



Detection of acute lymphoblastic leukemia using image segmentation and data mining algorithms

Vasundhara Acharya, B.E.,M.Tech¹ · Preetham Kumar² 

Received: 19 October 2017 / Accepted: 22 April 2019 / Published online: 14 June 2019
© International Federation for Medical and Biological Engineering 2019

Abstract

Blood is composed of white blood cells, red blood cells, and platelets. Segmentation of the blood smear cells and extraction of features of the cells is essential in the field of medicine. Acute lymphoblastic leukemia is a form of blood cancer caused due to the abnormal increase in the production of immature white blood cells in the bone marrow. It mostly affects the children below 5 years and adults above 50 years of age. Due to the late diagnosis and cost of the devices used for the determination, the mortality rate has increased drastically. Flow cytometry technique that performs automated counting fails to identify the abnormal cells. Manual recount performed using hemocytometer are prone to errors and are imprecise. The proposed work aims to survey different computer-aided system techniques used to segment the blood smear image. The primary objective here is to derive knowledge from the different methodologies used for extracting features from white blood cells and develop a system that would accurately segment the blood smear image by overcoming the drawbacks of the previous works. The objective mentioned above is achieved in two ways. Firstly, a novel algorithm is developed to segment the nucleus and cytoplasm of white blood cell. Secondly, a model is built to extract the features and train the model. The different supervised classifiers are compared, and the one with the highest accuracy is used for the classification. Six hundred images are used in the experimentation. InfoGainAttributeEval and the Ranker Search method are used to achieve the feature selection which in turn helps in improvising the classifier performance. The result shows the classification of the acute lymphoblastic leukemia into its three respective categories namely: ALL-L1, ALL-L2, ALL-L3. The model can differentiate between a normal peripheral blood smear and an abnormal blood smear. The extracted feature values of a cancerous cell and a normal cell are also shown. The performance of the model is evaluated using the test images stained with various stains. The proposed algorithm achieved an overall accuracy of 98.6%. The promising results show that it can be used as a diagnostic tool by the pathologists.

Keywords Acute lymphoblastic leukemia · Blood smear cells · Flow cytometry · Hemocytometer · White blood cell

1 Introduction

1.1 Motivation

Acute leukemia is a cancer of the blood or bone marrow. It is a heterogeneous group of neoplastic diseases which is categorized mainly into two main types which are myeloid and lymphoid. The research work focuses on the identification of lymphoid leukemia. In this case, it leads to the production of a large number of abnormal white blood cells termed as lymphoblasts. According to a survey, it is said that the probability of prediction of ALL in adults is reduced and the rate of mortality is high. The percentage of successful treatment is estimated to be 25–

✉ Vasundhara Acharya, B.E.,M.Tech
vasundhara.acharya@manipal.edu

Preetham Kumar
preetham.kumar@manipal.edu

¹ Department of Computer Science and Engineering, Manipal Institute of Technology (MIT), Manipal Academy of Higher Education(MAHE), Manipal, India

² Department of I&CT, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

40%. The different categories of acute lymphoblastic leukemia are ALL-L1, ALL-L2, and ALL-L3 [1].

The cells in ALL-L1 are regular shaped and small. The cells in ALL-L2 are large, and the nuclear membrane is irregular. The cells in ALL-L3 are giant and have a hole-like structure termed as vacuoles. In the market, the analysis of the blood smear cells is done using both automated and manual techniques. The automated procedures involve the use of laser sensors. The process is carried out with a small amount of sample. The sensor counts the number of cells passing through the aperture. The light detector and electrical impedance are the two primary sensors. The results obtained are exact, but it fails to identify the abnormal cells. Manual procedures are employed to overcome the above problem. In this case, the cell counting is achieved using counting chambers that hold the specific volume of the blood specimen. A blood film is prepared, and the different number of white blood cells is enumerated. The technique successfully identifies abnormal cells. The drawback here is the human's capability of visualizing the cell features. It is time-consuming and is less precise. The main downside of the manual approach is it is prone to sampling error. With the need for quality results, there arose a necessity for the automation of the whole process. It would reduce the burden on hematologists and would help in obtaining accurate results in a short period. In this research work, a new approach to automated detection is proposed.

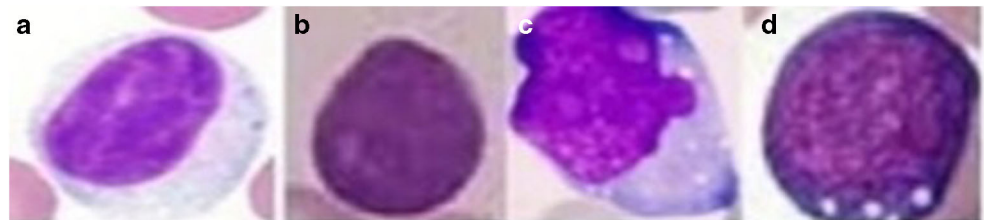
The advantage of using a computer system for analysis is it requires only one image and not the entire blood sample. The features like the count, texture, shape, and maturity level of white blood cells could aid in the diagnosis of diseases that range from inflammatory to leukemia. The precision of the result obtained using a computer system is much higher compared to the result obtained by human analysis. Blood smear image consists of white blood cells (WBCs), red blood cells (RBCs), and platelets distributed over the entire image. The primary role of computer-aided diagnosis tool is to achieve segmentation of the blood smear, feature extraction, and classification. The segmentation accuracy has a significant decisive influence on cancer identification. The region of interest characteristics primarily affects the cancer classification. In this paper, it aims to propose a model to segment the blood smear and diagnose acute lymphoblastic leukemia. Normal lymphocyte and different subtypes of acute lymphoblastic leukemia are shown in Fig. 1.

1.2 Related works

The features extracted from WBC play a vital role in the diagnosis of cancer. Thus, the primary step for the development of a CAD tool is the separation of WBC from the other components and its classification. Numerous studies regarding this are discussed here.

Afshar et al. [2] used the clinical parameters to diagnose leukemia. A sample test was conducted using SPSS software for 38 parameters. Based on the significance value, eight parameters were selected. The network was trained using the Levenberg-Marquardt learning algorithm with the parameters chosen. The trained network was later tested with test data. Alomari et al. [3] proposed an approach for segmentation of RBCs and WBCs using threshold technique. The image was converted to grayscale, preprocessed, and a threshold value of 64 was used to convert it to binary. Image complement, noise elimination, and fill operations were performed to improve the result of edge detection. Belekar and Chougule [4] used the mathematical morphology and SMMT (self-dual multiscale morphological toggle) operator for segmentation of nucleus and cytoplasm. For the segmentation of the nucleus, they utilized two approaches that are level set method and watershed algorithm. After the application of two procedures, they compared the results and identified that the application of the SMMT operator would improve the outcomes. They made use of the SMMT operator along with threshold operation to prevent the problem of leakage. For the detection of cytoplasm, they used granulometric analysis and mathematical morphology technique. A comparison between the two approaches was also made. Bhagvathi and Thomas [5] used the fuzzy logic for the segmentation of WBC. Initially, the RGB image obtained was converted using the grayscale image. Fuzzy rules were set to detect the circles. To identify the concentric circle, multiple radiant arguments were used. The result was an accumulator array. Image complement, noise elimination, and hole filling operation were performed. Based on the intensity values, the WBCs were separated from RBCs. Erosion operation was applied, and the WBC mask was obtained. Bhamare and Patil et al. [6] proposed a technique to obtain the WBCs from the blood smear images using the manually computed threshold. The input image was preprocessed by applying the contrast stretch operation. The resulting image was converted to grayscale and then to binary.

Fig. 1 a Normal lymphocyte. b ALL-L1 cell. c ALL-L2 cell. d ALL-L3 cell



Morphological dilation, noise elimination, and erosion operations [7] were applied to the resulting image. The output image had only white blood cells. Shape, color, and texture features were computed. Joshi et al. [8] proposed an algorithm for segmentation of white blood cells using Otsu's threshold. Operations like histogram equalization and noise elimination were performed. Otsu's threshold was applied to convert the image to binary. The resulting image was used for feature extraction. Area, perimeter, and circularity were computed. The classification of cells was done by using the K nearest neighbor algorithm. Marzuki et al. [9] proposed a technique to obtain the WBC nucleus using active contour technique. The input image was converted to grayscale. Active contour was used to determine the boundaries of the nucleus. The initial point was chosen randomly. It was used to determine the edges. Next, the roundness of the object was computed. Based on the value, the shapes of blood cell were classified. Mishra and Deshmukh [10] proposed a technique to identify the blast cells from the blood smear image. Background separation, noise elimination, and contrast adjustment were performed. The region of interest was the nucleus. The cells were extracted, and they were categorized as blast cell or normal cell. Mulik et al. [11] proposed an approach for identification of WBCs by using K-means clustering. The number of clusters chosen was 4. In one group, it would hold the WBC, another would hold the RBC, next would hold the cytoplasm, and last would hold the background. Based on WBC extracted, some features such as area, perimeter, eccentricity, and some texture features like energy and homogeneity were calculated. Support vector machine was used for classification. Nasir et al. [12] proposed a technique to extract the WBCs from the blood smear image by segmenting the image into three regions by constructing a histogram. The input image was converted to grayscale. The gray level histogram was built to obtain the threshold that would segment the image into three respective regions having the WBC, RBC, and background. Pixel values lying in the range 0–89 were WBCs. RBCs were present in the pixel range 90–170. The background pixels were present in the pixel range 171–255. The resulting image was converted to binary and unwanted particles were eliminated. The objects in the image were labeled using the 'bwlabel' function. A graph was plotted which had the WBC to RBC ratio. The chart was plotted for acute lymphoblastic leukemia cases, acute myeloid leukemia cases, and normal cases. Putzu and Ruberto [13] avoided the different steps of pre-processing and segmentation. They applied to pre-process at each stage of segmentation that provided a robust and straightforward solution for segmentation of WBC from other cells. They determined the threshold value using the Zack algorithm. The leukocytes that were grouped were removed by considering the roundness values of the connected components. If the roundness value was higher than a certain predetermined threshold, they were individual leukocytes

and did not require any separation. But if it was less than prescribed limit, then it needed the separation. Separation was performed by getting the information of the points of concavity and the information related to the positions of the maximum image in gray tones. For the detection of cytoplasm, the image was cropped to obtain individual leukocytes. Sadeghian et al. [14] proposed a framework for the detection of WBC. The color scale image obtained was converted to grayscale, and it was cropped to have one cell per image. Edge detection was performed to detect the edges of the nucleus. Snake algorithm found the connected boundary that was identified and it selected the nucleus. The cytoplasm was segmented using the Zack algorithm. Thanh et al. [15] proposed an approach to classifying acute leukemia using the convolution neural network. The network had four layers. The first three layers were used to detect the features, and the other two layers were used for classification. The model was able to classify leukemia into four types. Vaghela et al. [16] proposed a technique to identify and count the leukemia cells in the image. The author recognized the demerits involved in the watershed transform-based segmentation, K-means clustering algorithm, edge detection using histogram equalizing method, and linear contrast stretching. The image was converted initially to grayscale and then to binary using a threshold value. Area, perimeter, and other features were computed. Overlapping and not overlapping cells were determined using the major axis and minor axis value. Wang et al. [17] proposed a novel cell detection methodology that took into consideration both intensity and shape information of the cell to improve the quality of segmentation. The classification of the cells was done using the Online Support Vector Classifier (OSVC) which was adaptive to changing experimental conditions. A binarization process was done to obtain cell shape information. Next, both intensity and shape information was used for the local maxima generation. At last, the pixels of nuclei were made to move into the gradient vector field which would later converge at those local maxima. For the separation of the touching objects, a seeded watershed algorithm was applied to it. Before the binarization, the cell clusters were separated from the background. A data-driven background correcting algorithm which used the cubic B-spline was used to detect the background. In local maxima generation, the original image was added to the image obtained after the distance transform operation. The local maxima were later generated which became the cell centers. The cell detection problem was reduced by detecting local maxima in the gradient image. The threshold was calculated by using the number of pixels in the original cell. In each local region, only one local maximum was treated as the cell center. The cells were then segmented using the seeded watershed algorithm. The result showed that the seeded watershed algorithm had less over segmentation problem as compared to others. A comparison between the SVM (support vector machine), LASVM, OSVC

was made. The result showed that OSVC had the upper hand over the other two. Wang et al. [18] proposed an approach to identify the lymphoblast by considering both spectral and spatial features. Spatial feature extraction was achieved using the normalization and encoding method. Spatial features were determined using the support vector machine recursive feature elimination (SVM-RFE) algorithm. The identification was made using a marker-based learning vector quantization (MLVQ) neural network. Z. Saeedizadeh et al. [19] proposed an approach for the detection of WBCs using the bottleneck algorithm and K-means. The input image was converted from RGB to HSV color space. Contrast stretch and noise elimination operations were applied later. The global threshold was used to obtain the mask. The image was converted to LAB color space to obtain the nucleus mask. The same procedure was involved, but in the open operation, the objects with an area less than 100 pixels were eliminated. The clumped nuclei were removed using the bottleneck algorithm. Area and convex area were computed along with its ratio. Constraints were set on the ratio to identify touching nuclei. Features like nucleus to cytoplasm ratio and eccentricity were used to classify the normal cell from other marrow cells. Jiang et al. [20] proposed an approach to segment the WBC nucleus using by combining scale-space filtering and watershed clustering. WBC was separated from the remaining cells by utilizing the gray value. The nucleus region was identified using the scale space filtering approach. The cytoplasm region was extracted using the watershed clustering. The complexity in the image space was reduced by applying morphological operations. Kumar et al. [21] proposed a method to highlight the nucleus boundary, which helped in segmenting the nuclei from cell image. The edges of the nuclei were determined using the Teager energy operator. The cytoplasm was extracted using morphological operations which involved a lot of assumptions. Alreza and Karimian et al. [22] proposed a technique to identify the WBC from a blood smear using the color conversion model. The image was converted from RGB to CMYK color space. The authors determined that the WBC nucleus was more significant in contrast to the Y component. The threshold to convert the image to binary was computed using the Zack algorithm. The cytoplasm was later calculated by separating the WBC nucleus from the entire leukocyte region. Reta et al. [23] proposed an approach to extract cell regions using color and texture features. The synthetic and real leukemia images were considered for the experiment. The image was converted to CIE Lab color space to obtain the color features. The structural and stochastic features were extracted using the 2D World Decomposition model. The features were then given as input to the unsupervised binary Markov random field segmentation model. The accuracy of the model was evaluated by comparing the results obtained using the proposed approach with the manual segmentation results. Madhloom et al. [24] proposed a technique to localize

the cell present in the peripheral blood smear image. Several image processing techniques were integrated to achieve localization of the cell. The color features were combined with morphological reconstruction to segment and locate the lymphoblast. Houby [25] proposed an approach to diagnose cancer in different image modalities. Feature selection was performed using Ant Colony Optimization Algorithm. Different classifiers were used to achieve classification, and their performance was compared. Theera-Umpon [26], proposed a technique using fuzzy clustering to achieve cell segmentation. The cell was segmented into two regions namely: nucleus and non-nucleus region. The segmentation was carried out with fuzzy C-means algorithm. Mathematical morphology methods were applied to the resultant image. Naïve Bayes classifier was used to perform classification. Mohammed et al. [27] proposed a technique to segment the cells using the watershed algorithm and optimal thresholding. The problem of over and under-segmentation faced by the watershed algorithm was prevented by reducing the local minima by 1%. Mask subtraction was done to achieve the cytoplasm area. Suryani et al. [28] proposed a Fuzzy rule-based system to diagnose leukemia types. Color identification filters, canny edge detection, and threshold techniques were employed to achieve the separation of WBC from the surrounding RBCs. Features like the WBC area and nucleus to granule ratio were computed to differentiate between the different leukemia cases. Khobragade et al. [29] proposed a technique to identify leukemia by computing various geometrical and statistical features of the cells. The WBCs were separated from the other surrounding components of the blood by using the values of standard deviation and mean. They also identified that mean and standard deviation played a vital role in differentiating a blast from a normal cell. Cuevas et al. [30] proposed an approach to detect the presence of WBCs using the ellipse detector algorithm. The task of finding the presence of multiple WBCs in cluttered blood smear images was considered as a multi ellipse detection problem. Differential evolution algorithm was implemented to convert the problem of detection to an optimization problem. Candidate ellipses were identified using an objective function. Images with different complexities were used for experimentation. Ahasan et al. [31], used the modified watershed transform and morphological closing operation to locate the WBCs present in the blood smear. The touching WBC was separated using the marker-controlled watershed algorithm. The broken WBCs were reconstructed using the morphological closing operation. Area removal technique was used to eliminate the megakaryocytes from the WBC. Guo et al. [32] used a support vector machine (SVM) to achieve the image segmentation. Feature selection was made using sequential minimal optimization (SMO) algorithm. It produced good segmentation accuracy for images clicked under different

illumination conditions and was insensitive to blood smear staining. Sinha and Ramakrishnan [33] achieved an automated leukocyte segmentation using Gaussian mixture modeling (GMM) and EM (expectation maximization) algorithm. It was unsupervised and did not require any parameter tuning.

1.3 Comparative analysis

A comparative analysis of the state of the art segmentation methods and the proposed method is illustrated in Table 1. The advantages and drawbacks associated with the state of the art methods are discussed briefly in this section.

Table 1 Comparison of state of the art methods with the proposed method

Method	Image dataset	Accuracy	No. of images	Reference
Segmentation and classification using artificial neural network	Sina hospital of Hamadan	96.7%	131	Afsar et al. [2]
Segmentation using threshold technique	Image captured from CDC-DPDx	98.4%	100	Alomari et al. [3]
	Image captured from microscope Olympus BX51	92%	20	Sadeghian et al. [14]
Segmentation using Otsu's threshold algorithm	Public Image Dataset	93%	108	Joshi et al. [8]
Segmentation by applying pre-processing at every stage	ALL-IDB	92%	108	Putzu and Ruberto [13]
Segmentation and classification using convolution neural network	ALL-IDB	96.42%	108	Than et al. [15]
Segmentation using binarization and local maxima generation	Images captured from epifluorescence TE2000-E Eclipse microscope	99%	240	Wang et al. [17]
Segmentation using microscopic hyperspectral imaging technology	Ruijin Hospital, Shanghai, China	92.9%	135	Wang et al. [18]
Segmentation using K-means Clustering	Pathology department of Alzahra hospital, Isfahan, Iran	95.28%	50	Saeedzadeh et al. [19]
Segmentation using scale space filtering and watershed clustering	–	98.9%	45	Jiang et al. [20]
Segmentation using Teager energy operator and morphological operations	–	–	–	Kumar et al. [21]
Segmentation using the Zack's threshold and distance conversion technique	Imam Reza (AS) hospital, Mashhad	93%	29	Alreza and Karimian [22]
Segmentation using Markov random fields	IMSS San Jose Cells	95%	200	Reta et al. [23]
Segmentation using lymphoblast localization	University of Malaya Medical Center (UMMC)	95%	180	Madhloom et al. [24]
Segmentation using color features and morphological reconstruction	ALL-IDB	96.25	108	Houby [25]
Segmentation using Fuzzy C-means clustering and mathematical morphology	University of Missouri Ellis-Fischel Cancer Center	59.55%	431	Theera-Umpon [26]
Segmentation using watershed algorithm and optimal thresholding	Department of Pathology and Laboratory Medicine, University of Calgary and Calgary Laboratory Services (CLS), Calgary, AB, Canada	92% for nucleus segmentation 99.85% for cytoplasm segmentation	140	Mohammed et al. [27]
Segmentation using fuzzy rule-based system	Karang Anyar General Hospital	83.65%	104	Suryani et al. [28]
Segmentation using Otsu's threshold and Sobel operator for edge detection.	–	91%	–	Khobragade et al. [29]
Segmentation using differential evolution algorithm	ASH Image Bank	98.26%	50	Cuevas et al. [30]
Segmentation using marker controlled watershed and morphological closing operation	University of degli Studi di Milano, Italy	88.57%	70	Ahasan et al. [31]
Segmentation using GMM and EM algorithm	University of Kaiserslautern, Germany	80%	115	Sinha and Ramkrishnan [33]
Proposed method	Atlas of hematology ALL- IDB and the Internet	98.6%	600	

Alomari et al. [3] proposed a technique to segment and quantify the WBCs and RBCs. A threshold value was computed through experimental results. Although it achieved a good accuracy in quantification, the threshold value chosen failed to satisfy all the images. Belekar and Chougule [4] proposed a segmentation technique using SMMT operator and granulometric analysis. Even though it achieved a good segmentation accuracy, the cropping of the cells to have one cell per image was a problem. The process was time-consuming and tedious when many cells were present. Bhagvathi and Thomas [5] applied fuzzy logic to detect and count the WBCs. The edges of the cell were determined using the fuzzy rules. However, the work was not successful in generating accurate WBC masks for some images. Bhamare and Patil et al. [6] proposed a technique to separate the WBCs from the surrounding RBCs by considering the gray level intensity. The threshold was determined manually. However, the manually defined threshold was not precise. Joshi et al. [8] proposed an algorithm for segmentation of white blood cells using global Otsu's threshold. Circularity, perimeter, and area were computed. However, the work did not calculate the ratio between the nucleus area and cytoplasm area which plays a vital role in differentiating a lymphoblast from a lymphocyte. Mulik et al. [11] proposed a technique to segment the WBCs using the K-means algorithm. The number of clusters chosen was four. Even though it achieved good segmentation accuracy, the stage of separation of WBC cytoplasm from the surrounding components was not taken into consideration. Nasir et al. [12] proposed a technique to extract the WBCs from the blood smear image by constructing a histogram. The pixel values in the range 0–89 were considered as WBC. Even though it achieved reasonable accuracy, the threshold value failed to generate precise results for some images. Putzu and Ruberto [13] proposed an approach to segment the WBCs using the threshold value computed using Zack's algorithm. The grouped leukocytes were eliminated by calculating the roundness value. Although it achieved good accuracy in segmenting the WBC nucleus, WBC cytoplasm was extracted after cropping the image. The process of cropping was time-consuming. Sadeghian et al. [14] proposed a technique to segment the WBCs. Canny edge detection and GVF snake were used to extract the nucleus. The process involved the cropping of the cells. Even though it helped in eliminating the surrounding artifacts, the process was tedious when smear had a large number of cells. Thanh et al. [15] proposed an approach to detect acute leukemia using the convolution neural network. Although it was able to classify the cells as normal and abnormal, it did not discuss the classification of leukemia into its four types. Jiang et al. [20] proposed a technique to segment the WBC

nuclei using scale space clustering and watershed clustering. The method was successful in identifying the various WBC regions, but it involved the process of extracting sub-images by extending the area of the nucleus which was not very precise. Kumar et al. [21] proposed a new approach to segment the WBC nuclei and cytoplasm using Teager energy operator and morphological operation. The work was successful in identifying the nuclear boundary, but the cytoplasm detection involved a lot of assumptions which was not reasonable. Alreza and Karimian [22] proposed a technique to count the WBCs. The method was able to detect the WBCs present in the blood smear, but the method failed when the WBCs were present in groups. It also involved the process of cropping the cells to obtain the cytoplasm which was not a feasible option when the smear had a large cell population. Reta et al. [23] proposed an approach to extract cell regions using color and texture features. 2D world Decomposition model was used to obtain structural and stochastic features. Even though the work was successful in identifying the cell and nucleus, it worked accurately for images with a single cell. It failed to generate accurate results for images with cell population. Madhloom et al. [24] proposed an approach to localize and segment the lymphoblast in the blood smear image. Even though the method was able to localize the cells in the smear, the threshold value used to convert the image to binary failed to satisfy all the images. Houby [25] proposed an approach to classifying the cell as normal or abnormal. Even though it was successful in the cell classification, it involved the cropping of the region of interest which was tedious. Theera-Umpon [26] proposed a technique to segment the cell by combining the Fuzzy C-means clustering algorithm and mathematical morphology operation. The approach was able to generate better results when compared to the manual segmentation results. However, the method worked for images with a single cell. The computational complexity would also increase with the number of clusters. Mohammed et al. [27] proposed an approach to diagnosing chronic lymphocytic leukemia using the watershed algorithm and optimal thresholding. Even though good segmentation accuracy was achieved, the method was suitable for only one cell per image. It also failed if the cell (WBC) was tightly attached to the red blood cell. Khobragade et al. [29] proposed a technique to identify leukemia by computing various geometrical and statistical features of the cells. Even though good segmentation accuracy was obtained, the method did not take into consideration the computation of cytoplasm area which played a vital role in differentiating a blast from a normal cell. Cuevas et al. [30] proposed a methodology to detect the WBCs using the ellipse detection algorithm. Even though the technique was successful in detecting WBCs, it did not mention the feature extraction stage. Wang et al. [34] proposed a

technique to identify the WBCs using boundary support vectors. The method achieved good results when compared to InterOtsu's algorithm. But it did not consider any evaluation metrics to measure the performance of their algorithm. Ahasan et al. [31] proposed a technique to identify the WBCs from the peripheral smear using the modified watershed transform and morphological closing operation. The method was able to identify the WBC's nucleus, but it did not take into consideration the segmentation of cytoplasm of WBC. Guo et al. [32] proposed a technique to segment the smear image using support vector machine. Even though it was able to work for images with different stains, it was not able to generate precise results for nucleus detection due to the variation in its color. Sinha and Ramakrishnan [33] proposed a technique to segment the leukocytes in the blood smear using GMM and EM algorithm. Even though the algorithm did not require any parameter tuning and was completely unsupervised, it did not produce accurate results for all the stains. It also needs to be improvised to deal with clustered cells.

In contrast to the segmentation approaches proposed in the state of the art methods, the proposed segmentation algorithm is simple and produces precise results irrespective of the stain used and cell abnormalities. An overview of different segmentation algorithms that can be applied over blood smear images is presented in Table 1.

1.4 Limitations in the existing system

Determination of the stage of the tumor at the earliest plays a vital role in the field of medicine. Due to the complication of the disease, it goes undiagnosed. The high variability of the size, shape, location, and edge makes the diagnosis complicated. The present procedures incur a lot of cost and time. Manual processes are prone to sampling errors. Automated devices fail to identify abnormal cells. The delay in the diagnosis leads to further division of the stem cells which aggravate the condition. The accurate segmentation of the blood smear images cannot be achieved using a single algorithm due to its morphological complexity.

1.5 Contributions of the paper

The main aim of the work is to achieve an accurate diagnosis of malignancy in case of any cell abnormalities, variations in the location of cells and with any stains. The main contributions of the paper are as follows:

- An automated algorithm using K-medoids to segment the nucleus of all shapes and deformations from blood smear stained with different stains is proposed.
- An automated algorithm to segment the cytoplasm without manual cropping of cells is proposed. To the best knowledge of the authors, the algorithm to segment cytoplasm of all the cells without manual cropping is first of its kind in literature. Finally, we show that the proposed system can generate precise results for blood smear images stained with different stains.

Section 2 describes the proposed algorithm. Section 3 illustrates the results. Section 4 gives a brief description of the pros and cons of state of the art techniques. Section 5 concludes the paper.

2 Methodology

The proposed work intends to predict the stage of cancer using blood smear images. In the case of acute lymphoblastic leukemia, it plans to classify the instances into mainly three classes which are ALL-L1, ALL-L2, and ALL-L3. The features of the WBC play a vital role in the diagnosis of leukemia. The proposed model achieves its objectives by following two steps. Firstly, it segments the WBC nucleus and extracts its features. Secondly, it segments the WBC cytoplasm. To segment WBC cytoplasm and its nuclei, a three-part algorithm including nucleus segmentation, splitting of touching nuclei, and cell and cytoplasm segmentation is given. The proposed system is depicted in Fig. 2. The process of nucleus segmentation is explained in Sect. 2.2. The stage of separation of touching nuclei is specified in Sect. 2.2 under image post-processing step. The segmentation of cytoplasm is described in Sect. 2.3. Section 2.4 describes the feature extraction stage. The model used for classification and prediction is illustrated in Fig. 3. The performance of random forest, K-nearest neighbor (KNN), naïve Bayes, and the decision tree is measured regarding their accuracy in classifying the instances. The algorithm with the highest accuracy is chosen.

2.1 Description of the data set

The blood smear images are collected from Atlas of Hematology [35], ALL-IDB [36], and the Internet. The dataset consists of 600 images. Matlab with image processing toolbox is used to perform the feature extraction of blood smear image. Images with normal dimension and magnified dimension are considered for experimentation. The images with standard dimension have a size of 459×306 and resolution of 96 dpi. The magnified images have a size of 640×480 with a resolution of 96 dpi.

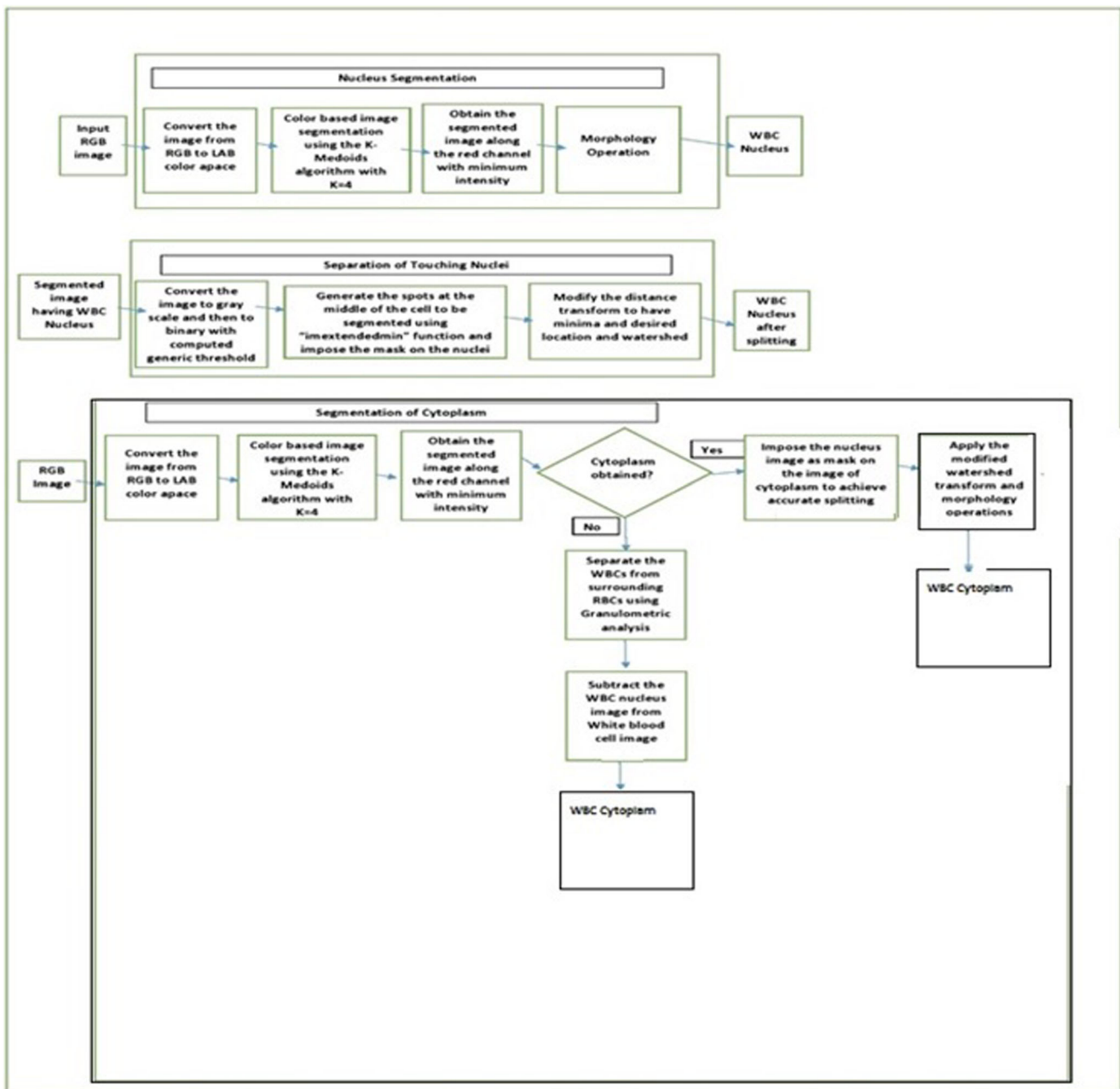


Fig. 2 Proposed system

2.2 Segmentation of nucleus of a white blood cell from blood smear image

Fig 4 depicts the significant steps required to extract the nucleus from the image with standard dimension and magnified dimension. Conversion of the image from RGB to LAB color space takes place in image pre-processing step. Color-based image segmentation using K-medoids algorithm takes place in image segmentation step. The application of morphology operation and separation of touching nuclei takes place in the image post-processing stage

- a. Image pre-processing: The image is converted from RGB to LAB color space. It is a color-opponent space which consists of a luminosity layer ‘L*,’ chromaticity-layer ‘a*’ indicating where the color falls along the red-green axis, and chromaticity-layer ‘b*’ indicating where the color falls along the blue-yellow axis. It is derived from CIE XYZ tristimulus values. Any color on the CIE chromaticity diagram can be considered to be a mixture of the three CIE primaries, X, Y, Z. That mixture may be specified by three numbers X, Y, Z called tristimulus values. The color is represented by the Eq. (1). The tristimulus values are

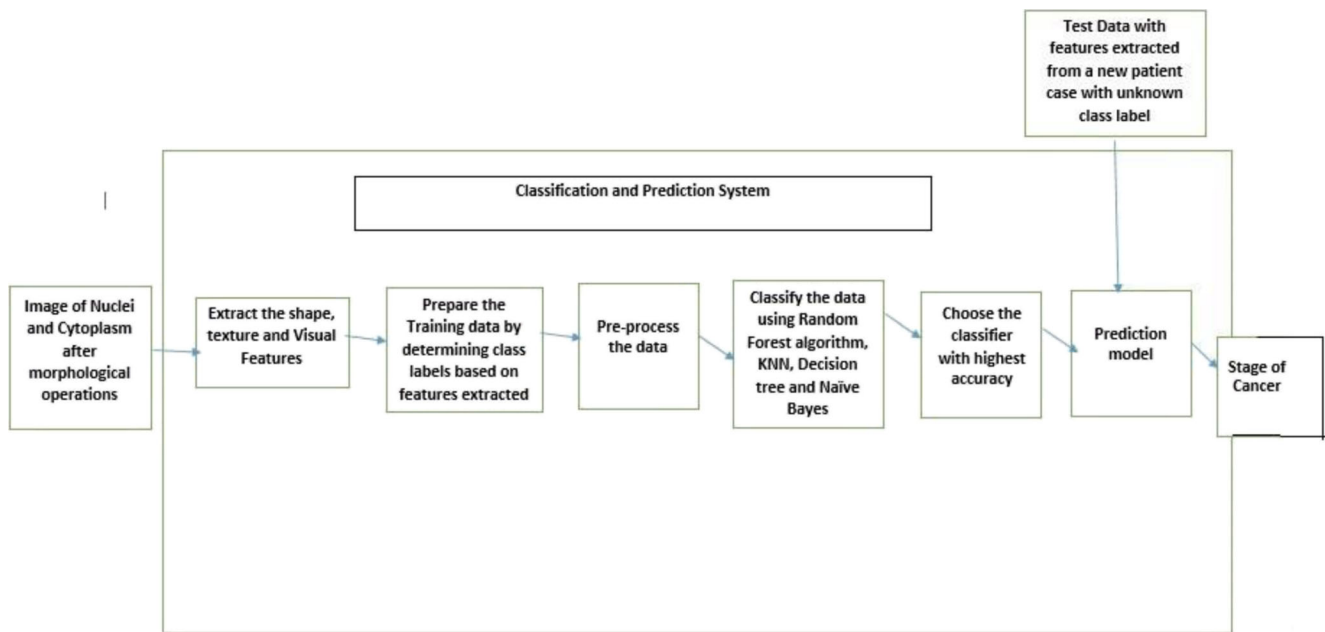


Fig. 3 Classification and prediction system

computed using the CIE primaries [37] and associated color matching functions.

$$c = xx + yy + zz \tag{1}$$

b. Image segmentation: Color image segmentation is performed using K-medoids algorithm. Initially, a comparison is made between K-means and K-medoids algorithm. The results proved that the K-means algorithm took less execution time when compared to the K-medoids algorithm. The main downside of the K-means algorithm is its less robustness. Since in the medical field accuracy is of utmost importance, the proposed work uses K-medoids algorithm to segment the image. The image is reshaped into a vector and is fed to the K-medoids algorithm. K-medoids clustering requires the number of clusters to be partitioned and a distance metric to quantify how close two objects are to each other. Euclidean distance is used to measure the similarity (distance metric). Since the color information exists in the ‘a*b*’ space, pixels with ‘a*’ and ‘b*’ values are chosen. K-medoids is applied to the image with 4 clusters to obtain WBC nucleus, WBC cytoplasm, RBCs, and background separately in each cluster. The pseudo code for the K-medoids algorithm is given in algorithm 1. The similarity in image regions is considered here. The total cost of configuration denotes the cost involved in swapping a pair of non-selected object ‘o’ and selected object ‘m.’ The difference in absolute error-value if a current representative object is replaced by a non-representative object determines the cost function. Clustering is done on color values of the pixels using

the distance formula. For every object in the input, K-medoids returns an index corresponding to a cluster. Every pixel in the image is labeled with an index corresponding to a cluster. The objects in the image are separated using the pixel labels. The mean value of the segmented images is obtained along the red channel. Separation of the nucleus from the cytoplasm takes place at this stage. After the analysis, it was found that the segmented image with minimum intensity along the red channel had the WBC nuclei (Fig. 5). Time consumed by both the algorithms concerning the number of white blood cells is depicted. The result obtained after the application of the K-medoids algorithm on the input image is shown in Fig. 6.

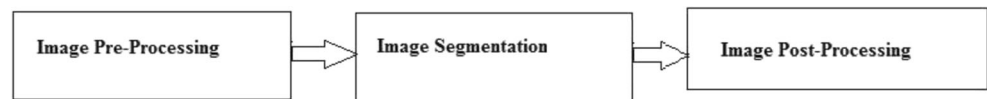
Algorithm 1 Pseudo code of K-medoids algorithm

Input: S (image as vector), K (number of cluster), distance metric. Output: Clusters

- 1 Initialize: randomly select k of the n data points as the medoids.
- 2 Associate each data point to the closest medoid. (“closest” here is defined using Euclidean distance).
- 3 for each Medoid m do
- 4 for each non-Medoid data point o: do
- 5 Swap m and o and compute the total cost of the configuration End for
- 6 End for
- 7 Select the configuration with the lowest cost.
- 8 Repeat steps 2 to 7 until there is no change in the Medoid.

c. Image post-processing: The segmented image having the nucleus is converted to grayscale. The image is

Fig. 4 Stages involved in the extraction of the nucleus of white blood cell



converted to binary by using a level value of 0.1. The level value of 0.1 is chosen after the experimental analysis. Area open function with a value of 200 is used to remove small particles. Area open function takes two parameters. The first one is the binary image and the second one is the number of pixels ‘P’. It intends to remove all connected components that have fewer than P pixels from the binary image BW, producing another output binary image. The connectivity value for two dimensions is 8 and 26 for three dimensions [38]. Border clear operation is performed as border pixels do not contribute much to the analysis. The primary purpose of the clear border operation is to suppress the structures that are lighter than their surroundings and the ones that are connected to the image border [39]. The touching objects are separated using modified watershed transform [40] with a mask having small spots present at the center of each cell to be segmented (to generate local minima only at the desired locations). The raw watershed transform tends to perform the over-segmentation. It is because even the tiny local minima become the catchment basin. In this work, the watershed transform is modified so that no minima occur at the filtered out location. After the operation, the touching objects are separated accurately. Boundaries of the object are accurately obtained using the “bwboundaries” function. The exterior boundaries of objects and holes inside the objects are detected using this function [41]. The resulting image is given as input to “regionprops” [42] to extract texture and shape features. The nucleus is labeled using the “bwlabeln” function [43]. It helps the pathologist to differentiate between the touching nuclei. They can relate the extracted feature values with labeled nucleus if needed. It returns a label matrix with labels for

connected components in the binary image. The mask values, level value, and area open function value are chosen after doing the experimental analysis over the images present in the training set. For the images with magnified dimension, a value of 500 is given as an input to the area open function. Figure 7 depicts the WBC nucleus extracted using the proposed algorithm.

2.3 Segmentation of cytoplasm of white blood cell from blood smear image

The nucleus to cytoplasm ratio plays a vital role in differentiation a lymphoblast from a lymphocyte. The WBC cytoplasm has to be extracted to compute the ratio. The proposed work aims to segment the cytoplasm from the cells with variations in shape, locations, and most importantly stained using different stains. Two different techniques are proposed to achieve the objective.

- (a) Obtain the cytoplasm by extracting the segmented image generated using the K-medoids algorithm.
- (b) Obtain the cytoplasm after separating the WBCs from surrounding RBCs utilizing the size-based analysis (granulometric analysis).
 - (a) Cytoplasm obtained from the cluster (using K-medoids algorithm): for some blood smear images, the WBC cytoplasm is obtained directly after the application of the K-medoids algorithm. It is mainly due to staining used. To achieve accurate results, the precise splitting of cytoplasm is a must. It is a two-step process. Firstly, the WBC nucleus mask obtained from the nucleus segmentation stage is imposed

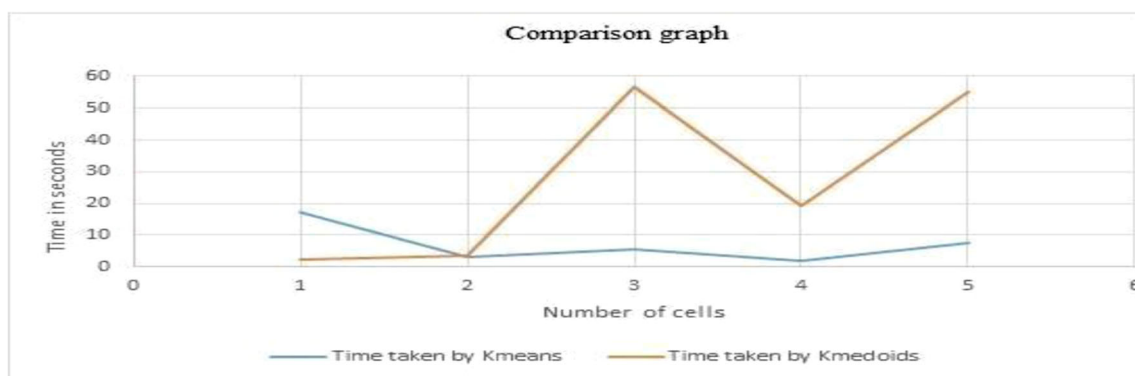


Fig. 5 Comparison between K-medoids and K-means algorithm concerning the time taken for execution

on the resulting image. Secondly, the image is modified so that the mask is forced to be the only local minima in the image. The watershed transform is later applied to the resulting image. The problem of over-segmentation faced by existing watershed transform can be overcome using this approach. The steps to accurately segment the cytoplasm obtained from the cluster are illustrated in Fig. 8. The pseudo code to achieve it is given in algorithm 2. An instance of extraction of cytoplasm using this technique from ALL-L2 type is depicted in Fig. 9.

Algorithm 2 Pseudo code to obtain the cytoplasm from the cluster

- Input: Segmented image along the red channel with second minimum intensity
 Output: Cytoplasm
1. The input image is converted to grayscale.
 2. Next, it is converted to binary using a level of 0.2. The level value is computed by analyzing the images in the training data set and is generic.
 3. The objects with an area less than 200 pixels are eliminated. The pixels at the border are eliminated using the clear border operation as they do not contribute to the feature computation. A median filter with the 2 * 2 neighborhood is used to eliminate noise.
 4. Holes are filled to enhance the segmentation. Modified watershed transform was applied to it.
 5. The resulting image is checked to see if cytoplasm is obtained accurately. If a precise result is obtained, the image is given as output. Else steps from 6 to 10 are executed.
 6. Image of cytoplasm obtained from step 1 is converted to grayscale. The contrast of the image is improved using Adaptive histogram equalization. The resulting image is converted to binary using the Otsu's threshold.
 7. Holes are filled in the image using imfill [44]. An area open function with value 10 is used to remove small objects.
 8. The perimeter of the image is obtained to identify the boundary accurately. The perimeter is laid on the equalized image with color specification values [0.2 0.2 0.2]. The color specification value is used to determine the color used for the overlay.
 9. The final image of the nucleus after segmentation stage is obtained. An extended maximum of the image is achieved using the "imextendedmax." The function returns the extended-maxima transform for the image, which is the regional maxima of the H-maxima transform [45]. It forms the mask.
 10. An area open function with a value of 3 to eliminate small unwanted particles is used. The mask is laid on the image with color specification values [1 1 1].
 11. The resulting image is complemented to have a better visualization. The watershed ridge lines are made to appear at locations where the mask value is non-zero using the "imimposemin" [46] function.
 12. Watershed transform is applied to the resulting image.
-

- (b). Compute the cytoplasm area after separating it from the surrounding red blood cells:

In some instances, cytoplasm may not be obtained directly as a result of the K-medoids algorithm. It mainly happens due

to variation in staining characteristics. The granulometric analysis is performed to achieve the results. The medical studies prove that WBCs are larger than surrounding RBCs. Granulometry is an approach used to compute the size of RBCs utilizing a series of mathematical operations. Once the size is calculated, the WBCs are separated from the surrounding RBCs. The pseudo code to obtain the area of all the objects in the image is given in algorithm 3

Algorithm 3 Pseudo code to obtain the area of all the objects in the input image

- Input: Input image
 Output: Cytoplasm
1. The input image is converted to grayscale, and the contrast of the image is enhanced using the stretch limits in the range 0.1–0.7. [47]
 2. Histogram equalization is applied to the resulting image [48]. In this, the intensity histogram is made to get the desired shape by altering the image. The dynamic range and contrast of the image are modified here. It is used in the correction of non-linear effects introduced by a display system and image comparison processes.
 3. The resulting image is converted to binary using the global Otsu's threshold and is complemented.
 4. The noise is eliminated using an area open function with value 200. A value of 7000 is used for images with magnified dimension.
 4. The touching objects are separated using the modified watershed transform.
 5. The area of the objects in the image is calculated using the "regionprops" function and is placed in an array in descending order.
-

The global Otsu's threshold uses a threshold that helps to reduce the intraclass variance of the black and white pixels. It is the weighted sum of variances of the two classes. It is computed using Eq. 2.

$$\sigma_w^2(t) = w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t) \tag{2}$$

The probabilities of the two classes are given by the weights w_0 and w_1 . The threshold is represented by t . Variances of the two classes are given by σ_0^2 and σ_1^2 . The class probability is computed from the L bins of the histogram. It is computed using the Eqs. 3 and 4.

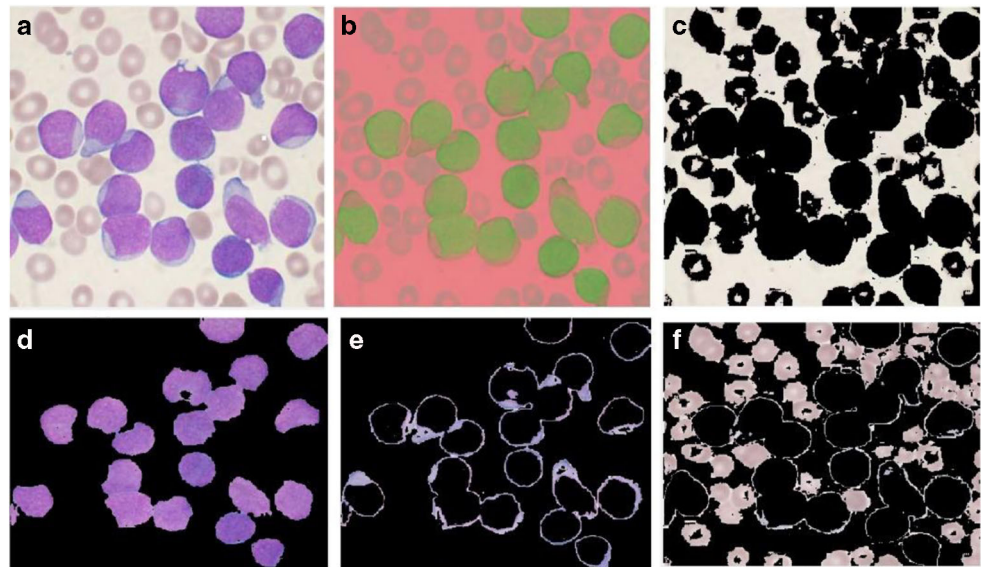
$$w_0(t) = \sum_{i=0}^{t-1} p(i) \tag{3}$$

$$w_1(t) = \sum_{i=t}^{L-1} p(i) \tag{4}$$

The intra class variance is minimized as given in Eq. 5.

$$\sigma_b^2(t) = \sigma_2 - \sigma_w^2(t) = w(\mu_0 - \mu_T^2)_+ w_1(\mu_1 - \mu_T^2) \tag{5}$$

Fig. 6 **a** Input Image. **b** The image in LAB color space. **c** Cluster 1 has a background. **d** Cluster 2 having the WBC nucleus. **e** Cluster 3 having the WBC cytoplasm. **f** Cluster 4 having surrounding objects present in the blood smear including RBC



where w is the class probabilities and μ_{-} is the class mean. The class mean is computed using Eqs. 6, 7, and 8.

$$\mu_0(t) = \sum_{i=0}^{t-1} i.p(i)/w_0 \quad (6)$$

$$\mu_1(t) = \sum_{i=t}^{L-1} i.p(i)/w_1 \quad (7)$$

$$\mu_{-T} = \sum_{i=0}^{L-1} i.p(i)/w_1 \quad (8)$$

An XOR operation is used to separate the white blood cells from red blood cells. The procedure is given in algorithm 4.

Algorithm 4: Pseudo code for the separation of RBCs from other components in the blood

Input: Binary image having all the cells obtained after image post-processing

Output: Image with only RBCs

1. Acquire the number of WBC nucleus present in the input image.
 2. for 1 to Number_of_WBC_nucleus do
 3. Apply XOR operation on the binary image with lower bound value as an area open function with two parameters. One is the binary image, and the other is one. Upper bound value as another area open function with two parameters. One is the binary image, and the other is the area of the cells that is computed using algorithm 3.
 4. End for
-

The cytoplasm is extracted using Eqs. (9) and (10).

White blood cell

$$= \text{Image with all the cells} - \text{Image with only RBCs} \quad (9)$$

WBC Cytoplasm

$$= \text{Image with WBC} - \text{Image with WBC nucleus} \quad (10)$$

Figure 10 shows the cytoplasm extracted by following the steps shown in the above two algorithms.

2.4 Feature extraction

Feature extraction forms the significant step in the development of computer-aided diagnosis tool. In the proposed work, shape features, visual features, and texture features are extracted. Table 2 depicts the features extracted. The description is given below.

2.4.1 Shape features

Eccentricity This parameter is used to measure to what extent the shape of a nucleus deviates from being circular. Lymphoblasts are eccentric, and Lymphocytes are concentric. The cancerous cells will have eccentricity below 0.6, and noncancerous cells will have above 0.6. The eccentricity value is fixed after consulting the domain experts. The Eq. (11) shows the computation of eccentricity.

$$\text{Eccentricity} = \frac{\text{Length of Major axis}}{\text{Length of Minor axis}} \quad (11)$$

Elongation It specifies the lengthening of a cell during the cell differentiation process. The pseudo code to obtain elongation value is given in algorithm 5.

Algorithm 5 Pseudo code for the calculation of elongation factor

Input: Binary image of white blood cells output: elongation value of cells.
 1 Remove unwanted particles using an area open function with value 500.
 The author determines the value of area open function after experimentation over the images in training dataset. Label the objects and identify their boundary.
 2 Measure the centroid value of each object and obtain the number of Blobs.
 3 for each BLOB do
 4 Obtain the x coordinate of the centroid and the y coordinate of the centroid.
 5 Get the distances of boundary pixels from the centroid.
 6 Scan the boundary to identify the pixel that is farthest and closest from the centroid.
 7 Obtain the maximum and minimum value of the distance.
 8 Record the above values in an array.
 9 End for
 10 for each entry in the array do
 11 Calculate the elongation value by estimating the ratio between the maximum distance and minimum distance.
 12 End for

Form factor It is the ratio between area and perimeter. The Eq. (12) computes the form factor.

$$Form\ factor = (4 * \pi * area) / (perimeter * perimeter) \quad (12)$$

Centroid Represents the center mass of the region. It is the arithmetic mean of all (x, y) coordinates in the region R. It is computed using Eq. (13). Cancerous cells will not have an extensive variation between x and y coordinates. Healthy cells will have a full variation.

$$\bar{x} = \frac{1}{|R|} * \sum_{(u,v) \in R} u, \bar{y} = \frac{1}{|R|} * \sum_{(u,v) \in R} v \quad (13)$$

Orientation It is the angle between the x-axis and the major axis of the ellipse that has the same second moments as the region. It describes the direction of the major axis that runs through centroid and along the widest part of the region.

Major axis length It gives the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region. It is computed using the Eq. (14).

$$\sqrt{(x1-x2)^2 + (y1-y2)^2} \quad (14)$$

where (x1, y1) and (x2, y2) are the two end points of the major axis.

Area Actual number of pixels in the region. It is computed using the Eq. (15).

$$A(R) = |R| = N \quad (15)$$

Solidity Ratio between area and convex area. It is computed using Eq. (16).

$$Solidity = \frac{Area}{convex\ area} \quad (16)$$

Equivalent diameter Returns diameter of a circle with the same area as the region. It is computed using the Eq. (17).

$$Equivalent\ diameter = 2 * \sqrt{\frac{Area}{\pi}} \quad (17)$$

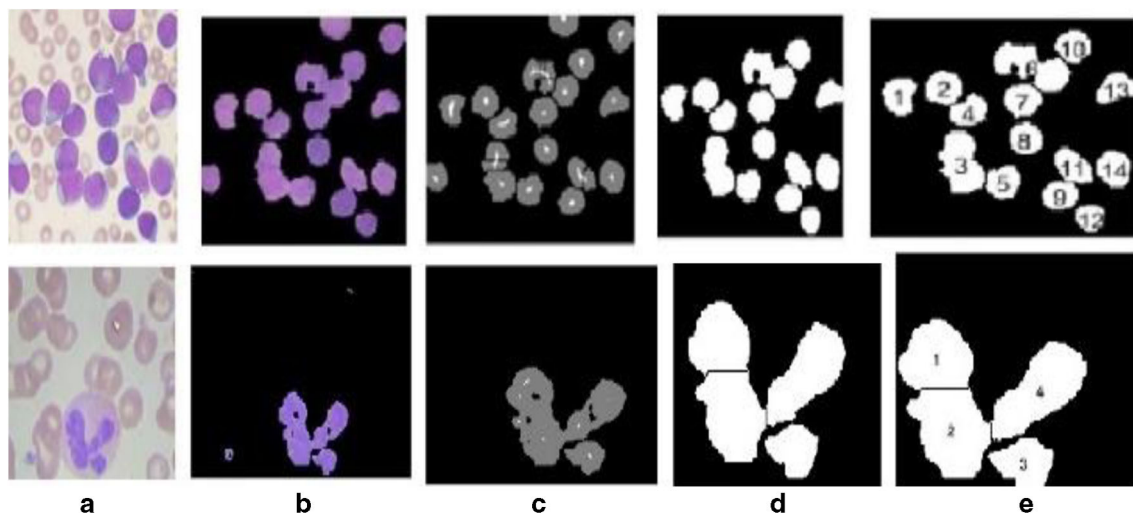


Fig. 7 Extraction of nuclei **a** blood smear image and **b** WBC nuclei. **c** Mask with small spots at the center of the cell. **d** Binary image of nuclei after post-processing. **e** Labeled nuclei

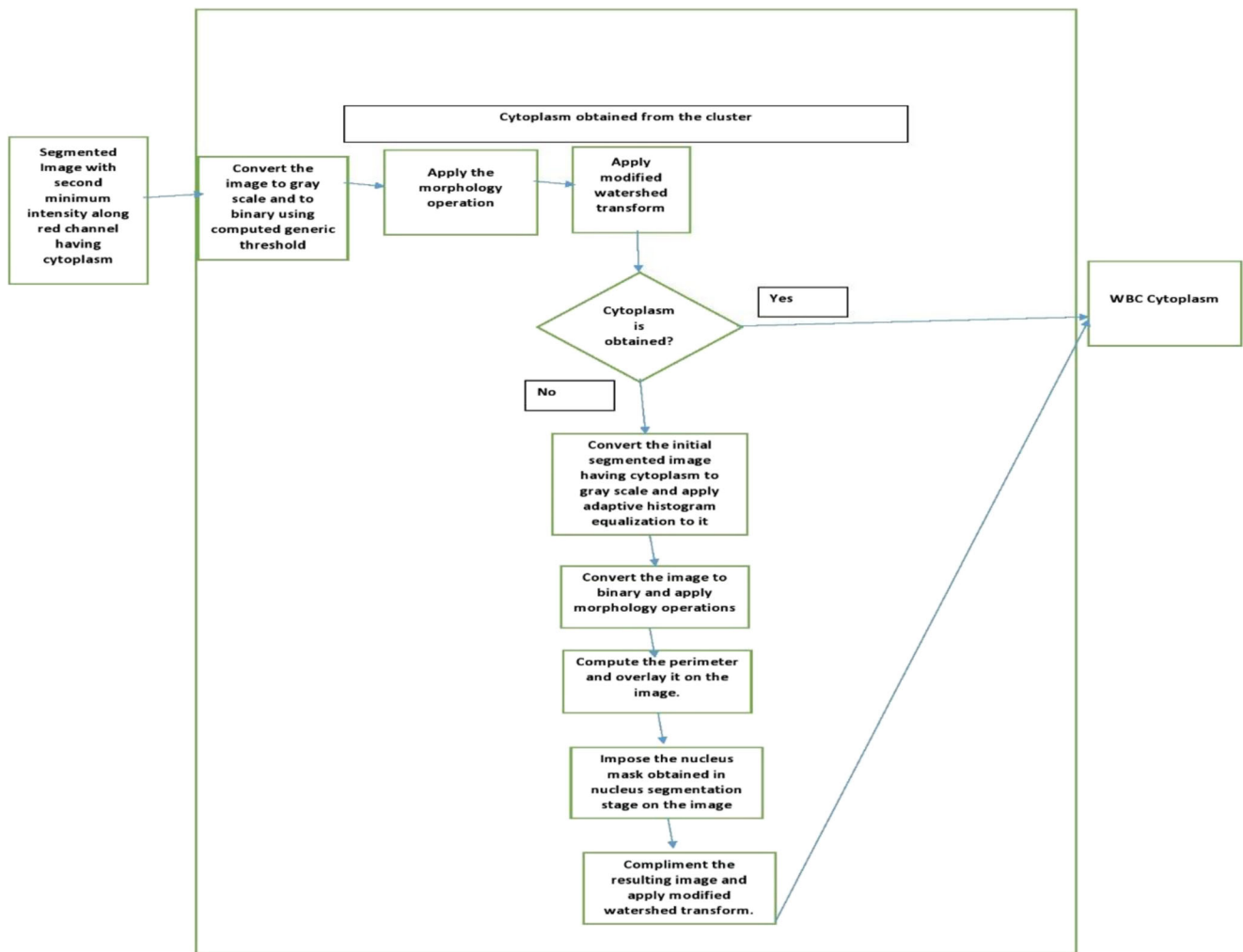


Fig. 8 The steps to accurately segment cytoplasm obtained from the cluster

Perimeter Specifies the distance around the boundary of the region. It is computed using the Eq. (18).

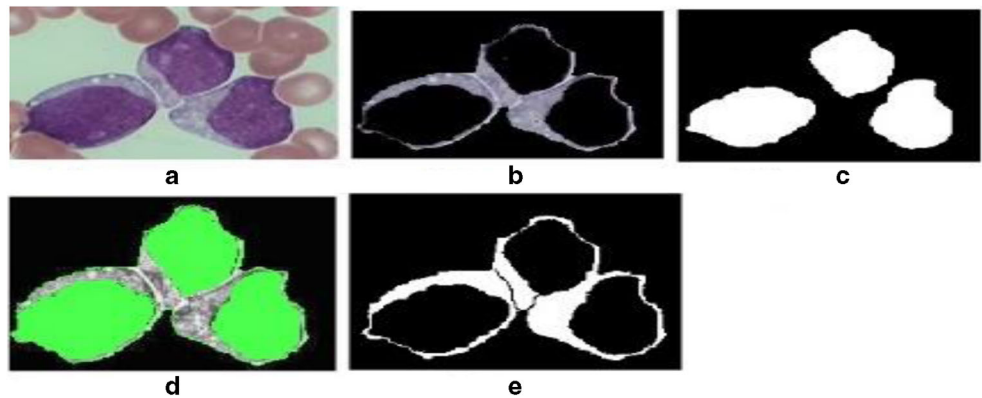
$$B_p = \sum_{i=1}^N d_i \tag{18}$$

where B_p is the perimeter of the object or shape and d_i is the pixel count.

Minor axis length Specifies the length of the minor axis of the ellipse that has the same normalized second central moments as the region. It is computed using the Eq. (19).

$$\sqrt{(x1-x2)^2 + (y1-y2)^2} \tag{19}$$

Fig. 9 **a** Input Image. **b** Image of the cytoplasm obtained from the K-medoids algorithm. **c** Segmented image of WBC nucleus used as mask. **d** Imposition of the mask. **e** Cytoplasm after the accurate splitting



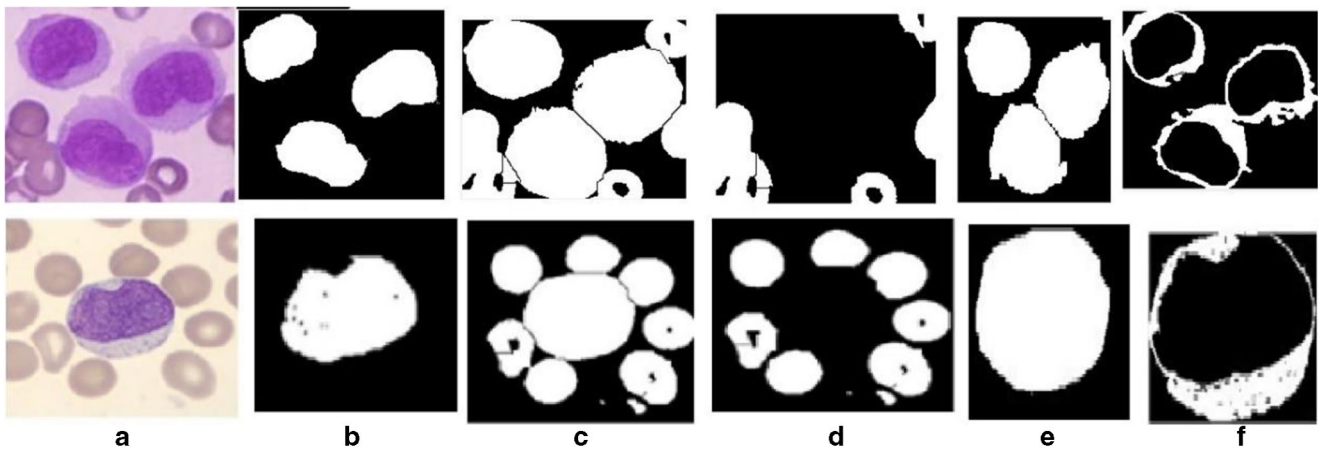


Fig. 10 **a** Original Image. **b** WBC nucleus. **c** Binary image with all the cells. **d** Binary image with only RBCs as a result of XOR operation. **e** WBC cell. **f** WBC cytoplasm

where (x_1, y_1) and (x_2, y_2) are the two end points of the minor axis.

Roundness The measure of how closely the shape of an object approaches that of a circle. It is computed using the Eq. (20).

$$Roundness(R) = 4 * \pi * A(R) / P^2(R) \tag{20}$$

Radius It measures the radius of the cells. It is computed using the Eq. (21).

$$Radius = Equivalent\ diameter / 2 \tag{21}$$

Cytoplasm area Measures area of the cytoplasm which is used in calculating the N/C ratio.

N/C ratio Specifies the ratio between the area of the nucleus and cytoplasm. It is computed using the Eq. (22). It plays a significant role in the determination of the stage of a tumor. If the N/C ratio is greater than 0.85, then it is classified as ALL-L1, and if it is below 0.85 and above 0.5, it is classified to be either ALL-L2 or ALL-L3 type depending upon the other features. Healthy cells will have a ratio between 0.1 and 0.5. The range of N/C values is obtained after consulting the domain experts (doctors).

$$\frac{N}{C} \text{ ratio} = \frac{Area\ of\ the\ nucleus}{Area\ of\ cytoplasm} \tag{22}$$

2.4.2 Visual features

The pathologist decides these features by seeing the cell.

Size of the blast Specifies the size of the cell that if the cell is large or small.

The shape of the membrane Specifies if the nuclear membrane is regular shaped or irregular. ALL-L1 have regularly shaped membrane compared to the other two.

Vacuoles Specifies if hole-like structures are present in the nucleus. ALL-L3 has the hole-like structures in the blasts.

2.4.3 Pixel value measurements

Maximum intensity Returns a scalar that specifies the value of the pixel with the highest intensity in the region.

Minimum intensity Returns a scalar that defines the value of the pixel with the least intensity in the region.

Table 2 List of extracted features

Id	Feature	Id	Feature
1	Eccentricity	15	Roundness
2	Centroid	16	Elongation
3	Orientation	17	Entropy of image
4	Major axis length	18	Mean intensity
5	Area	19	Contrast of image
6	Solidity	20	Correlation of image
7	Equivalent diameter	21	Energy of image
8	Perimeter	22	Homogeneity
9	Minor axis length	23	Radius
10	Maximum intensity	24	Cytoplasm area
11	Minimum intensity	25	N/C ratio
12	Form factor	26	Size of blast
13	Shape of membrane	27	Vacuoles
14	Hausdorff dimension	28	Class

Mean intensity Specifies the mean of all the intensity values in the region.

These features are computed using the “regionprops” function. This function is used to measure properties for each region in the image. The output is a struct array where one element is present per region. The field names represent the measured properties.

2.4.4 Texture features

Entropy of the image It is a statistical measure of randomness that can be used to characterize the texture of the input image. It is computed using the Eq. (23).

$$Entropy = -\sum p * \log_2(p) \tag{23}$$

where p indicates the count of the histogram.

Contrast of the image It is the intensity contrast between a pixel and its neighbor over the whole image. The contrast of the image is computed using Eq. 24.

$$Contrast = \sum_{i,j=0}^{N-1} P_{i,j}(i,j)^2 \tag{24}$$

P_{ij} Element i,j of the normalized symmetrical GLCM [49].
 N Number of gray levels in the image

Correlation of the image: Returns a measure of how correlated a pixel is to its neighbor over the whole image. The correlation of image is computed using Eq. 25.

$$Correlation = \frac{\sum_{i,j=0}^{N-1} P_{i,j}(i-\mu)(j-\mu)}{\sigma^2} \tag{25}$$

P_{ij} Element i,j of the normalized symmetrical GLCM [49]
 μ represents the GLCM mean. It is computed as given in Eq. (26).

$$\mu = \sum_{i,j=0}^{N-1} i.P_{i,j} \tag{26}$$

σ^2 Represents the variance of the intensities of all reference pixels in the relationships that contributed to the GLCM. It is computed as given in Eq. (27).

$$\sigma^2 = \sum_{i,j=0}^{N-1} P_{i,j}(i-\mu)^2 \tag{27}$$

Energy of the image: Specifies the uniformity of the image. It is computed using the Eq. (28).

$$Energy = \sum_{i,j=0}^{N-1} P_{i,j}^2 \tag{28}$$

Table 3 Ranked attributes

Information gain	Attribute name
0.9453	N/C ratio
0.6963	Size of blast
0.6326	Membrane
0.2517	Cytoplasm area
0.2297	Vacuoles
0.2273	Area
0.2223	Equivalent diameter
0.2223	Radius
0.1637	Minor axis length
0.0929	Perimeter
0.0815	Major axis length
0.0735	Solidity
0.0629	Mean intensity
0.0617	Elongation
0.0435	Minimum intensity
0	Centroid y
0	Centroid x
0	Maximum intensity
0	Orientation
0	Form factor
0	Eccentricity
0	Energy of image
0	Contrast of image
0	Correlation of image
0	Hausdorff dimension
0	Roundness
0	Entropy
0	Homogeneity
0	Id

Homogeneity (a) Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. It is computed using the Eq. (29).

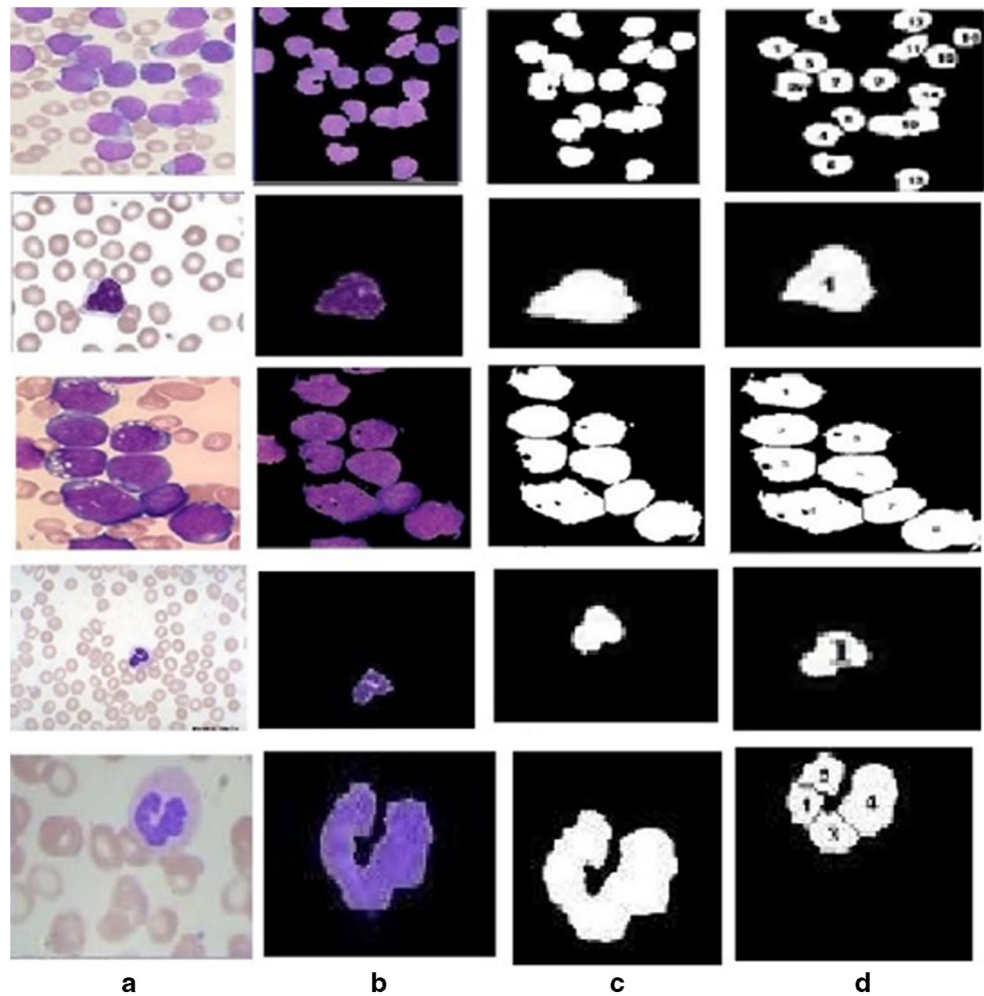
$$\sum_{i,j=0}^{N-1} P_{i,j} / (1 + (i-j)^2) \tag{29}$$

2.4.5 Fractal dimension

Hausdorff dimension: Fractal dimension is a statistical quantity that indicates how completely a fractal appears to fill space. Box-counting algorithm is used here. It partitions the structure space with square boxes of d-dimensional grid and size ‘r.’ The r value determines the number of nonempty boxes required to cover the fractal structure. It is shown in Eq. (30).

$$N((r) \sim r^{-D}) \tag{30}$$

Fig. 11 **a** Input image. **b** The nucleus of immature white blood cell. **c** Binary image of the nucleus before the separation of touching objects. **d** Labeled nuclei



In the box-counting algorithms, the number of boxes covering a point set is a power law function of the box size. Equation (31) shows the computation of box counting dimension.

$$-D = \lim_{r \rightarrow 0} \log N(r) / \log r \tag{31}$$

They indicate the roughness of the object. The Hausdorff dimension of a cancer cell will be above 1. Healthy cells will have a value below 1. The values are set after consulting the domain expert.

2.5 Data preprocessing

Data pre-processing is an essential step in the data mining process. Data gathering methods are loosely controlled, resulting in out-of-range values and impossible values. This stage plays a significant role in getting accurate results. Since the maximum, minimum, and mean intensity are obtained

from the segmented image of white blood cells some missing values appear which are replaced with value 126 for maximum intensity, a value of 25 for minimum intensity and value 72 for mean intensity. These values are the attribute means.

2.6 Feature selection

Feature selection methods can be used to achieve efficient data reduction. It helps in reducing the measurement cost, elimination of redundant features, complexity reduction, and a decrease in computational cost. It is useful for finding accurate data models. The subset of attributes is assessed using the InfoGainAttributeEval, and the Ranker Search method is used. The infoGainAttributeEval evaluates the worth of an attribute by measuring the information gain concerning the class. It is computed as given below:

$$\begin{aligned} \text{InfoGain}(\text{Class}, \text{Attribute}) \\ = H(\text{Class}) - H(\text{Class} | \text{Attribute}) \end{aligned} \tag{32}$$

The ranker search method computes the individual evaluations to rank the attributes. It is used along with attribute evaluators such as Chi-square, GainRatio, and InfoGain. The ranker search ranks the attributes using their individual evaluations. Two methods namely, Filter and Wrapper, are used. Filter method is not dependent upon the exact learning algorithm [50]. The wrapper method is expensive and it depends upon the learning algorithm [51]. It is used along with attribute evaluators. The attributes are discarded based on a threshold value. The method ranks the attributes with high or low value. The rank of the attributes generated by Ranker search method is given in Table 3. The ranks are placed in decreasing order. The attributes like Centroid_y, Eccentricity, Centroid_x, maximum intensity, orientation, form factor, roundness, and Id had zero information gain. These features were excluded during the classification and prediction as they were irrelevant.

2.7 Classification and prediction of unknown class label

In this research work, the classification algorithm classifies the instances of acute lymphoblastic leukemia into

the L1/L2/L3/normal cell. The work aims to compare the different classifiers namely: Naive Bayes, Decision Tree, K-nearest neighbor, and random forest. The algorithm with the highest accuracy is chosen for the prediction. The efficiency of the algorithms is computed and tabulated in the results section.

3 Results

In this study, 600 blood smear images were used. All images are in RGB color space and saved in the JPG format. For this experiment, images stained with Leishman stain [52] and Wright's stain [53] are used. The normal sized images have the dimension of 459×306 and resolution of 96 dpi. The magnified images have a size of 640×480 with a resolution of 96 dpi. Out of the 600 images used, 192 images belonged to ALL-L1 category. A total of 103 images belonged to normal category. A total of 228 images belonged to ALL-L2 type and 77 images belonged to ALL-L3 type.

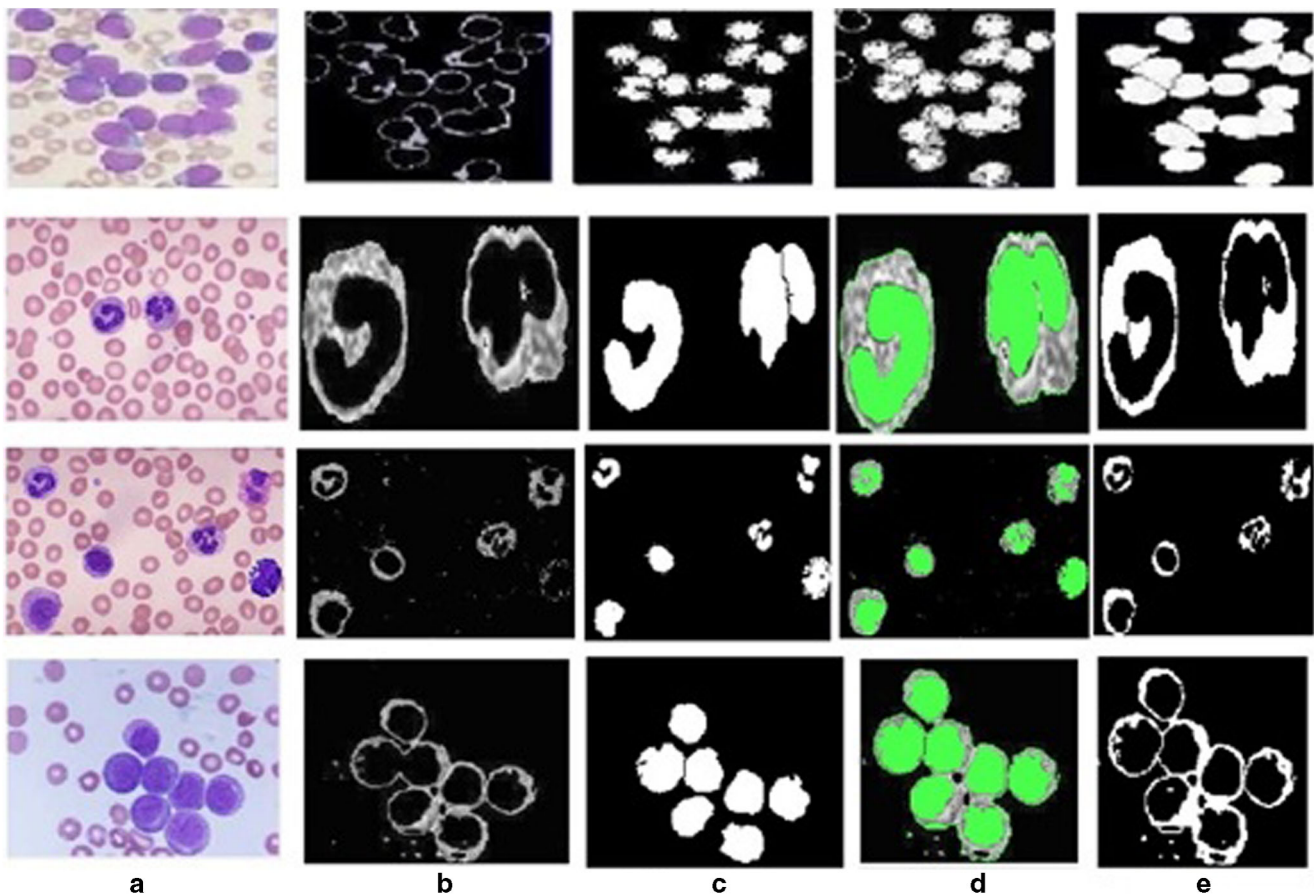
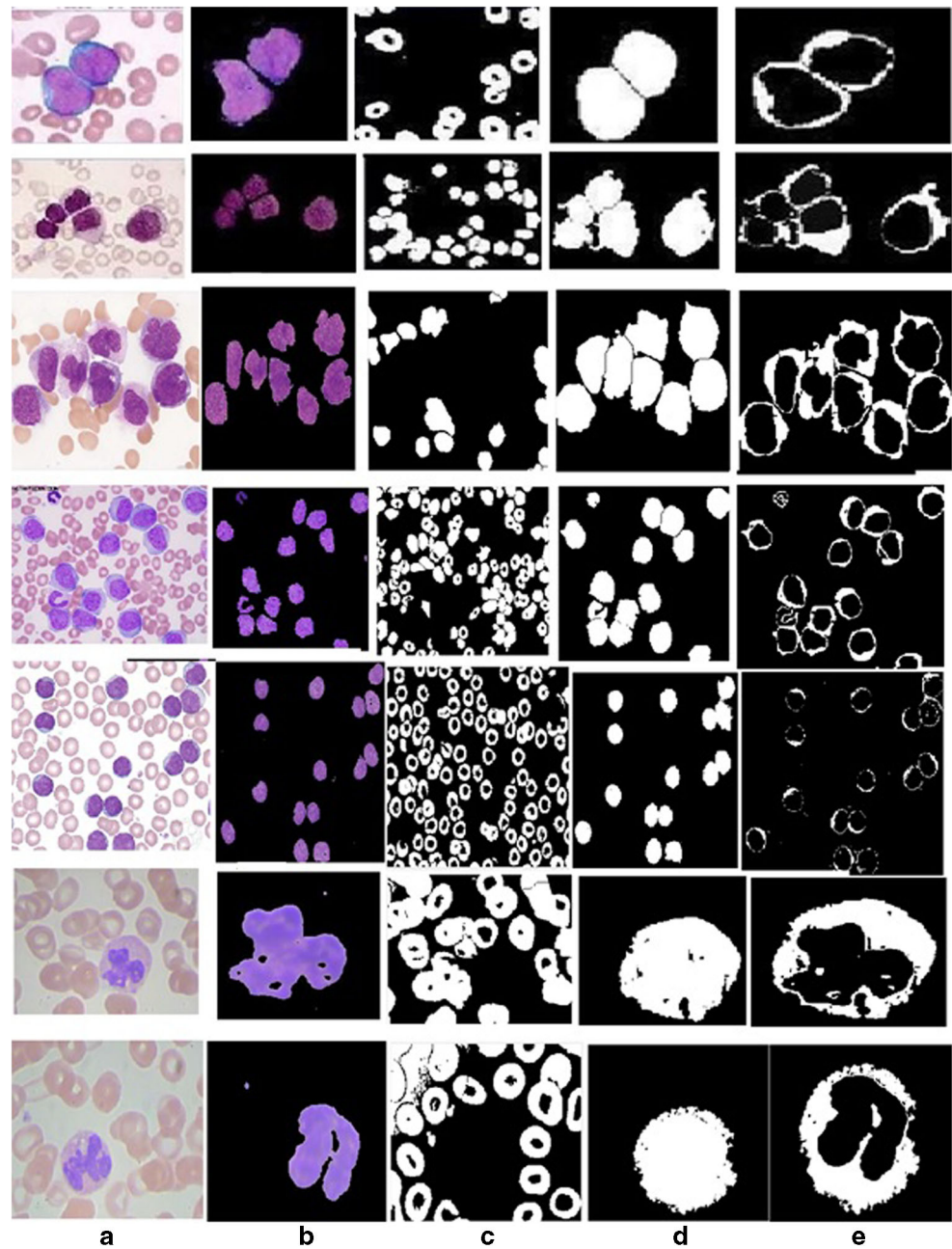


Fig. 12 **a** Input image. **b** Image of the cytoplasm obtained from the K-medoids algorithm. **c** Nucleus mask. **d** Nucleus mask imposed on the image of the cytoplasm. **e** Cytoplasm after accurate splitting

Fig. 13 Cytoplasm extraction after separation from surrounding RBCs. **a** Input image. **b** Nucleus of the WBC. **c** The binary image of red blood cells. **d** The binary image of WBC. **e** Binary image of the cytoplasm



3.1 Result of segmentation of WBC nucleus from blood smear

Nuclei have variable shapes in numerous types of leukocytes. Finding an essential technique to form modeling and segmenting the nucleus has continually been a challenge for scientists. Lymphoblast count raises to a higher level in case of acute lymphoblastic leukemia. The lymphoblast's also consumed the nutrients and space by squeezing out the WBC from the bone marrow. Nucleus features play a vital role in differentiating a lymphoblast cell from a lymphocyte. This work intends to segment WBC nucleus from blood smear

stained with various stains. The result of segmentation on some images in the dataset is shown in this section. In ALL-L1 type, the blasts have minimal cytoplasm, variable nuclear size, and chromatin density. The nuclear contour is irregular. In ALL-L2 type, the cell has irregular membrane and cytoplasm is scanty. The cells, in this case, are giant and have large nucleolus. It is found mostly among the adults. In ALL-L3 type, the cells are large and varied with vacuoles. In case of a normal peripheral blood smear, the cell has irregular membrane and cytoplasm is abundant. Normal lymphocytes seen on a blood smear are small and are relatively homogeneous. Their nucleus appears dense or coarse and clumped with

ridges of chromatin and parachromatin. Figure 11 depicts the result of nucleus extraction. The effect of segmentation on some images of ALL-L1, ALL-L2, ALL-L3 type, and normal blood smear present in training set is shown here.

3.2 Result of segmentation of WBC cytoplasm from blood smear

In acute lymphoblastic leukemia, the cytoplasm occupies about 20–30% of the cell. The nucleus holds around 80–90% of the region. In ALL-L1 type, the blasts have minimal cytoplasm. In ALL-L2 type, the cell has scanty cytoplasm. In ALL-L3 type, the cells contain a moderate volume of cytoplasm, and prominent vacuoles are present. In case of a normal peripheral blood smear, the cell has abundant cytoplasm. The feature of cytoplasm plays a vital role in the diagnosis of malignancy. Figure 12 depicts the result of segmentation using the first approach of cytoplasm extraction as explained in Sect. 2.3. The result of segmentation on some images in the training set is shown here.

Figure 13 depicts the cytoplasm extracted after separating the surrounding RBCs. The separation is done by taking into consideration the size of RBCs. The process of segmentation

of the cytoplasm is explained in Sect. 2.3. The result of segmentation on some sample images from the training set is shown here.

3.3 Observations

In this section, the feature values of lymphoblast and lymphocyte are recorded. Recorded values of two images from the training sample are shown. The feature values extracted are tabulated in Table 4. The observations reveal that the blasts are more eccentric than the normal cells. The energy values indicate that healthy cells are more uniform than blasts. The higher value of elongation for the blasts shows its tendency to lengthen. The value of the N/C ratio provides a clear differentiation between the lymphoblast and the lymphocyte.

3.4 Performance evaluation of the classifier

In the proposed work, an image-level analysis is used as a benchmark [54]. The confusion matrix, sensitivity, specificity, positive predictive value, and negative predictive value are computed for every class (ALL-L1, ALL-L2, ALL3, and normal) to measure the performance of the classifier. The sensitivity is

Table 4 Extracted feature values of lymphoblast and lymphocyte


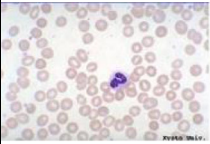
Feature	Lymphoblast		Lymphocyte
			
Eccentricity	0.6918	0.4676	0.701706
Centroid	(84,124)	(126,85)	(420.92,274.78)
Orientation	-58.60	25.99	37.18
Major Axis Length	79.864	65.25	66.87
Area	3332	2871	2324
Solidity	0.8954	0.9397	0.9039
Equivalent Diameter	61.13	60.46	54.396
Perimeter	258.07	210.56	195.657
Minor Axis Length	57.668	57.681	47.641
Maximum Intensity	174	168	234
Minimum Intensity	11	13	36
Form Factor	0.6287	0.8138	0.7629
Roundness	19.98	15.44	16.4723
Elongation	2.79	2.26	1.366
Entropy	1.9797	1.9797	0.166
Mean Intensity	133.67	119.38	105.69
Contrast	0.1080	0.1080	0.014
Correlation	0.9612	0.9612	0.886
Energy	0.0765	0.0765	0.9867
Homogeneity	0.983	0.983	0.9984
Radius	32.56	30.23	27.198
Cytoplasm area	4510	2871	4851
N/C ratio	0.7388	0.983	0.4790
Hausdorff Dimension	1.1799	1.1799	0.8324
Size of Blast	Large	Large	Small
Shape	Irregular	Irregular	Irregular
Vacuoles	No	No	No

Table 5 Performance evaluation of the classifiers

Algorithm	Random forest				Decision tree				Naïve Bayes				KNN			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
Confusion matrix																
A	185	1	0	6	182	4	0	6	170	10	6	6	176	8	0	8
B	05	98	0	0	7	96	0	0	25	78	0	0	13	90	0	0
C	0	0	228	0	0	0	228	0	8	5	215	0	0	0	223	5
D	0	0	5	72	0	0	7	70	2	6	10	59	0	0	6	71
True Positive _A	185				182				170				176			
True Negative _A	403				401				373				395			
False Positive _A	05				07				35				13			
False Negative _A	07				10				22				16			
Sensitivity _A	96.3%				94.7%				88.5%				91.6%			
Specificity _A	98.7%				98.28				91.4%				96.8%			
Positive Predictive Value _A	97.3%				96.26%				84%				93.1%			
Negative Predictive Value _A	99.01%				97.56%				94.43%				96.1%			
Accuracy _A	98%				97.1%				85.5%				95.1%			
True Positive _B	98				96				78				90			
True Negative _B	496				493				476				489			
False Positive _B	01				04				21				08			
False Negative _B	05				07				25				13			
Sensitivity _B	95.14%				93.2%				75.7%				87.3%			
Specificity _B	99.7%				99.1%				95.7%				98.3%			
Positive Predictive Value _B	98.9%				96%				81.2%				91.8%			
Negative Predictive Value _B	99%				98.6%				95%				97.4%			
Accuracy _B	99%				98.1%				95.6%				96.5%			
True Positive _C	228				228				215				223			
True Negative _C	367				365				356				366			
False Positive _C	05				07				16				06			
False Negative _C	0				0				13				05			
Sensitivity _C	100%				100%				94.2%				97.8%			
Specificity _C	98.6%				98.1%				95.6%				98.3%			
Positive Predictive Value _C	97.8%				97%				93%				97.3%			
Negative Predictive Value _C	100%				100%				96.4%				98.6%			
Accuracy _C	99.16%				98.3%				95.1%				98.16%			
True Positive _D	72				70				59				71			
True Negative _D	517				517				517				510			
False Positive _D	06				06				6				13			
False Negative _D	05				07				18				06			
Sensitivity _D	93.5%				90%				90%				92.2%			
Specificity _D	98.85%				98.8%				98.8%				98.83%			
Positive Predictive Value _D	92.3%				92.1%				90.07%				84.5%			
Negative Predictive Value _D	99%				98.6%				96.6%				98.8%			
Accuracy _D	98.16%				97.83%				96%				96.8%			

calculated using the Eq. 33. The specificity is computed using Eq. 34. The positive predictive value is calculated using Eq. 35. The negative predictive value is computed using the Eq. 36. The accuracy of the model is computed using Eq. 37.

$$Sensitivity = TP / (TP + FN) \tag{33}$$

$$Specificity = \frac{TN}{TN + FP} \tag{34}$$

$$Positive\ Predictive\ Value = TP / (TP + FP) \tag{35}$$

$$Negative\ Predictive\ Value = \frac{TN}{TN + FN} \tag{36}$$

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{37}$$

In the confusion matrix, ‘A’ represents the ALL-L1 type. ‘B’ indicates Normal. ‘C’ indicates ALL-L2 type. ‘D’ indicates ALL-L3 type. Where TP represents the true positive, TN represents the true negative, FP represents the false positive, and FN represents the false negative. True positives (TP) indicate the number of images that are diagnosed as an ALL-type by the proposed system and the domain expert also diagnoses it as an ALL-type. False positives (FP) indicate the number of images that are diagnosed as an ALL-type by the proposed system but not by the domain expert. False negative

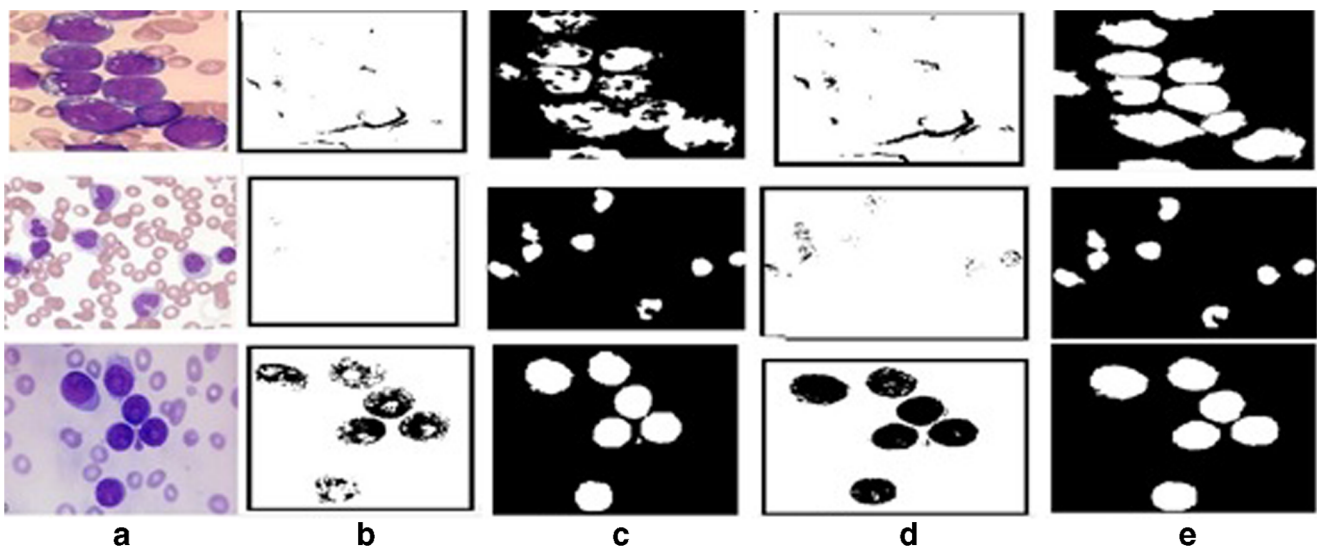


Fig. 16 **a** Blood smear image. **b** WBC nucleus obtained using the technique given by Alomari et al. [3]. **c** WBC nucleus obtained using the technique given by Joshi et al. [8]. **d** WBC nucleus obtained using the

technique given by Nasir et al. [12]. **e** WBC nucleus obtained using the proposed method. **f** WBC nucleus obtained using the technique given by Madhloom [24]

(FN) shows the number of ALL-type images detected by the domain expert but not by the proposed method. The type may be L1, L2, L3, or normal. The values of the metrics are presented in Table 5.

The result shows that the Random Forest algorithm outperforms the other supervised algorithms. Hence, the proposed work uses it for classification and prediction. Random Forest is an ensemble learning method for classification and regression. It operates constructing a multitude of decision trees at training time and predicts the unknown class labels at the prediction stage. When a new input vector needs to be

anticipated, it puts the vector down the tree, and each tree provides votes. The forest chooses the classification with maximum votes. Here, the authors have chosen 100 iterations as it is the default and the tree depth is kept unlimited.

3.4.1 Prediction made by the proposed system

To evaluate the performance of the proposed system, an instance of the ALL-L3 case is fed into the system. The image is segmented, and extracted features are stored in ARFF (Attribute-Relation File Format) [55]. WEKA

```
@relation finaltestfile

@attribute Major_Axis_length numeric
@attribute Area numeric
@attribute Solidity numeric
@attribute Equi_diameter numeric
@attribute Perimeter numeric
@attribute Minor_Axis_length numeric
@attribute Minimum_Intensity numeric
@attribute Cytoplasm_Area numeric
@attribute N/C_ratio numeric
@attribute Roundness numeric
@attribute Elongation numeric
@attribute Radius numeric
@attribute 'Mean Intensity' numeric
@attribute Size_of_Blast {L,S}
@attribute Membrane {R,I}
@attribute Vacuoles {N,Y}
@attribute Class {L1,L2,L3}
@data
32.671954,622,0.898844,28.141695,99.06,25.799001,183,33,979,0.635342,1,14.078047,76.483553,L,I,Y ?
33.190714,682,0.953846,29.467768,95.595,26.783043,181,45,1210,0.563636,1.066154,14.733884,80.738552,L,I,Y ?
28.636405,578,0.974705,27.128075,84.67,25.981609,134,45,732,0.789617,0.986882,13.564038,73.904425,L,I,Y ?
27.110648,497,0.966926,25.155517,78.021,23.607906,163,40,740,0.671622,0.974542,12.577759,69.96945,L,I,Y ?
34.592663,731,0.945666,30.508001,99.126,27.591039,204,9,1034,0.706963,1.069527,15.254,78.476574,L,I,Y ?
34.74651,648,0.937771,28.723844,95.562,24.478956,50,0,1030,0.629126,1.121319,14.361922,25,L,I,Y ?
```

Unknown class labels

Fig. 14 Test data with unknown class label

```

@relation finaltestfile_predicted

@attribute Major_Axis_Length numeric
@attribute Area numeric
@attribute Solidity numeric
@attribute Equi_diameter numeric
@attribute Perimeter numeric
@attribute Minor_Axis_length numeric
@attribute Minimum_Intensity numeric
@attribute Cytoplasm_Area numeric
@attribute N/C_ratio numeric
@attribute Roundness numeric
@attribute Elongation numeric
@attribute Radius numeric
@attribute 'Mean Intensity' numeric
@attribute Size_of_Blast {L,S}
@attribute Membrane {R,I}
@attribute Vacuoles {N,Y}
@attribute 'prediction margin' numeric
@attribute 'predicted Class' {L1,N,L2,L3}
@attribute Class {L1,N,L2,L3}

@data
32.671954,622,0.898844,28.141695,99.06,25.799801,183,23,979,0.63
5342,1,14.070847,76.483553,L,I,Y,-0.34,L3,?
33.190714,682,0.953846,29.467768,95.595,26.783043,781,45,1210,0.
563636,1.066154,14.733884,80.738552,L,I,Y,-0.34,L3,?
28.636405,578,0.974705,27.128075,84.67,25.981609,134,45,732,0.78
9617,0.986882,13.564038,73.904425,L,I,Y,-0.32,L3,?
27.110648,497,0.966926,25.155517,78.021,23.607906,163,40,740,0.6
71622,0.974542,12.577759,69.96945,L,I,Y,-0.3,L3,?
34.592663,731,0.945666,30.508001,99.126,27.591039,204,9,1034,0.7
06963,1.069527,15.254,78.476574,L,I,Y,-0.34,L3,?
34.74651,648,0.937771,28.723844,95.562,24.478956,50,0,1030,0.629
126,1.121319,14.361922,25,L,I,Y,-0.34,L3,?
    
```

Fig. 15 Predicted class labels

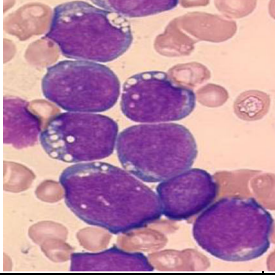
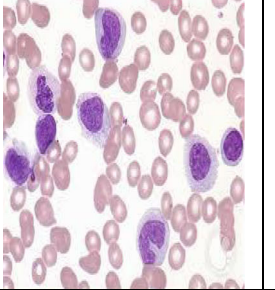
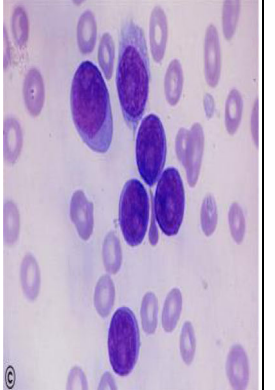
(Waikato Environment for Knowledge Analysis) is used as the platform to implement random forest algorithm. The tool accepts the input in ARFF format. The test file is created by extracting features from a blood smear image diagnosed as ALL-L3 type. The image is fetched from the test dataset. Figure 16 describes the test file. The class label is unknown here. In the file, ALL-L1 is labeled as L1, ALL-L2 as L2, ALL-L3 as L3, and normal as N. Figures 14 depicts the ARFF file with unknown class labels and Figure15 depicts the class label predicted by the system after applying the knowledge gained from training phase.

4 Discussion

1. The proposed model performed well for images with different stains and clicked under different illumination conditions. K-medoids generated accurate results when compared to SVM and GMM. SVM was not able to extract nuclei due to its variation in colors [56]. GMM did not generate accurate results for

- images colored with different stains. It failed to obtain the cytoplasm in some cases which plays a vital role in diagnosis of the disease [].
2. The proposed model has achieved an accuracy of 98.6% (average of accuracies of every class).
 3. The method uses only 600 samples of blood smear images which are stained with different stains.
 4. The methods described in related works section achieved a standard accuracy. But the techniques involved the manual cropping of cells in obtaining the cytoplasm area. The proposed algorithm automated the process of cropping of cells. It proved to be a great help for the pathologists when multiple cells existed.
 5. The method introduced a new automated technique to separate the surrounding RBCs from the WBCs which gave precise results.
 6. The touching cells are isolated using a modified watershed transform. The cytoplasm was obtained by placing the nucleus mask over the cell to achieve the perfect split. This technique helped to get the cytoplasm region when the cells were firmly attached.

Table 6 Number of WBC nuclei obtained using the proposed technique, state of the art methods and ground truth

Image	Segmentation using the technique in [4]	Segmentation using the technique in [19]	Segmentation using the technique in [27]	Segmentation using the technique in [44]	Segmentation using the proposed method	Ground truth determined by the pathologist
	19	29	27	10	11	11
	0	09	12	12	8	07
	25	07	10	08	06	06

7. The model is validated using the test images. The results obtained are verified with the domain expert.
8. The work was implemented on a computer with an Intel Core 2.5GHz processor and 4GB RAM. The feature selection step helped to reduce the training time. The classification using random forest took 1.03 s to classify the samples. Thus, the proposed technique can be considered computationally less difficult and fast. Therefore, it may be regarded as to deploy in hospitals and clinics.
9. Afsar et al. [2] have already used ANN on a dataset obtained from the hospital for automated detection of acute leukemia. The authors intend to try ANN and different deep neural network architectures on a bigger dataset to diagnose acute leukemia as the future work. Also, the

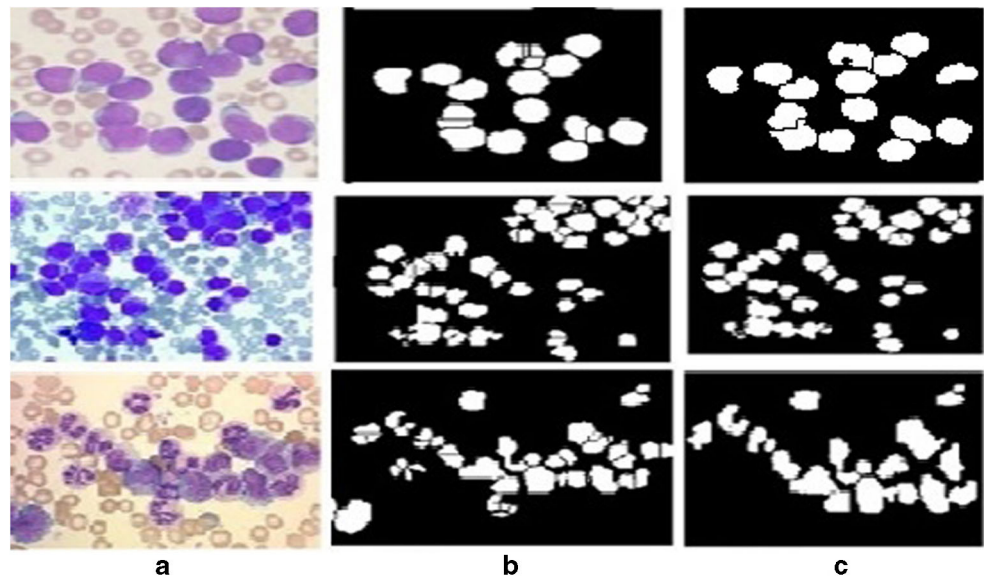
proposed technique can also be employed to detect the acute myeloid leukemia and its subtypes.

The effect of using the proposed algorithm is discussed below.

4.1 Effect of using K-medoids algorithm and modified watershed transform to segment WBC nucleus

In this section, the effect of using K-medoids algorithm to segment the WBC nucleus is discussed. The result obtained before and after the modified watershed transform is also shown. It is then compared with the results obtained using state of the art segmentation techniques. The results show that

Fig. 17 **a** Blood smear image. **b** Touching nuclei separation obtained using existing watershed transform. **c** Touching nuclei separation obtained using modified watershed transform



the use of K-medoids to segment WBC nucleus produces promising results when compared to the state of the art methods. The result of segmentation on sample images of

ALL-L3 type and ALL-L2 type are shown here. Figure 16 depicts the result of segmentation of WBC nucleus using the different methods and the proposed method. The result

Table 7 WBC count obtained using watershed transform, modified watershed transform, and ground truth

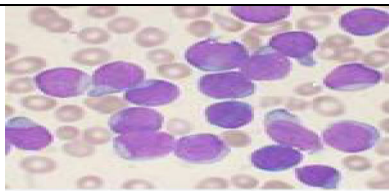
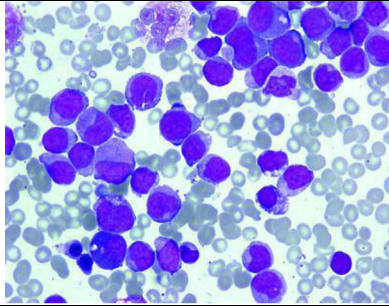
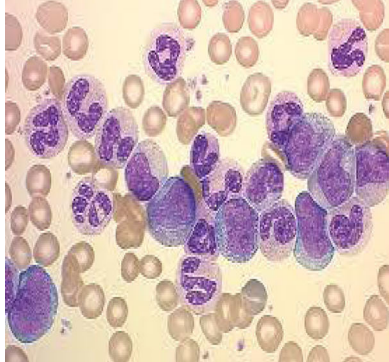
Image	No of WBC nuclei using watershed transform	No of WBC nuclei using modified watershed transform	Ground Truth
	22	15	19
	78	37	40
	56	23	22

Fig. 18 **a** Blood smear image. **b** Cytoplasm obtained as a result of using K-means algorithm for segmentation. **c** Nuclei mask. **d**. Cytoplasm obtained as a result of using K-medoids algorithm for segmentation and nucleus mask imposition

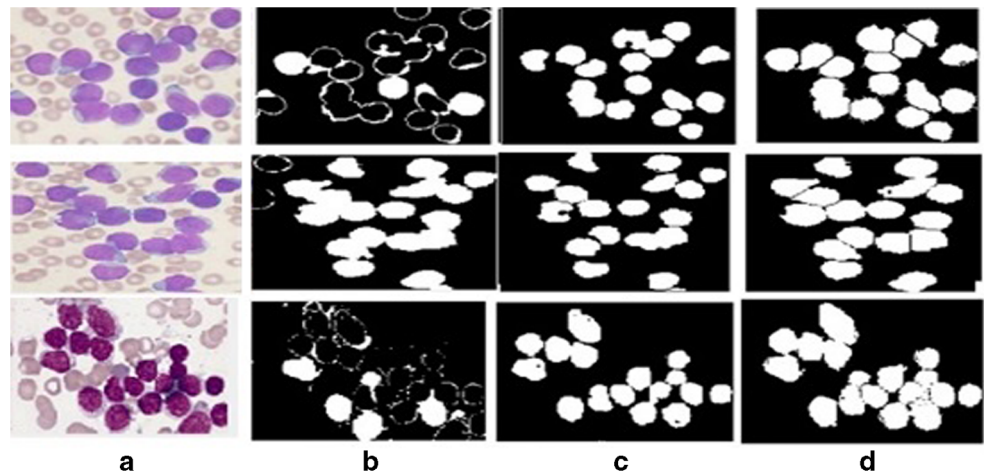
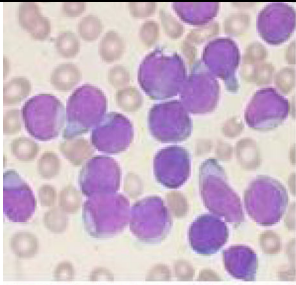
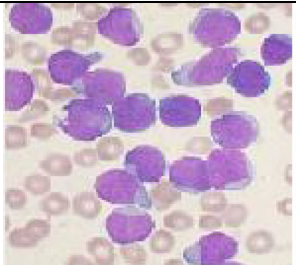
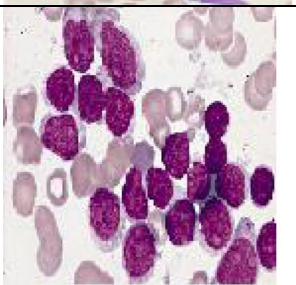


Table 8 Count of WBC nucleus and WBC cytoplasm

Image	No. of WBC nuclei	No. of Cytoplasm region obtained using K-Means algorithm	No. of cytoplasm region obtained using K-medoids algorithm and imposition of nucleus mask	Ground Truth
	15	13	15	WBC Nuclei: 16 WBC cytoplasm :16
	16	9	16	WBC Nuclei:16 WBC Cytoplasm: 16
	19	10	19	WBC nuclei:19 WBC cytoplasm:19

obtained after the last stage of every segmentation technique is depicted. The number of WBC nuclei obtained using each of the technique, and the proposed method is presented in Table 6. The ground truth is given by the pathologist.

The problem of over-segmentation faced by the watershed transform is overcome in this work. The modified watershed transform is explained in Section 2.2 under image post-processing stage. The result obtained after the final segmentation stage is depicted here. Fig. 17. Depicts the touching nuclei separation achieved using the existing watershed transform and the proposed method. The results shown proves that the proposed method produces promising results when compared to the current technique. Table 7 represents the count of WBC nuclei obtained using watershed transform, modified watershed transform, and ground truth. One downside of the approach used here is the loss of cells present at the border. Due to inaccurate extraction of cells present at the edge, the count of WBC nuclei obtained using the proposed method varies from the ground truth.

4.2 Effect of imposing nucleus mask on the cytoplasm

In this section, a comparison is made between the result obtained after applying K-means algorithm to the input image and the proposed method. Figure 18 shows the cytoplasm extracted using the above two techniques. The count of WBC nucleus and WBC cytoplasm obtained using the methods mentioned above is tabulated in Table 8. The result of segmentation on three sample images is shown. All instances belong to the ALL-L2 type. The region count is computed using the “regionprops” function.

5 Conclusion

A novel approach for leukemia detection is proposed. In comparison with the other algorithms, the innovative approach extracts the cytoplasm from the images without involving manual cropping procedure. Cytoplasm extraction plays a vital role in segmentation since the ratio between the nucleus area, and cytoplasm area is a pathologically significant feature for cancer detection. Texture, shape, and visual features are extracted. The features are ranked using the Ranker Search method. Different algorithms are compared, and the one with the highest accuracy is chosen for classification and prediction. The proposed algorithm takes in to account the images stained with different stains and cells with various abnormalities for efficient detection of the malignancy. The segmentation approach is fully automatic without the need to crop the image explicitly. The work achieved an accuracy of 98.6%. The pitfall of the algorithm is it requires manual intervention

to diagnose the presence of vacuoles. It can be automated by incorporating the hole detection algorithm. The model can be made more robust by training it with a dataset having additional images clicked under different illumination conditions. Border cells that were located at extreme corners are not accurately dealt. Future work involves precise extraction of border cells and also the sub classification of acute lymphoblastic leukemia into its respective phenotypes namely pre-T type, pre-B type, and mature B.

References

1. Karen Seiter (2016) ALL types. In: Acute Lymphoblastic Leukemia Staging. Available via <https://emedicine.medscape.com/article/2006661-overview>. Accessed 16th Nov, 2017
2. Afshar S, Abdolrahmani F, Tanha FV, Seif MZ, Taheri K (2011) Recognition and prediction of leukemia with artificial neural network (ANN). *Med J Islam Repub Iran* 25(1):35–39
3. Alomari YM, Azma RZ, Abdullah SNHS, Omar K (2014) Automatic detection and quantification of WBCs and RBCs using iterative structured circle detection algorithm. *Computational and Mathematical Methods in Medicine* 2014:1–14. <https://doi.org/10.1155/2014/979302>
4. Belekar SJ, Chougule SR (2015) WBC segmentation using morphological operation and SMMT operator—a review. *International Journal of Innovative Research in Computer and Communication Engineering* 3(1):434–440
5. Bhagvathi SL, Thomas NS (2016) An automatic system for detecting and counting RBC and WBC using fuzzy logic. *ARPN-JEAS* 11(11):6891–6894
6. Bhamare, Miss. Madhuri G. and D.S.Patil (2013) Automatic blood cell analysis by using digital image processing: a preliminary study. *Int J Eng Res Technol* 2(9):3137–3141.
7. Erosion operation (2017) <https://in.mathworks.com/help/images/ref/imerode.html>. Accessed 16th November, 2017
8. Joshi MD, Karode AH, Suralkar SR (2013) White blood cells segmentation and classification to detect acute leukemia. *International journal of emerging trends Technology in Computer. Science* 2(3): 147–151
9. Marzuki NIC, Mahmood NH, Razak MAA (2015) Segmentation of white blood cell nucleus using active contour. *J Teknol* 74(6):115–118. <https://doi.org/10.11113/jt.v74.4675>
10. Mishra, S. J., & Deshmukh, A.P. (2014) Detection of Leukemia in Human Blood Sample based on Microscopic Images. *International Journal of Advanced Research in Electronics and Communication Engineering*, 1(3) :10-14.
11. Mulik V, Bhilare PM, Alhat S (2016) Analysis of acute lymphoblastic leukemia cells using digital image processing. *International Journal for Scientific Research and Development* 4(2):70–72
12. Nasir A, Mustafa N, Nasir NFM (2009) Application of thresholding technique in determining ratio of blood cells for leukemia detection. In *Proceedings of the international conference on man-machine systems (ICoMMS 2009)*, 1–6
13. Putzu L, Ruberto CD (2013) White blood cells identification and counting from microscopic blood image. *International journal of medical, health, biomedical, bioengineering and Pharm Eng* 7(1): 20–27

14. Sadeghian F, Ramli AR, Seman Z, Khahar BHA, Saripan MI (2009) A Framework for White Blood Cell Segmentation in Microscopic Blood Images Using Digital Image Processing. *Biological Procedures Online* 11(1):196–206. <https://doi.org/10.1007/s12575-009-9011-2>
15. Thanh TTP, Pham GN, Park JH, Moon KS, Lee SH, Kwon KR (2017) Acute leukemia classification using convolution neural network in clinical decision support system. In *Proc Of Computer Science & Information Technology*, 49–53. <https://doi.org/10.5121/csit.2017.71305>
16. Vaghela HP, Pandya M, Modi H, Potdar MB (2015) Leukemia detection using digital image processing techniques. *International Journal of Applied Information Systems* 10(1):43–51
17. Wang M, Zhou X, Li F, Huckins J, King RW, Wong STC Novel cell segmentation and online learning algorithms for cell phase identification in automated time-lapse microscopy. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 65–68. <https://doi.org/10.1109/isbi.2007.356789>
18. Wang Q, Wang J, Zhou M, Li Q, Wang Y (2017) Spectral-spatial feature-based neural network method for acute lymphoblastic leukemia cell identification via microscopic hyperspectral imaging technology. *Biomedical Optics Express* 8(6):3019–3028
19. Saeedizadeh Z, Dehnavi A, A.M., Rabbani H (2015) Automatic recognition of myeloma cells in microscopic images using bottleneck algorithm, modified watershed and SVM classifier. *J Microsc* 261(1):46–56. <https://doi.org/10.1111/jmi.12314>
20. Jiang K, Jiang QX, Xiong Y (2003) A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering. *Mach Learning Cybnetics* 5:2820–2825
21. Kumar BR, Joseph DK, Sreenivas TV (2002) Teager energy based blood cell segmentation. In *proceedings of the 14th international conference on digital signal processing*, Santorini, Greece, 619–622
22. Alreza ZKK, Karimian A (2016) A. Design a new algorithm to count white blood cells for classification leukemic blood image using machine vision system. In *Proceedings of the 6th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Iran, 251–256
23. Reta C, Gonzalez JA, Diaz R, Guichard JS (2011) Leukocytes segmentation using Markov random fields. *Adv Exp Med Biol* 696:345–353
24. Madhloom HT, Kareem SA, Ariffin H (2011) An image processing application for the localization and segmentation of lymphoblast cell using peripheral blood images. *J Med Syst* 36:2149–2158
25. Houbay EMFE (2018) Framework of computer aided diagnosis Systems for Cancer Classification Based on medical images. *J Med Syst* 42:157–167
26. Theera-Umpon N (2005) Patch-based white blood cell nucleus segmentation using fuzzy clustering. *ECTI Transactions on Electrical Eng., Electronics, and Communications* 3(1):15–19
27. Mohammed EA, Mohamed MMA, Naugler Christopher, Far BH (2013) Chronic lymphocytic leukemia cell segmentation from microscopic blood images using watershed algorithm and optimal thresholding. In *Proc. of the 26th IEEE Can Con El Comp En(CCECE)*:1–5
28. Esti Suryani and Wiharto Wiharto and Nizomjon Polvonov (2015) Identification and Counting White Blood Cells and Red Blood Cells using Image Processing Case Study of Leukemia. *International Journal of Computer Science & Network Solutions* 2(6):35–49.
29. Khobragade S, Mor DD, Patil CY (2015) Detection of leukemia in microscopic white blood cell images. In *Proc. of the 2015 international conference on information processing (ICIP)* (ICIP): 435–440. <https://doi.org/10.1109/INFOP.2015.7489422>
30. Cuevas E, Diaz M, Manzanares M, Zaldivar D, Perez-Cisneros M (2013) An improved computer vision method for white blood cells detection. *Computational and Mathematical Methods in Medicine*: 1–14
31. Ahasan R, Ratul AU, Bakibillah ASM White Blood Cells Nucleus Segmentation from Microscopic Images of strained peripheral blood film during Leukemia and Normal Condition. In *Proc. of the 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*: 361–366. <https://doi.org/10.1109/ICIEV.2016.7760026>
32. Guo N, Zeng L, Wu Q (2007) A method based on multispectral imaging technique for white blood cell segmentation. *Comput Biol Med* 37:70–76
33. Sinha N, Ramakrishnan AG (2002) Blood cell segmentation using EM algorithm. In *proc. Third Indian conference on computer vision. Graphics & Image Processing (ICVGIP)*:1–6. <https://doi.org/10.1109/ICIEV.2016.7760026>
34. Wang M, Chu R (2009) A novel white blood cell detection method based on boundary support vectors. In *proc. of the 2009 IEEE international conference on systems. Man and Cybernetics*:2595–2598
35. Nivaldo Medeiros(n.d.) Blood Smear Database. In: *Atlas*. Available via <http://www.hematologyatlas.com/principalpage.htm>. Accessed 16th November, 2017. <http://www.hematologyatlas.com/principalpage.htm>. Accessed 16 Nov 2017
36. Labati RD, Piuri V, Scotti F (2011) ALL -IDB: the acute lymphoblastic leukemia image database for image processing. In *Proc. of the 2011 IEEE international conference on image processing* 2045–2048. <https://doi.org/10.1109/ICIP.2011.6115881>
37. C.I.E primaries (n.d.) <http://hyperphysics.phy-astr.gsu.edu/hbase/vision/cieprim.html>. Accessed 16th November, 2017
38. Area open function (2015) <https://in.mathworks.com/help/images/ref/bwareaopen.html>. Accessed 16th November, 2017
39. Border clear (2015) <https://in.mathworks.com/help/images/ref/imclearborder.html>. Accessed 16th November, 2017
40. Watershed Transform (2013) <https://blogs.mathworks.com/steve/2013/11/19/watershed-transform-question-from-tech-support/>. Accessed: 16th November,2017
41. Gonzalez, R. C., R. E. Woods, and S. L. Eddins (2004) *Digital Image Processing Using MATLAB*. New Jersey, Pearson Prentice Hall
42. Measurement properties (2017) <https://in.mathworks.com/help/images/ref/regionprops.html>. Accessed: 16th November, 2017
43. Label the components (2015) <https://in.mathworks.com/help/images/ref/bwlabel.html>. Accessed: 16th November, 2017
44. Filling holes (2015) <https://in.mathworks.com/help/images/ref/imfill.html>. Accessed:16th November, 2017
45. Extended maxima (2012) <https://in.mathworks.com/help/images/ref/imextendedmax.html>. Accessed 16th November, 2017
46. Impose minima (2003) <https://in.mathworks.com/help/images/ref/imimposemin.html>. Accessed: 16th November, 2017
47. Subrajeet Mohapatra (2013) Hematological image analysis for acute lymphoblastic leukemia detection and classification. Doctoral Dissertation, National Institute of Technology Rourkela
48. Histogram equalization (2015). Histeq function. In: *Histogram Equalization*. Available via <https://in.mathworks.com/help/images/histogram-equalization.html>. Accessed: 16th Nov, 2017 <https://in.mathworks.com/help/images/histogram-equalization.html>. Accessed: 16 Nov 2017

49. GLCM (2006) http://support.echoview.com/WebHelp/Windows_and_Dialog_Boxes/Dialog_Boxes/Variable_properties_dialog_box/Operator_pages/GLCM_Texture_Features.htm#About_the_GLCM_and_textures. Accessed 16th November, 2017
50. Almuallim H, Dietterich TG (1991) Learning with many irrelevant features. In: Proc. Ninth National conference on Artificial intelligence (AAAI-91), Anaheim, CA 2: 547–552
51. Kira K, Rendell LA. (1992) The feature selection problem: traditional methods and a new algorithm. In: Proc. AAAI-92, San Jose, CA 122–126
52. MyHematology(2017) Leishman Stain . In : MyHematology. Available via <https://myhematology.com/red-blood-cells/leishman-stain/> . Accessed: 16th Nov, 2017
53. Wright Stain Method Technical Data Sheet (2005) Wright Stain . In: Technical Data Sheets. Available via. <https://www.emsdiasum.com/microscopy/technical/datasheet/26060.aspx>. Accessed 16 Nov 2017
54. ALL-IDB (2011) <https://homes.di.unimi.it/scotti/all/results.php>. Accessed 16th January,2019
55. Attribute-Relation File Format (ARFF) (2008) ARFF. In: Attribute-Relation File Format (ARFF). Available via. <https://www.cs.waikato.ac.nz/ml/weka/arff.html>. Accessed 16 Nov 2017
56. Mohammed EA, Mohamed MM, Far BH, Naugler C (2014) Peripheral blood smear image analysis: a comprehensive review. Journal of Pathology Informatics 5(1):9. <https://doi.org/10.4103/2153-3539.129442>
57. Adjust operation (2017) <https://in.mathworks.com/help/images/ref/imadjust.html>. Accessed:16th November, 2017

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Vasundhara Acharya is an Assistant Professor in Department of Computer Science and Engineering at Manipal Institute of technology, Manipal Academy of Higher Education, Manipal. Her areas of interest are Medical Image Processing, Artificial Intelligence, Machine Learning and Information Security. She received her Master's Degree in Software Engineering from MIT, Manipal. She received her Bachelor of Technology degree in Information Science from NMAMIT, Nitte and plans on pursuing a Doctorate in the field of Bioinformatics.

Preetham Kumar is a Deputy Registrar-Academics(Technical) at MAHE and Professor in the Department of Information & Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal. His research interests include Data Mining, Advanced Database Management Systems, Operating Systems, Software Architecture, Software Engineering. He received his PhD in Data Mining from National Institute of Technology, Karnataka.