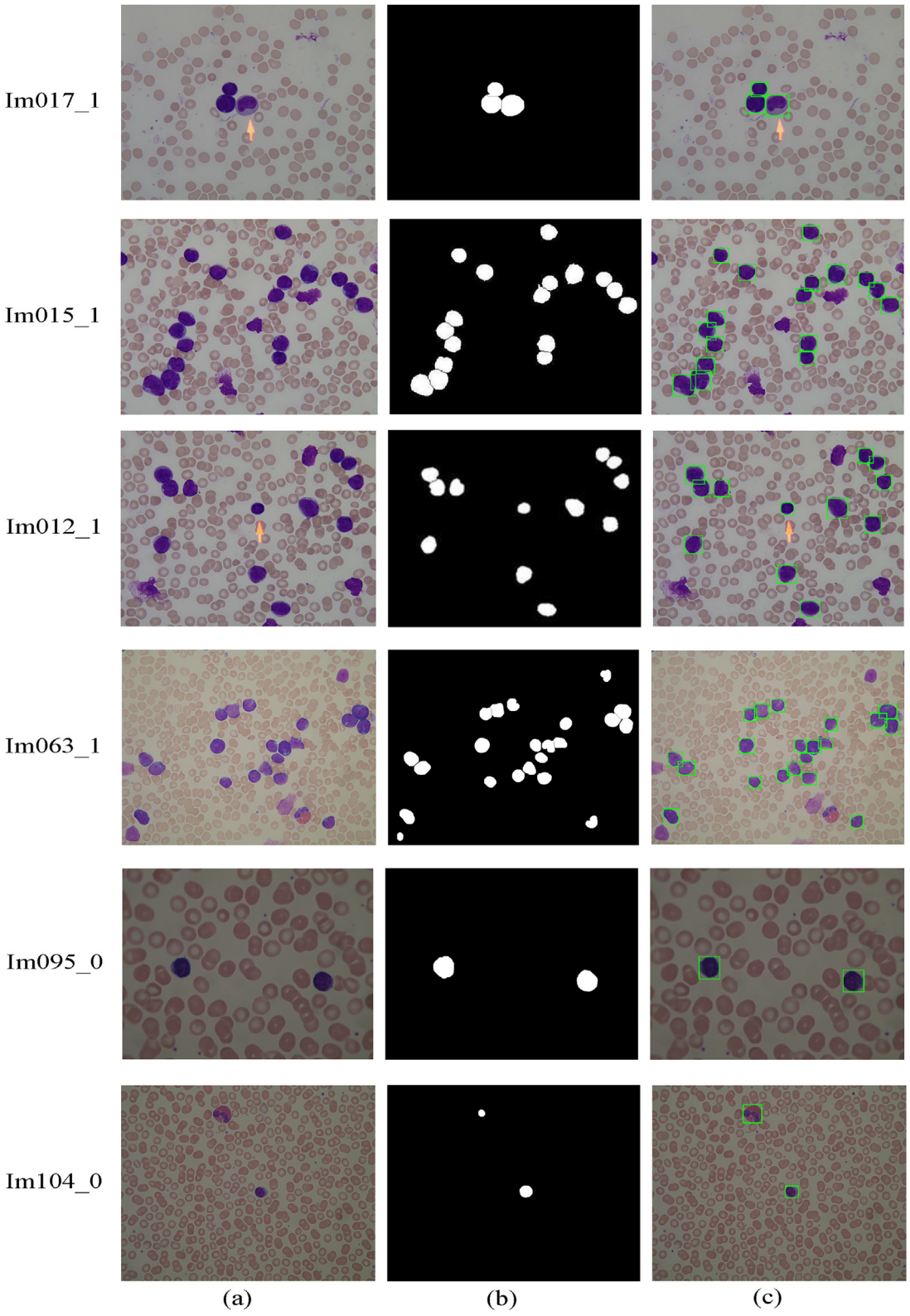

Fig. 4. Results for separation of grouped WBCs (a) original image, (b) image after applying marker-based watershed segmentation, (c) image after WBC separation. (Images have been taken from ALL-IDB1 dataset.)

(a) Perform the Fourier transform to $f(a, b)$ to find the fourier coefficients.

$$F(u, v) = \sum_{a=0}^{M-1} \sum_{b=0}^{N-1} f(a, b) e^{-j2\pi\left(\frac{au}{M} + \frac{bv}{N}\right)} \quad (2)$$

(b) Partition the Fourier coefficients $F(u, v)$ and the square root of the number of points in the partition is multiplied with

it. Again perform an inverse Fourier transform. The calculated voice image can be represented by,

$$S(a', b', u_a, u_b) = \frac{1}{\sqrt{2^{p_a+p_b-2}}} \sum_{u=-2^{p_a-2}}^{2^{p_a-2}-1} \sum_{v=-2^{p_b-2}}^{2^{p_b-2}-1} F(u+u_a, v+u_b) e^{j2\pi}\left(\frac{ua'}{2^{p_a-1}} + \frac{vb'}{2^{p_b-1}}\right) \quad (3)$$

where, $u_a = 2^{p_a-1} + 2^{p_a-2}$ (horizontalfrequency)
$u_b = 2^{p_b-1} + 2^{p_b-2}$ (verticalfrequency)

(c) Obtain the DOST co-efficients and store it in an empty feature matrix *EM*.

The DOST of a $M \times N$ image gives $M \times N$ coefficients, and each coefficient is considered as a feature as they contain the frequency information. So, the total number of voice frequencies generated from the lymphocyte image is equal to the size of the lymphocyte. The details of feature extraction step in described in Algorithm 1.

3.3. Feature reduction

Linear discriminant analysis (LDA) is used for various problems like face recognition, cancer detection, multimedia information retrieval, etc. [23]. The primary objective of LDA is to find the projection F that enhances the relationship between class scatter S_b, and the within-class scatter S_w. It can be written as,

$$argmax(F)\frac{|FS_bF|}{|FS_wF|} \quad (4)$$

The LDA algorithm encounters numerous challenges for a very high dimensional data. The scatter matrices formed are very large. It is a challenging task to manage such huge matrices. Such matrices face a small issue named as small sample size (SSS). In this case, the size of the sample set is less than the dimension of the original feature set. To solve these problems, we have used PCA before LDA approach to lessen the dimension of the feature vector. PCA is applied S_w will no longer be in singular form. After that, LDA can be applied easily to get a reduced feature set as given in equation (5)

$$RFM = (X_{PCA})_{LDA} \quad (5)$$

where X is the input vector (features), and *RFM* is the reduced vector after implementing the reduction algorithm. PCA is a technique that is utilized for several purposes [24]. When PCA is employed, the feature vector gets decreased to a size of $n \times (n-1)$, where n is the total number of samples. Later, the dimension is further reduced by applying LDA. This method creates *RFM*. The steps for feature reduction is given in Algorithm 2.

3.4. Classification

We have proposed a combination of Adaboost and Random Forest for the classification of normal and abnormal cells to improve accuracy. AdaBoost is a new ensemble method used for the prediction in the classification task. It is a learning algorithm that combines many classifiers having a higher error rate and produces another resulting classifier having lower error rate [25,26]. We have investigated the performance of the Adaboost algorithm in which random forest is taken as a weak learner to resolve several classification issues [27]. In this paper Adaboost.M1 [25] algorithm is used. It takes a training dataset S of n samples, i.e. $S = (x_i, c_i)$, $i = 1, \ldots, n$, where each x_i represents the feature vector of the dataset and $c_i \in C = \{1, 0\}$ is the class label. The number 1 and 0 corresponds to the abnormal (positive) and normal (negative) state respectively. A base learning algorithm is used by the AdaBoost algorithm in a series of the cycle. In all the cycles, each training sample is assigned a weight that symbolizes the expectation of the sample being taken for the training set during the experiment. Initially, all the samples are given a fixed weight. After training, the weight is renewed based on the prediction results. Easy samples that are correctly classified by the weak learner get low weights, and hard samples that are the incorrectly classified get higher weights. Hence, it focuses on the samples with more weights in the training set, which seem to be harder for the weak learner. This process will continue for each cycle. AdaBoost at last linearly combines all the weak hypotheses into a single final hypothesis.

Random forest (RF) is one of the most extensively employed and robust techniques, which has recorded a greater accuracy rate among recent ML algorithms. It is suitable for training large set of data with a guarantee of estimating the most appropriate features required for classification. It is a combination of tree-structured classifiers where the value of each tree depends on an arbitrary vector that is sampled separately in the forest [28]. Some points that are needed for the creation of each tree in the forest is described below:

- If there are n samples present in the dataset, then the trees are made up by picking n samples randomly along with replacement of samples in each turn from the original dataset.
- From z number of features present in the dataset, it selects $r < z$ for each node in the tree, which is required to split the node. The value of r is held fixed through out the development of the forest.
- Then each tree is grown to its possible extent.

After building a forest, an input vector is assigned based on the majority of votes as an outcome of all the trees. We have employed random forests as a base learner of AdaBoost algorithm to analyze microscopic images. The detailed steps of classification process are described in Algorithm 3 .

4. Experimental results and discussions

For the analysis of leukemia, peripheral blood smear is taken from an open database ALL-IDB [29] which is again divided into two different types namely, ALL-IDB1 and ALL-IDB2. The ALL-IDB1 database contains 108 images. Among 108 images in ALL-IDB1, 49 images are classified as abnormal, and the rest 59 images are the normal one. The number of candidate lymphoblast presents in that 49 images is approximately found to be 510. And the total number of leukocytes from all 108 images using marker-based watershed segmentation is found as 799 which include both normal and abnormal images. The ALL-IDB2 dataset is produced by cropping the images from ALL-IDB1 having less dimension which is essentially utilized for testing reasons. ALL-IDB2 represents 50% of the total images as lymphoblasts. The proposed scheme is validated on the ALL-IDB1 dataset. These specimens were gathered by the specialists 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 having resolution 1712×1368.

A 5×5 cross-validation(CV) procedure is being applied to make the classifier more generalize. Table 2 shows the CV setting of the dataset we have used.

Table 2
k-fold cross validation settings for ALL-IDB dataset.

Dataset	Total number of cells	Training images		Testing images	
		Malignant	Benign	Malignant	Benign
ALL-IDB1	799	421	257	75	46

Table 3
Defination of performance measure.

Measure	Definition
TPR	$TP/(TP+FN)$
TNR	$TN/(TN+FP)$
$F-Score$	$2TP/(2TP+FN+FP)$
Precision	$TP/(TP+FP)$
Accuracy	$(TP+TN)(TP+TN+FP+FN)$

4.1. Performance evaluation of proposed DOST+PCA+LDA+ADBRF scheme

The experiments were carried out on a personal PC with an internal memory of 4 GB, running under Windows 8 operating system using Matlab toolbox. The performance of the proposed method is analyzed with other existing methods regarding true positive rate (TPR), true negative rate (TNR), and accuracy. The performance measures are illustrated in Table 3. TPR estimates the overall range of abnormal lymphocytes properly classified out of normal lymphocytes. TNR calculates the correctly classified normal cells. Accuracy measures the total number of correctly classified cells.

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 negative classes,

FP (False Positive): incorrect classification rate for negative classes

4.2. DOST based Feature extraction

The proposed work employs 2-dimensional discrete orthonormal *S*-transform (DOST)to extract texture features from the WBC sub-image. When DOST is applied to an image of size 256×256, it gives a feature matrix of 256×256 coefficients and each coefficient is considered as a feature. The output of the feature extraction step gives a feature vector of size 799×65536 (Algorithm 1). These features act as primary features. All the features extracted are not important for the classification purpose. So, a reduction technique is used to reduce the dimension of the feature set. Therefore, PCA+LDA (Algorithm 2) is employed to decrease the feature vector size to 35 which are the first 35 principal components.

4.3. Classification

The decreased 35 features with an output class level were given to the classifier. The optimum parameters of random forest classification are found out by varying the number of trees, and the features are selected for every node. It has been discovered the RF with 35 trees offers appropriate outcome. The performance of the proposed scheme was studied with the varied range of features (PCs), and the accuracies are shown in Figure 5. The highest accuracy was achieved with the first 35 principal components. The performance metrices regarding sensitivity, specificity, precision, $F-score$, and accuracy of ADBRF classifier is tabulated in Table 4. From Table 4, we can view that the proposed scheme is much more suitable than the different classifiers on the available dataset. Additionally, the training error rate on the number of trees is shown in Figure 6. It can be observed that 35 trees show a mean squared error of 0.006 which is the least one. The success of ADBRF classifier

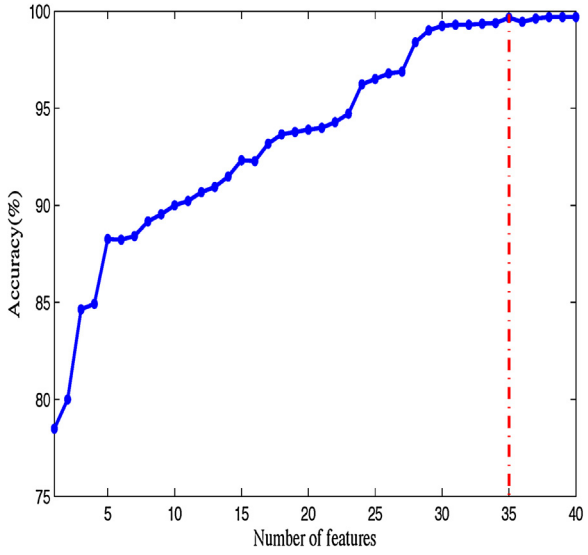

Fig. 5. Performance evaluation with respect to number of features.

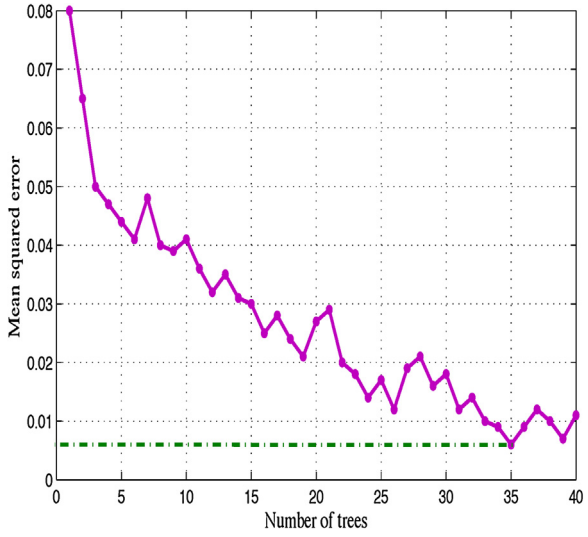

Fig. 6. Mean squared values for different trees in ADBRF method.

Table 4
Performance measure of different classifiers over 5-fold.

Classifier	Sensitivity (%)	Specificity (%)	$F-Score$(%)	Precision (%)	Accuracy (%)
k-NN ($k=5$)	95.19	94.77	95.96	96.75	95.038
BPNN	96.26	97.82	97.42	98.64	96.85
SVM-L	98.66	98.69	98.92	99.20	98.67
RF	100.00	97.81	99.33	98.69	98.67
ADBRF	100	99.12	99.73	99.46	99.66

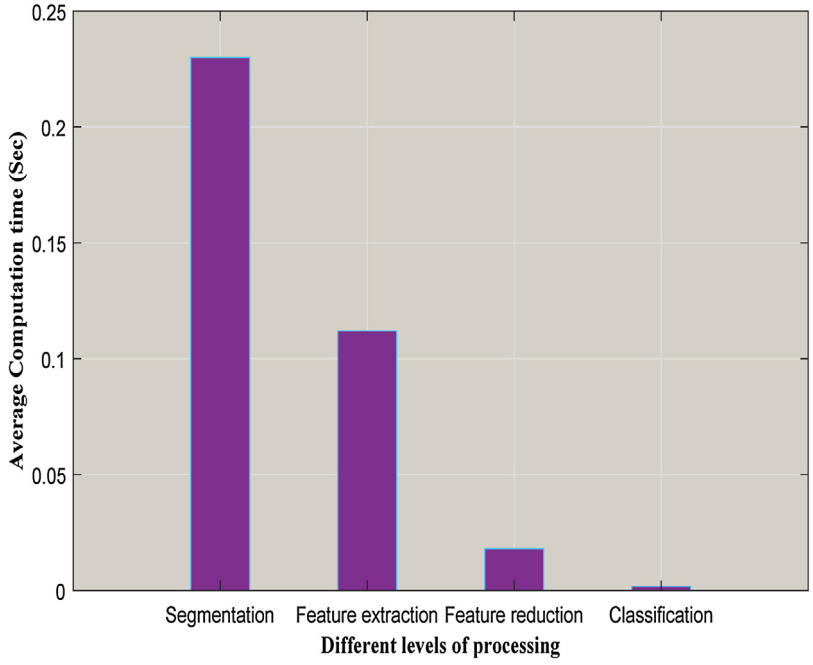

Fig. 7. Average computation time at different stages of the proposed scheme.

Table 5
ADFRF Parameter list

Parameter	Values	Description
T	35	Total number of trees
r	5	size of random vector
P	25	no. of iteration
w_1	$12 \times 10^{(}-4)$	Initial weight
Z_p	$2\sqrt{err_p(1-err_p)}$	Normalization constant

Table 6
Fold-wise result for each run of 5× 5 CV procedure.

Run	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Accuracy (%)
Run-1	121	121	120	120	121	99.66
Run-2	121	120	119	120	121	99.33
Run-3	121	121	121	120	120	99.66
Run-4	121	120	121	119	120	99.33
Run-5	121	121	121	121	121	100.00
						99.59

is determined by choosing the parameters properly. The optimum parameter with a lower error rate taken for the experiment is tabulated in Table 5.

Ultimately, a parallel investigation has been performed on the number of features and accuracy achieved by several methods and given in Table 8. It is inferred that the method proposed in [16]give 93.2% of accuracy with a large feature set. The method proposed by [15] can significantly reduce the number of features with greater accuracy. However, due to the ensembling technique used for classification, the time complexity is very high. The authors in [30] have taken shape and color features to with the help of multi-layer perceptron network to get a classification accuracy of 95.70%. Our suggested system uses 35 features which give the highest accuracy of 99.66%. The accuracy achieved by the proposed DOST+PCA+LDA+ADBRF scheme over ALL-IDB1 dataset is listed in Table 6. Table 7 shows the 5-fold CV procedures for the test instances using ADBRF classifier. Computation time is found to be one of the most important aspects of a CAD system. Figure 7 demonstrates the time required in the different stages of the proposed system. For each leukocyte sub-image, the computation

Table 7
Performance analysis of 5-fold CV procedure using ADBRF classifier.

Fold	Testing instances	TP	FN	TN	FP	Accuracy (%)
Fold-1	121	75	0	46	0	100.00
Fold-2	121	75	0	46	0	100.00
Fold-3	121	75	0	45	1	99.17
Fold-4	121	75	0	45	1	99.17
Fold-5	121	75	0	46	0	100.00
						99.66

Table 8
Classification performance comparison of suggested system with some other existing schemes.

Authors	Features	Classifier	Number of features	Accuracy (%)
Putzu et al. [16]	Morphological + colour + textural	*SVM – P*	131	93.2
Mohapatra et al.[15]	Morphological + colour + textural	EOC_5	33	99
Aimi et al. [30]	Shape + colour	*MLP*	42	95.70
Mishra et al. [31]	DCT	*SVM – L*	90	89.76
Mishra et al. [19]	GLCM+PCA (nucleus)	*RF*	40	98.04
	GLCM+PPCA (nucleus)	*RF*	40	99.004
Mishra et al. [32]	GLRLM	*SVM*	44	96.97
Proposed method	DOST+PCA	*RF*	35	96.7
	DOST+PCA+LDA	*RF*	35	98.67
	DOST+PCA+LDA	*ADBRF*	35	99.66
	DOST+PCA+LDA (cytoplasm)	*ADBRF*	35	94.61

times required for segmentation, feature extraction, feature reduction, and classification stage is found to be 0.23s, 0.112s, 0.018s, and 0.0017s respectively. The total time taken for the classification of a leukocyte sub-image is found to be 0.3617s.

Algorithm 1. DOST based feature extraction.

Require: n: total number samples
Ensure: Feature_matrix: $X[1:n, 1:m]$
Function *dost*() computes the DOST coefficients of sample images and *resize*() is used to set the dimension of each leukocyte.
Create $EM[1:Z, 1:Z]$ which is used to hold DOST co-efficients and a feature vector (FV)
Initialise Z (pixel value), $s \longleftarrow 1$
$m \longleftarrow Z \times Z$(Features extracted from a leukocyte sub-image (IP))
for $k \longleftarrow 1$ to n **do**
  Get IP_k
  $IP_k \longleftarrow resize(IP_k, Z)$
  $EM_k[1:Z, 1:Z] \longleftarrow dost(IP_k)$
  for $q \longleftarrow 1$ to Z **do**
    for $l \longleftarrow 1$ to Z **do**
      $FV_k[1, s] \longleftarrow EM_k[q, l]$
      $s \longleftarrow s+1$
    end for
    Reset $s \longleftarrow 1$
    for $p \longleftarrow 1$ to m **do**
      $X[k, p] \longleftarrow FV_k[1, p]$
    end for
  end for
end for

Algorithm 2. PCA+LDA based feature reduction.

Require: $X[1:n, 1:m]$: (n: Number of samples, m: Number of features)
Ensure: $RFM[1:n, 1:z]$: Reduced feature matrix
Function *lda*() and *pca*() calculates the principal co-efficients in the reduced space
$F[1:n, 1:n-1] \longleftarrow pca(X)$
Return F
Create an empty matrix $RFM[1:n, 1:z]$
Choose z
$RFM[1:n, 1:z] \longleftarrow lda(F, z)$
Return RFM

Algorithm 3. ADBRF Classifier

Require: S: Training dataset with n samples, $S = (x_i, c_i)$ for $i = 1, 2, \ldots, n$ and $x_i \in RFM$ with class labels $c_i \in C$
RFM and C represents the input and output class respectively.
base learner: Random forests
P: total number of iterations
x: testing samples
T: Number of trees
θ: input vector used for tree construction
Ensure: D_{fin}: Final classifier decision
Weight initialization: $w_1(i) = \frac{1}{n}$ for all i
for $p \longleftarrow 1$ to P **do**
  for $i \longleftarrow 1$ to T **do**
    Generate a vector θ_t with weight $w_p(i)$
    $S_t \longleftarrow bootstrapSamples(S)$
    $S_t \longleftarrow buildTreeClassifier(S_t, \theta_t)$
    prediction result based on voting
  end for
  Weak hypothesis, $h_p(x_i) \rightarrow C_i \in \{0, 1\}$
  Calculate the error
  $$err_p \longleftarrow Pr_i w_p[h_p(x_i) \neq C_i] = \sum_{i:h_p(x_i) \neq C_i} w_p(i)$$
  Choose a parameter $\eta_p \longleftarrow \frac{1}{2}\ln\left(\frac{1-err_p}{err_p}\right) > 0$
  update weight,
  $$w_{p+1}(i) \longleftarrow \frac{w_p(i)}{Z_p} \times \begin{cases} e^{-\eta_p} & \text{if } h_p(x_i) = C_i \\ e^{\eta_p} & \text{if } h_p(x_i) \neq C_i \end{cases}$$
  Z_p is a normalization constant
end for
Return D_{fin} $sign\left(\sum_{p=1}^{P} \eta_p h_p(x)\right)$

5. Conclusion

In this paper, we have proposed a DOST+PCA+LDA based CAD system for the detection of ALL. This scheme uses 2-dimensional discrete orthonormal S-transform for extracting texture features from the microscopic images. PCA+LDA is harnessed to reduce the dimension of the extracted features. The significant features selected are the supplied to a combination of AdaBoost and random forest classifier for achieving better performance. The effectiveness of the proposed scheme is demonstrated through comparison over different classifier. The method achieves an accuracy of 99.66% on the ALL-IDB1 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, CA: A 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] S. Mishra, B. Majhi, P.K. Sa, A survey on automated diagnosis on the detection of leukemia: a hematological disorder, in: 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), IEEE, 2016, pp. 460–466.
[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. Pediatr. Hematol. Oncol. 33 (6) (2011) 475–479.
[5] 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.
[6] T. Singh, Atlas and Text of Hematology, Avichal Publishing Company, New Delhi, 2010, pp. 136.
[7] T.-C. Lin, R.-S. Liu, Y.-T. Chao, S.-Y. Chen, Classifying subtypes of acute lymphoblastic leukemia using silhouette statistics and genetic algorithms, Gene 518 (1) (2013) 159–163.
[8] N. Zong, M. Adjouadi, M. Ayala, Artificial neural networks approaches for multidimensional classification of acute lymphoblastic leukemia gene expression data, WSEAS Trans. Inf. Sci. Appl. 2 (8) (2005) 1071–1078.
[9] M.E. Ross, X. Zhou, G. Song, S.A. Shurtleff, K. Girtman, W.K. Williams, H.-C. Liu, R. Mahfouz, S.C. Raimondi, N. Lenny, et al., Classification of pediatric acute lymphoblastic leukemia by gene expression profiling, Blood 102 (8) (2003) 2951–2959.
[10] H.J. Escalante, M. Montes-y Gómez, J.A. González, P. G ómez-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.
[11] D.J. Foran, D. Comaniciu, P. Meer, L.A. Goodell, Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy, IEEE Trans. Inf. Technol. Biomed. 4 (4) (2000) 265–273.
[12] V. Piuri, F. Scotti, Morphological classification of blood leucocytes by microscope images, Computational Intelligence for Measurement Systems and Applications 2004. CIMSA. 2004 IEEE International Conference on, IEEE (2004) 103–108.
[13] F. Scotti, Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images, IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (2005) 96–101.
[14] H.A. Halim, M.Y. Mashor, R. Hassan, Automatic blasts counting for acute leukemia based on blood samples, Int. J. Res. Rev. Comput. Sci. 2 (4) (2018).
[15] 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.
[16] L. Putzu, G. Caocci, C. Di Ruberto, Leucocyte classification for leukaemia detection using image processing techniques, Artif. Intell. Med. 62 (3) (2014) 179–191.
[17] J. Angulo, J. Klossa, G. Flandrin, Ontology-based lymphocyte population description using mathematical morphology on colour blood images, Cell. Mol. Biol. 52 (6) (2006) 2–15.
[18] G. Zack, W. Rogers, S. Latt, Automatic measurement of sister chromatid exchange frequency, J. Histochem. Cytochem. 25 (7) (1977) 741–753.
[19] S. Mishra, B. Majhi, P.K. Sa, L. Sharma, Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection, Biomed. Signal Process. Control 33 (2017) 272–280.
[20] C. Gonzalez, R.E. Woods, Digital Image Processing, Nueva Jersey, 2018.
[21] L. Mansinha, R. Stockwell, R. Lowe, Pattern analysis with two-dimensional spectral localisation: applications of two-dimensional s transforms, Phys. A 239 (1) (1997) 286–295.

[22] S. Drabycz, R.G. Stockwell, J.R. Mitchell, Image texture characterization using the discrete orthonormal S-transform, J. Digit. Imag. 22 (6) (2009) 696–708.
[23] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data–with application to face recognition, Pattern Recogn. 34 (10) (2001) 2067–2070.
[24] M. Bishop, Pattern recognition, Mach. Learn. (2018) 128.
[25] Y. Freund, R.E. Schapire, et al., Experiments with a new boosting algorithm, Int. Conf. Mach. Learn. 96 (1996) 148–156.
[26] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.
[27] Y. Freund, R. Schapire, N. Abe, A short introduction to boosting, J. Jpn. Soc. Artif. Intell. 14 (771-780) (1999) 1612.
[28] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[29] ALL-IDB dataset for ALL classification, ALL-IDB dataset for ALL classification, http://crema.di.unimi.it/fscotti/all/.
[30] A.N. Aimi, M.Y. Mashor, H. Rosline, Classification of acute leukaemia cells using multilayer perceptron and simplified fuzzy ARTMAP neural networks, Int. Arab J. Inf. Technol. 10 (4) (2013) 356–364.
[31] S. Mishra, L. Sharma, B. Majhi, P.K. Sa, Microscopic image classification using DCT for the detection of acute lymphoblastic leukemia (ALL), in: Proceedings of International Conference on Computer Vision and Image Processing, Springer, 2017, pp. 171–180.
[32] S. Mishra, B. Majhi, P.K. Sa, F. Siddiqui, GLRLM based feature extraction for acute lymphoblastic leukemia (ALL detection, 2018.