Microscopic Image Classification Using DCT for the Detection of Acute Lymphoblastic Leukemia (ALL)

Sonali Mishra, Lokesh Sharma, Bansidhar Majhi and Pankaj Kumar Sa

Abstract Development of a computer-aided diagnosis (CAD) system for early detection of leukemia is very essential for the betterment of medical purpose. In recent years, a variety of CAD system has been proposed for the detection of leukemia. Acute leukemia is a malignant neoplastic disorder that influences a larger fraction of world population. In modern medical science, there are sufficient newly formulated methodologies for the early detection of leukemia. Such advanced technologies include medical image processing methods for the detection of the syndrome. This paper shows that use of a highly appropriate feature extraction technique is required for the classification of a disease. In the field of image processing and machine learning approach, Discrete Cosine Transform (DCT) is a well-known technique. Nucleus features are extracted from the RGB image. The proposed method provides an opportunity to fine-tune the accuracy for the detection of the disease. Experimental results using publicly available dataset like ALL-IDB shows the superiority of the proposed method with SVM classifier comparing it with some other standard classifiers.

Keywords Acute Lymphoblastic Leukemia · Discrete Cosine Transform · Watershed segmentation · CAD system

S. Mishra (✉) · L. Sharma · B. Majhi · P.K. Sa
Pattern Recognition Research Lab, Department of Computer
Science and Engineering, National Institute of Technology, Rourkela 769008, India
e-mail: smishra.nitrkl@gmail.com

L. Sharma
e-mail: lksharma1064@gmail.com

B. Majhi
e-mail: bmajhi@nitrkl.ac.in

P.K. Sa
e-mail: pankajksa@nitrkl.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 459,
DOI 10.1007/978-981-10-2104-6_16

1 Introduction

Acute Leukemia is a rapidly increasing disease that affects mostly the cells that are not yet fully developed. Acute Lymphoblastic Leukemia (ALL) is a significant ailment caused by the unusual growth and expansion of white blood cells [1]. ALL begins with the abnormalities starting from the bone marrow, resulting in reducing the space for red blood cells. The ALL blasts become so numerous that they flood through the red blood cells and platelets. As the cells build up, they reduce the immunity to fight with the foreign material. Hence, it is essential to treat the disease within a short span of time after making a diagnosis. As per the survey done by American Cancer Society, it has approximated that, in 2015 a total of 1,658,370 has been diagnosed, out of which 589,430 died in the US. In India, the total number of individuals suffering from leukemia was estimated to be 1,45,067 in 2014. Furthermore, as per the Indian Association of blood cancer and allied diseases, among all the cancers which is dangerous and can cause death, leukemia constitute one-third of the cases. ALL is mostly seen in children below 14 years [2].

ALL is identified with the excessive production of immature lymphocytes that are commonly known as lymphoblasts. The uncontrolled manufacture of lymphoblasts puts a stop to the formation of blood in the marrow, which eventually leads to the cause of death. The recognition of the blast cells in the bone marrow of the patients suffering from acute leukemia is a crucial step in the recognition of the development stage of the illness and proper treatment of the patients. The percentage of blasts are an important factor to define the various stages of lymphoblastic leukemia. According to the French–American–British (FAB) standard, three different types of acute lymphoblastic leukemia are classified based on the morphology of blast cells [3].

The morphological identification of acute leukemia is mainly performed by the hematologists [4]. The process begins by taking the bone marrow sample from the patient's spine. Wright's staining method is applied to make the granules visible during analysis [5]. This process involves many drawbacks, such as slowness of the analysis, a very low accuracy, requirement of an extremely skilled operator, etc. The identification by experts is reliable, but automated tools would be useful to support experts and also helpful in reducing the cost. The primary goal of this work is to analyze microscopic images by designing a computer-aided diagnosis system (CAD) to support medical activity.

This paper presents a new hybrid technique for acute leukemia classification from the microscopic images based on machine learning approaches. The proposed method mainly consists of three different steps namely, Segmentation, feature extraction, and classification of the disease. Over the years, many automatic segmentation techniques have been proposed for the disease, still they fail to segment the overlapping blood cells. This scheme utilizes discrete cosine features and support vector machine (SVM) for classification of normal and malignant cells.

The rest of the paper is organized as follows: Sect. 2 deals with some of the highly regarded works on the detection of leukemia from the blood smear along with some segmentation schemes. Section 3 presents the proposed work. Section 4 gives a comparative performance study of the proposed method with existing schemes. Finally the concluding remarks are provided in Sect. 5.

2 Related Work

A careful study on automatic blood cell recognition reveals that numerous works have been reported since early 2000. All these existing techniques said, giving a near perfect performance under certain constraints. Various segmentation and feature extraction techniques have been examined for the same. The following paper gives an overview of the different segmentation and feature extraction techniques based on their category.

Scotti [6] has proposed a method for automated classification of ALL in gray level peripheral blood smear images. As per the experiments conducted by them on 150 images, it has been concluded that lymphoblast recognition is feasible from blood images using morphological features. Gupta et al. [7] have proposed a suitable support vector machine-based technique for the identification of three types of lymphoblasts. The classification accuracy for the childhood ALL has been promising but needs more study before they are used for the adult. Escalante et al. [8] have suggested an alternative approach to leukemia classification using ensemble particle swarm model selection. Manually isolated leukemia cells are segmented using Markov random fields. This method is useful for ALL versus AML (Acute Myeloblastic Leukemia) classification. Putzu et al. [9] have proposed a scheme for automated detection of leukemia using image processing techniques. The segmentation procedure produces a lower segmentation rate and can be further improved. Mohapatra et al. [10] have suggested an ensemble classifier system to classify a blood cell into normal lymphocyte and an unhealthy one (lymphoblast). The scheme yields good accuracy for detecting the disease, but failed to detect the disease for grouped cell present in an image.

3 Proposed Work

The proposed method comprises of different stages such as the acquisition of images, preprocessing, segmentation of overlapping cells, feature extraction, and classification of the image into a normal (Benign) and abnormal (malignant) one. Figure 1 shows an overall block diagram of the proposed method. Each stage is discussed below in brief. The main contribution in this article is the use of discrete cosine transform (DCT) coefficients in SVM classifier for classification.

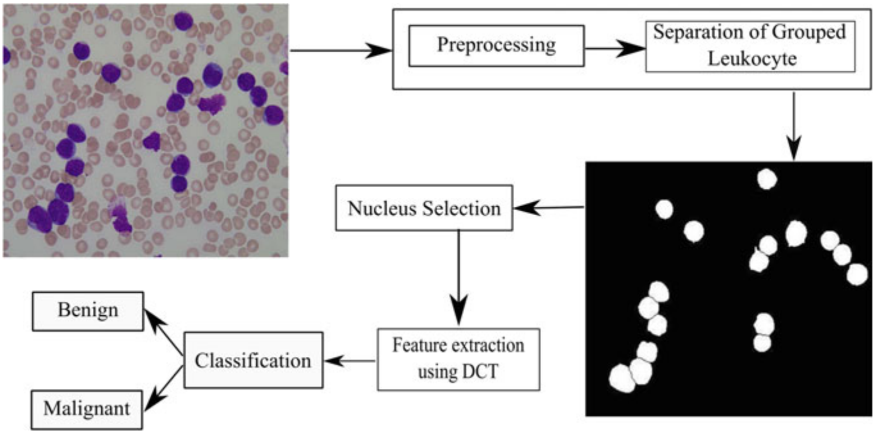

Fig. 1 Block diagram of the method for the classification of ALL

3.1 Preprocessing

Due to the presence of noise in the microscopic images under the diverse lighting conditions, the image requires preprocessing prior to segmentation. To generate a better quality image, Weiner filtering followed by contrast enhancement with histogram equalization is used.

3.2 Separation of Grouped Leukocyte Using Segmentation

Segmentation is a critical step for correct classification of the objects. Microscopic images are typically in RGB color space. It is very difficult to achieve accurate segmentation in the color image. So the RGB image is converted into Lab color space to reduce the dimension with the same color information. Color-based clustering mainly uses the Lab color space for the segmentation purpose.

Due to the presence of overlapped and grouped objects, marker-based watershed segmentation algorithm [11] is used for separating grouped and overlapped objects. In Lab space, component ‘a’ contains the highest information about the nucleus. So further processing is done on the component ‘a’. After separating the objects from an image, all the lymphocytes are extracted using the bounding box technique for the detection of ALL. Finally, the single lymphocyte sub-image is used in the next process for feature extraction.

3.3 Feature Extraction

The fundamental step of the proposed scheme is to calculate the DCT features from the lymphocyte sub-images of size $M \times N$. The cosine transform generates $M \times N$ DCT coefficients. Since the DCT has the energy compaction property and the energy coefficient are in descending order, the higher order coefficients are significant and the lower coefficient can be neglected. Hence, very few coefficients retain the energy of the whole image and reconstruct the original image with minor loss of information. This particular behavior of DCT has been exploited to use few DCT coefficients as a feature for classification of normal and abnormal cells in the image. Also, the feature extraction capacity of the DCT coupled with fast computation time has made it a worldwide candidate for pattern recognition. The general equation for a 2D ($M \times N$ image) DCT is defined by the following equation [12]:

$$F(u,v) = \frac{1}{\sqrt{MN}}\alpha(u)\alpha(v)\sum_{x=1}^{M}\sum_{y=1}^{N} f(x,y) cos\left[\frac{(2x+1)u\pi}{2M}\right] cos\left[\frac{(2y+1)v\pi}{2N}\right] \tag{1}$$

Gray scale coordinate of the image of size $M \times N$ is represented by, $f(x,y)$, where $1 \leq \text{x} \leq M 1 \leq \text{y} \leq N$. $\alpha(w)$ can be defined as,

$$\alpha(w) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{for } w = 1 \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

The detailed step for feature extraction is articulated in Algorithm 1.

Algorithm 1 Feature Extraction Algorithm

```
Require: Samples of $n$ lymphoblasts lymphocytes from the segmentation step
Ensure: $X[n : m]$: Feature matrix, $S[1 : n, 1 : m+1]$: New dataset
1: Compute DCT features using function $dct2()$ from the microscopic images
2: Store the features in the $X[n \times m]$
3: Append another vector $Y$ to the $X$ and assign a class level for each sample
4: Form a new dataset, $S$=($x_i$, $y_i$),$x_i \in X$, $y_i \in Y$, $1 \leq i \leq n$
```

3.4 Classification

Classification is the task of assigning an unknown feature vector, to one of the known classes. Each classifier has to be built up in such a way that a set of inputs must produce a desired set of outputs. In this paper support vector machine (SVM), a classifier is used for differentiating normal and malignant cells. Support vector machine is a very powerful supervised learning technique that was first introduced by Vapnik [13]. It is a two class supervised classifier. It uses a nonlinear mapping function for transforming the input data into a high-dimensional feature space by creating a hyperplane between the two categories. The entire set of measured data is divided into training and testing data. Here, images from ALL-IDB1 is used for both training and testing purpose. The extracted features from the above step can be classified using three other standard classifiers, i.e., Naive Bayesian [14], KNN [15], BPNN [16], and SVM. The basic steps of classification process are given in Algorithm 2.

Algorithm 2 Classifier System

```
Require: $S$: Training dataset with $N$ samples, $S = (x_i, y_i)$ for $i = 1, 2, \dots, N$ and $x_i \in X$ with class
    labels $y_i \in Y$
    $X$ and $Y$ represents the input and output class respectively.
Perform K-fold cross validation
for each classifier, (j= 1 to $M$) do
    Train the classifier
    Calculate the performance measures on test images
end for
```

4 Experimental Setup and Results

The experiments are carried out on a PC with 3.40 GHz Core-i7 processor and 4 GB of RAM, running under Windows 8 operating system. The proposed algorithm is simulated using MATLAB 2013 toolbox. Specimen blood samples are collected from a public database ALL-IDB [17, 18] having two distinct versions, namely, ALL-IDB1 and ALL-IDB2 containing 108 and 260 images, respectively. The ALL-IDB1 dataset is used for the purpose of training and testing. Among them, 59 are normal images and 49 are affected blood cells that consist of at least one lymphoblast. The ALL-IDB2 dataset is made from ALL-IDB1 by cropping the images having less dimension that is mainly used for testing purpose. The ALL-IDB is a public database employed in the field of medical image processing for the research purpose and detection of the tumor in the blood cells. Each image present in the dataset is represented using three primary colors (red, green, blue), and the image is stored with a size of 1368×1712 array.

To make the classifier more stable and more generalize, a fivefold cross-validation (CV) procedure is utilized. In this work, the abnormal (malignant) and normal (benign) images have been considered to be in the positive and negative class, respectively. Sensitivity is the probability that an investigative step is positive while the patient has the disease, whereas specificity is the probability that a diagnostic report is negative while the patient has not got any disease. For a given sample, a training system leads to four possible categories that are described in Table 1.

Table 1 Confusion matrix

	Positive	Negative	Performance measure
Positive	**True Positive (TP)**	**False Positive (FP)**	Positive predictive value = TP/(TP + FP)
Negative	**False Negative (FN)**	**True Negative (TN)**	Negative predictive value = TN/(TN + FN)
Performance measures	True Positive Rate (Sensitivity) = TP/(TP + FN)	True Negative Rate (Specificity) = TN/(TN + FP)	Accuracy = (TP + TN)/Total number of samples

4.1 Results and Discussion

The first step after the acquisition of the image is to clean the image to differentiate the lymphoblasts from the other component of the cell like RBC, platelets, etc. Weiner filter is being used to reduce noise present in the image along with the contrast enhancement. The next step is the segmentation process. Grouped cells of an image are separated using watershed segmentation. Conventional watershed segmentation is used along with the use of a marker. Marker-controlled watershed transform is a two-way process. It consists of two types. Internal markers represent the blast cell nucleus, and external markers represent the boundary to separate the grouped cells. Figure 2 represents the overall steps associated with the segmentation process.

The next step is to find the DCT coefficients. Here, the DCT coefficients have been taken as the feature for the classification process. Due to the energy compaction properties of the DCT, less computational time is required. The number of extracted

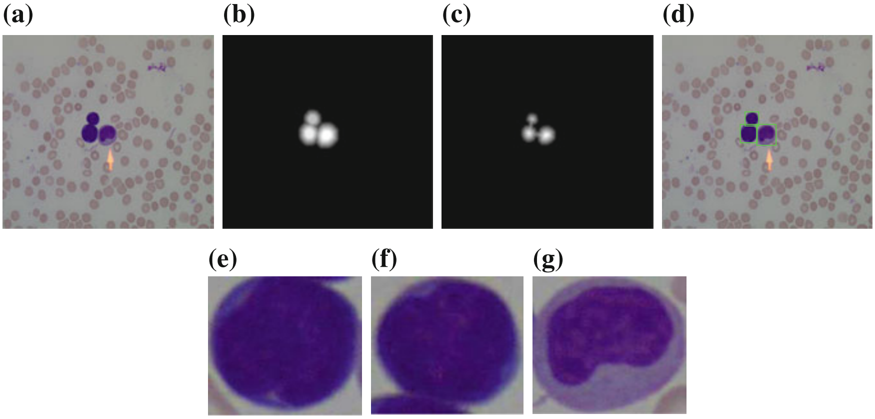

Fig. 2 Different segmentation steps: **a** original image, **b** image after using external marker, **c** image after using internal marker, **d** detected lymphoblasts, **e**–**g** detected sub-image using bounding box

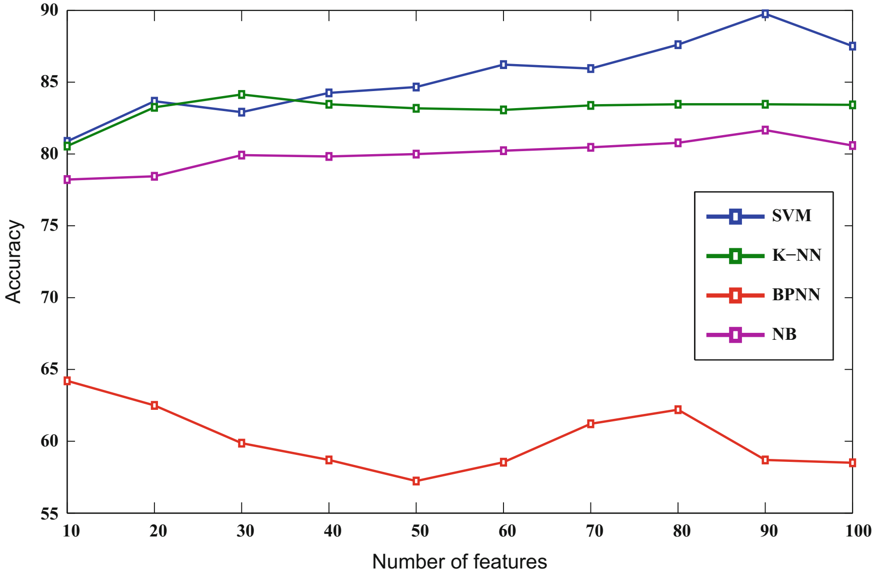

Fig. 3 Accuracy rate with the increase in number of features

Table 2 Comparison of accuracy of various classifiers over fivefold

Classifier	Fold					Average accuracy (%)
	1	2	3	4	5	
NB	78.49	83.78	81.23	83.45	81.35	81.66
KNN	82.85	80.59	85.51	84.23	84.12	83.46
BPNN	58.54	54.26	60.85	56.95	63.20	58.7
SVM	85.21	94.32	85.59	88.62	95.06	89.76

features is found to be 90. In Fig. 3, the classification accuracies of different classifiers with some features are portrayed. It is observed from that, all the classifiers show maximum accuracy at feature number 90. The acquired features are fed to different classifiers to get different performance measures. Tables 2, 3, and 4 present the fold-wise result of the fivefold cross-validation procedure for the determination of accuracy, sensitivity, and specificity for different classifiers. The optimum has been achieved with an accuracy of 89.76 % using the SVM classifier. The obtained values of sensitivity and specificity are found to be 84.67 % and 94.61 % respectively. Note that all the schemes are tested on the same dataset ALL-IDB1.

Table 3 Comparison of average sensitivity of various classifiers over fivefold

Classifier	Fold					Average Sensitivity (%)
	1	2	3	4	5	
NB	88.62	87.32	83.65	91.25	92.66	88.70
KNN	99.80	96.32	97.35	99.21	99.21	98.38
BPNN	77.38	83.25	85.23	79.36	77.98	80.64
SVM	89.36	85.69	88.22	79.24	80.84	84.67

Table 4 Comparison of average specificity of various classifiers over fivefold

Classifier	Fold					Average Specificity (%)
	1	2	3	4	5	
NB	70.32	78.61	76.55	80.29	71.13	75.38
KNN	62.89	68.25	69.41	73.56	72.04	69.23
BPNN	35.23	39.58	36.24	38.22	39.18	37.67
SVM	96.23	95.37	94.35	88.39	98.71	94.61

5 Conclusion

In this work, a hybrid system for the automatic classification of leukemia using microscopic images has been proposed. This system first applies Wiener filter followed by histogram equalization to preprocess the image. The watershed segmentation algorithm has been utilized to correctly separate the lymphocytes sub-image from the preprocessed image. The DCT-based feature is used for deriving a set of features from the sub-images. Subsequently, SVM is used to classify the images as benign and malignant. The simulation results show the efficacy of the proposed scheme while testing the system with SVM classifier. The classification accuracy on dataset ALL-IDB1 is found to be 89.76 % using SVM. However, there is a scope to reduce the computational overhead of the feature extraction step, and also works can be further extended toward the extraction of cytoplasm from the blood cells.

References

1. Siegel, R., Naishadham, D., Jemal, A.: Cancer statistics, 2013. CA: a cancer journal for clinicians 63.1, 11–30 (2013)
2. Kulkarni, K.P., Arora, R.S., Marwaha, R.K.: Survival outcome of childhood acute lymphoblastic leukemia in India: a resource-limited perspective of more than 40 years. Journal of pediatric hematology/oncology 33.6, 475–479 (2011)
3. Singh, T.: Atlas and text of hematology. Avichal Pub-lishing Company, New Delhi 136 (2010)
4. Saraswat, M., Arya, K.V.: Automated microscopic image analysis for leukocytes identification: A survey. Micron 65 20–33 (2014)

5. Wright, J. H.: The histogenesis of the blood platelets. Journal of Morphology. Vol. 3. No. 1. (1910)
6. Scotti, F.: Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. IEEE International Conference on Computational Intelligence for Measurement Systems and Applications. (2005)
7. Gupta, L., Jayavanth, S., Ramaiah, A.: Identification of different types of lymphoblasts in acute lymphoblastic leukemia using relevance vector machines. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 6675–6678 (2008)
8. Escalante, H.J., et al.: Acute leukemia classification by ensemble particle swarm model selection. Artificial intelligence in medicine 55.3 163–175 (2012)
9. Putzu, L., Caocci, G., Ruberto, C.D.: Leucocyte classification for leukaemia detection using image processing techniques." Artificial intelligence in medicine 62.3 179–191 (2014)
10. Mohapatra, S., Patra, D., Satpathy, S.: An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. Neural Computing and Applications 24.7-8 1887–1904 (2014)
11. Parvati, K., Rao, P., Das, M.M.: Image segmentation using gray-scale morphology and marker-controlled watershed transformation. Discrete Dynamics in Nature and Society (2008)
12. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE Transactions on Computers,100.1 90–93 (1974)
13. Vapnik, V.N., Vapnik, V.: Statistical learning theory. Vol. 1. New York: Wiley, (1998)
14. Duda, R.O., Hart, P.E., Stork, D.: Pattern classification. John Wiley & Sons, (2012)
15. Acharya, T., Ray, A.K.: Image processing: principles and applications. John Wiley & Sons, (2005)
16. Rumelhart, D.E., Hinton, G. E., Williams, R.J.: Learning representations by back-propagating errors. Cognitive modeling 5.3 (1988)
17. Labati, R.D., Piuri, V., Scotti, F.: All-IDB: The acute lymphoblastic leukemia image database for image processing. 18th IEEE international conference on Image processing (ICIP), (2011)
18. ALL-IDB Dataset for ALL Classification. http://crema.di.unimi.it/~fscotti/all/