Detection and Classification of Acute Lymphocytic Leukemia

Pradeep Kumar Das
Dept. of Electronics and Communication Engineering
National Institute of Technology Rourkela
Rourkela, 769008, India
Email: pdas391@gmail.com

Priyanka Jadoun
Dept. of Electronics and Communication Engineering
National Institute of Technology Rourkela
Rourkela, 769008, India
Email: priyankajadoun01@gmail.com

Sukadev Meher
Dept. of Electronics and Communication Engineering
National Institute of Technology Rourkela
Rourkela, 769008, India
Email: smeher@nitrkl.ac.in

***Abstract*—The research work aims to develop an automated detection and classification method for acute lymphocytic leukemia (ALL). Extraction of lymphocytes is accomplished by the color based k-means clustering technique. Then, shape, texture, and color features are extracted from the segmented image. Gray-level co-occurrence matrix (GLCM) and gray-level run-length matrix (GLRLM) algorithms are used to extract the features of nucleus. Moreover, Principal component analysis (PCA) is applied for dimensional reduction. Finally, an SVM (support vector machine) with an RBF kernel is employed to classify WBCs. The proposed method yields promising results with 96.00% accuracy and 92.64% sensitivity.**

Keywords: Classification; Contrast Limited Adaptive Histogram Equalization; K-means clustering; Leukemia; Support Vector Machine.

I. Introduction

Leukemia is a malignant blood growth that affects leukocytes. As indicated by the French-American-British (FAB) classification model [1], acute leukemia can be either acute lymphocytic leukemia (ALL) or acute myeloid leukemia (AML). Here, we discuss ALL, which affects a gathering of leukocytes called lymphocytes. It is a rapidly-progressive blood-cancer. It destroys the immune system. Hence, it leads to death if untreated in right-time. Microscopic study of a blood film is a crucial phase in leukemia detection. The microscopic examination is time taking, intuitively subjective, and controlled by clinical insight and experienced hematopathologists. To avoid above-explained problems, there is a requirement to develop an efficient methodology for the susceptible analysis of peripheral blood samples. Segmentation of image is a crucial issue in the mechanized hematological examination and should be precisely completed [2]–[4].

Liao and Deng [5] have presented a segmentation approach depends on straight forward thresholding pursued by shape identification. They assume that the cells are round fit as a fiddle. Angulo et al. [6] have presented a two-phase segmentation approach depending on the thresholding and paired-sifting. This method shows better segmentation execution regarding cytoplasm and nucleus. However, two-phase segmentation makes it slower. Madhukar et al. [7] have employed k-means clustering to detect nuclei, whereas they have extracted shape features and GLCM based texture features. Finally, they have applied SVM to classify lymphocytes [7]. Mohapatra et al. [20] have employed ensemble-classifier to successfully classify healthy and unhealthy lymphocytes. Putzu et al. [21] have presented a SVM based leukocyte classification approach to classify healthy and unhealthy leukocytes. Moreover, they have emphasized to extract crucial shape, color, and GLCM based texture features.

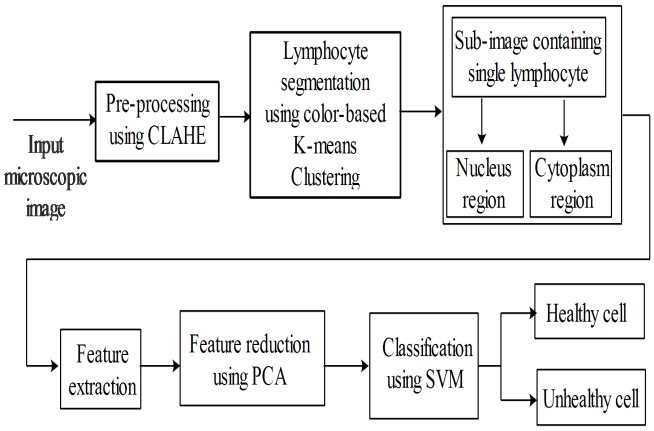

Fig. 1. Schematic of proposed work

Mishra et al. [8] have suggested a new approach to effectively classify ALL. For effective segmentation, they have used marker-based-watershed-segmentation, whereas they have applied GLCM based feature-selection technique. Finally healthy- and unhealthy- WBCs are classified using random forest [8]. Mishra et al. [3] have suggested a GLRLM based feature extraction approach to extract efficient texture features. Then they have employed SVM to successfully classify lymphocytes. Mishra et al. [22] have presented an efficient Adaboost algorithm with RF classifier to effectively classify healthy and unhealthy lymphocytes. They have extracted efficient texture features by employing discrete orthogonal S-transform. On the other hand, Vogado et al. [23] have employed convolutional neural network (CNN) to efficiently extract crucial features. Then they apply SVM to classify healthy and unhealthy leukocytes.

II. Proposed Method

Fig. 1 displays the schematic of proposed-method. The detailed explanation is given in subsequent sections.

A. Data Acquisition

The first phase of any system is images acquisition. The main motive is to obtain a set of image data with clarity and high-resolution [9]. In this research, we employ ALL-IDB1 dataset [10], to validate the proposed method. It contains 108 images, among which 49 are ALL affected [10]. Some ALL-IDB1 dataset images are displayed in figure2.

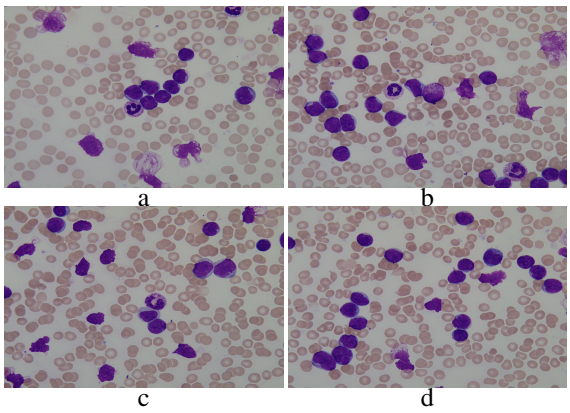

Fig. 2. Sample images from Dataset

B. Image pre-proessing

Raw image data captured directly from a camera can have various problems. If the image pre-processing can be used wisely, then it can solve many problems which lead to ultimate feature detection, because the final aim of any research is to detect the features of the object properly [11]. Here, Contrast limited Adaptive Histogram Equalization(CLAHE) technique is applied to enhance the image quality. It is explained in the followed section.

1) ***CLAHE****:* CLAHE is a modification of AHE. AHE suffers from over-amplification of the noise, particularly in homogeneous-region. CLAHE can solve this problem by confining the amplification. This is done by clipping the histogram based on clip-limit before calculating cumulative-distribution-function (CDF) [12], [24], [25]. In CLAHE, the transformation function is directly proportional to the CDF of pixels in neighborhood, as like AHE [12], [24], [25]. In CLAHE, amplification is directly proportional to the slope of the transformation function, which is directly proportional to the slope of CDF. Hence, it is directly proportional to PDF and histogram also. So, amplification is directly proportional to the histogram.

CLAHE works on small-regions *(tiles)*. The contrast of every *tile* is improved to achieve the desired output histogram [12], [24], [25].

2) Algorithm of CLAHE: Input: A RGB Color Image Output: CLAHE Enhanced Image

a. Convert RGB to L∗a∗b color space.

b. Divide the L component of image into several 8x8 blocks.

c. Compute the histogram of each block.

d. Fix the clip limit for the histogram.

e. Modify each histogram by specifying the transformation function.

f. Every histogram is converted so as not to exceed the specified section(clip) limit. The mathematical expression for the CLAHE with uniform distribution can be explained as follows.

$$s = [s_{\max} - s_{\min}] \times P(f) + s_{\min} \tag{1}$$

Where, P(f) is CPD(cumulative probability distribution)
$s_{\max}$ is maximum intensity
$s_{\min}$ is minimum intensity
s is computed intensity

Exponential-distribution for gray-level is mathematically written as:

$$s = s_{\min} - \left(\frac{1}{c}\right) \times \ln(1 - p(f)) \tag{2}$$

where c symbolizes clip parameter.

g. The neighboring-tiles are combined by bilinear interpolation. The image gray-values are adjusted depending on the modified-histograms.

C. Lymphocyte Segmentation

Lymphocyte cell segmentation is a vital step in automatic leukemia detection. This step requires the separation of blast cells from background components and further fragmentation in the nucleus and cytoplasm [13]. The nucleus of lymphocyte gives information about the chromatin and chromatin density profile. Hence, we can not ignore any part of the lymphocyte cell because it will lead to a loss of information and false diagnosis, which is not a good sign. Extraction of the lymphocyte cell from the peripheral blood smear is challenging because of the following reasons:

1. A picture of blood contains a combination of many blood components.
2. Adjacent cell problem: Red blood cell adjacent to lymphocyte cell and two lymphocytes cell.
3. Variation in the color around the circumference of the lymphocyte cell and nearby objects.

To make segmentation more precise, we use following steps.

1. Localization of lymphocyte
2. Transformation of color
3. Segregation of grouped lymphocyte
4. Separation of nucleus and cytoplasm

1) ***Algorithm of color-based k-means segmentation****:*

a. Read an image from the dataset.

b. Convert it from RGB to LAB.

c. Apply k-means-clustering to classify colors in a a*b* space.

d. Clustering the image into three parts according to the color value.

e. Since the cell nucleus is darker than the cytoplasm, it is separated from the cytoplasm by applying k-means clustering on L layer.

f. Apply some morphological operation to obtain a clean result image.

D. Feature Extraction

The proposed scheme emphasizes the extraction of significant features: shape-, color-, and texture- features, which have a key role in classification stage. Here, we extract GLCM and GLRLM based texture features [14].

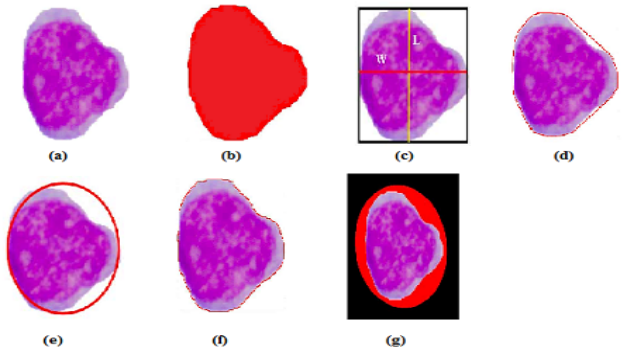

Fig. 3. Shape features: (a) Original lymphocyte cell (b) Area (c) Rectangular bounding box (d) Convex Hull (e) circularity (f) perimeter (g) minimum bounding ellipse

1) Shape Feature: It is created to illustrate a shape property: size, area, perimeter, axis-length, and convex-area. The descriptor attempts to determine the geometric appearance of the cells of the lymphocyte. In this research, the shape features are extracted from an entirely explosive cell as well as from its nuclei. On the other way, features of the cytoplasm are not be used as solitary features since most of the cytoplasm features are already exist in the lymphocyte cell. For the extraction of knowledgeable Shape feature of the lymphocyte cells and its nuclei, a total number of 13 shape descriptors are used. However, these features can be considered as a primary descriptor of an image; they are simple and always applicable. Some of the shape features: area, rectangular bounding box, convex hull, circularity, perimeter, and minimum bounding ellipse, for a particular lymphocyte cell is displayed in Fig. 3.

2) Texture Feature: It is a key feature for the automatic leukemia detection. For capturing the features of the nucleus chromatin pattern, we use GLCM and GLRLM based texture features.

GLCM: It is a popular statistical texture analysis technique to extract texture-information from an image [8], [15], [16], [21]. It is a second-order-statistical-feature. It utilizes the spatial relationship among reference and neighboring pixel [8]. The computation of GLCM depends on the number of gray levels n, the distance between Pixels (d), and the angle (θ). Computational cost is directly proportional to the number of gray level for calculating texture feature. Quantization merges comparable gray levels into the image and the effects of noise decrease to a certain extent. It is very important to make a note that if the texture patterns come from noise or artifacts, then the texture data could not be adequately represented by the GLCM [8]. The distance (d) is used to specify the distance between a pair of pixels. Similar to the distance parameter, the direction of the analysis is also an important parameter. The mostly used directions are 0^o, 45^o, 90^o, 135^o.

$$P(k,l) = \sum_{c=1}^{X}\sum_{d=1}^{Y} \left\{ \begin{array}{ll} 1, & if\ f(c,d) = 1\ and\ f(c+\delta c, d+\delta d) = j \\ 0, & otherwise \end{array} \right\} \tag{3}$$

where (c,d) and $(c+\delta c, d+\delta d)$ denote the locations of the reference and neighborhood pixel, respectively. $(\delta c, \delta d)$ illustrates the distance of an element $P(k,l)$ in GLCM consisting intensities k and l, respectively.

To make the classification more effective, each entry of GLCM should have a probability value [8]. Hence, we focus the estimation of normalized GLCM. It is defined as:

$$Q(k,l) = P(k,l) / \sum_{c=1}^{L-1}\sum_{d=1}^{L-1} P(k,l) \tag{4}$$

GLRLM: It is an efficient texture-analysis technique. It determines the length of homogeneous runs for every gray-value. Run length is defined as the number of adjacent-pixels that have a similar intensity in a specific direction [3]. Tang [17] has presented a GLRLM based feature extraction approach for effective classification. They have extracted 11 crucial texture-features. Moreover, we have employed GLRLM to extract texture-features of the nucleus of a lymphocyte-image. GLRLM will result in 11 features in four directions; in total, it will generate a vector of 44 features.

3) Color features: The chromatic composition of an image can be seen as a color distribution in the sense of the probability. So, easiest and most commonly used representation of color content or information is a color histogram. It represents the probabilities inherent to the intensities of the three color channels and captures the overall color distribution of an image. In this research, color features are developed from different color spaces: RGB and HSV. Both the color image was disintegrate into its three distinct color channels having the red, green, and blue values in the RGB and hue, saturation, and value in the HSV color space. Then the histogram of each color-band of the image is evaluated. Therefore totally 36 Color features are extracted for every sub-image of the nucleus.

E. Feature Reduction

The extracted feature-vector is of large dimension. For effective classification, we require to reduce its dimension significantly. Hence, we employ PCA based feature reduction technique to extract more significant features. Then these features are normalized, whose values vary between 0 to 1.

F. Classification

1) Separation of testing and training data: For effective classification, the normalized data has to be segregated into

two different sets: the training set and the testing set. ALL-IDB1 and ALL-IDB2 datasets contain 108 and 260 images, respectively. We use 70% and 30% of ALL-IDB1 dataset for training and testing purposes, respectively.

2) Classification using SVM: SVM is a supervised machine learning classifier. The SVM classifier is formally defined for the separation of the hyperplane. For a given labeled training data, it results in an optimal-hyperplane, which categorizes new classes. Powerful learning and better generalization ability make it more popular [18]. The basic procedure for the classification of the data using SVM can be summarized as follows.

i Prepare the dataset of the feature vector and divide it into the training and testing set.

ii Prepare validation set out of training set (K fold).

iii Selection of the feature using an appropriate method.

iv Finding the best hyper-parameter.

v Testing the model with the test data set.

vi Visualize the hyperplane.

III. Result and Discussion

The experiments are done using a PC having 3.1 GHz, Intel Core-i5 processor and 4 GB RAM. The simulation is done using MATLAB 2017a.

A. Pre-processing

Each of microscopic-image is preprocessed to enhance the quality. From Fig. 4, we visualize that the CLAHE based preprocessed image has more contrast and more clear vision than the original one.

1) Lymphocyte segmentation: Color-based-segmentation using k-means clustering is applied to the preprocessed image to segment the lymphocyte. The resultant image is shown in Fig. 5. From the figure, we observed that the desire WBCs are segmented at the third cluster. From these figures, we notice that the color based segmentation using k-means clustering illustrates very good performance.

2) Feature Extraction and Classification: In this stage, the crucial features: shape, texture, and color features, are extracted from the segmented image. GLCM and GLRLM algorithms are used to extract the feature vector of the nucleus. From the texture features, we notice that the feature value of healthy- and unhealthy- cells differ.

The proposed scheme (GLCM+GLRLM+PCA+SVM), has been carried out on the resultant segmented image, to effectively classify healthy- and unhealthy- lymphocytes. The following performance measures: sensitivity, specificity, and accuracy [2] are used to validate the proposed method.

$$Sensitivity = \frac{TP}{FN + TP} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$Accuracy = \frac{TN + TP}{FN + TN + FP + TP} \tag{7}$$

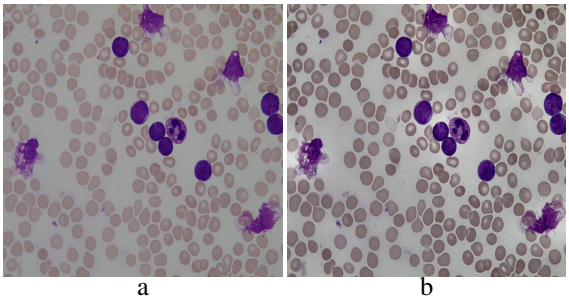

Fig. 4. Preprocessing: (a) Original image; (b) Resultant CLAHE preprocessed image

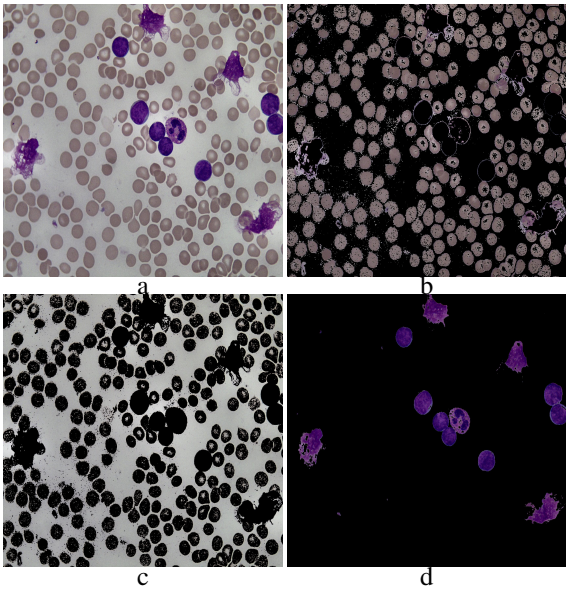

Fig. 5. Color-based-segmentation by employing k-means-clustering : (a) Preprocessed image, whereas (b), (C), and (d) represent images in cluster 1, 2, and 3, respectively

Where, TP: True-Positive

TN: True-Negative

FN: False-Negative

FP: False-Positive We employ 10-fold-cross-validation to make the classifier more reliable. SVM is trained using 700 cells (extracted from the ALL-IDB1 dataset-images). Among these cells, 465 are unhealthy, and 235 are healthy. Similarly, for the testing purpose, 300 cells are used; in which 178 are unhealthy, and 122 are healthy cells. Here, we extract 60 texture-features using GLCM and GLRLM. Moreover, we extract 13 Shape-features and 18 color-features. PCA algorithm is applied to select 50 crucial features, which is 55% of the total features. Finally, SVM, KNN, and back-propagation neural network (BPNN) [19] are used to classify WBCs. The average computation time (CT) required per lymphocyte for pre-processing and segmentation are 0.03s and 0.26s, respectively. Whereas, CT needed per lymphocyt for feature-extraction, feature-reduction, and classification are 0.099s, 0.012s, and 0.005s, respectively. The average computation-time of a lymphocyte is 0.406s. TableI demonstrates the superiority of SVM over KNN and BPNN, with better accuracy, sensitivity, and specificity.

TABLE I. PERFORMANCE MEASURE FOR DIFFERENT CLASSIFIER

classifier	Sensitivity(%)	Specificity(%)	Accuracy(%)
KNN	86.23	89.90	91.20
BPNN	88.50	90.56	93.50
SVM	92.64	93.06	96.00

IV. CONCLUSION

This paper presents an efficient scheme to detect healthy- and unhealthy(ALL)- lymphocytes. The presented CLAHE dynamically increases the contrast level of the image and also improves the image quality successfully. Then, leukocytes are extracted using color based k-means clustering algorithm. The scheme employs GLCM and GLRLM to extract texture features. Moreover, it also emphasizes to extract shape and color features. Finally, SVM with RBF kernel is employed to classify WBCs into healthy and ALL affected cell. From the experiment, we observed that the proposed scheme displays superior performance with 96% of accuracy and 92.64% sensitivity. As future work, the performance can be further improved using larger datasets.

REFERENCES

[1] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, and C. Sultan, "Proposals for the classification of the acute leukemias French-American-British (FAB) co-operative group," *British Journal of Hematology*, vol. 33, no. 4, pp. 451–458, Aug. 1976.

[2] P. K. Das, S. Meher, R. Panda, and A. Abraham, "A Review of Automated Methods for the Detection of Sickle Cell Disease," *IEEE Reviews in Biomedical Engineering*, vol. 13, pp. 309–324, 2020.

[3] S. Mishra, B. Majhi, and P. K. Sa, "GLRLM-Based Feature Extraction for Acute lymphocytic Leukemia (ALL) Detection," *In Recent Findings in Intelligent Computing Techniques*, pp. 399–407, 2018.

[4] M. Ghosh, D. Das, S. Mandal, C. Chakraborty, M. Pala, A. K. Maity, S. K. Pal, and A. K. Ray, "Statistical pattern analysis of white blood cell nuclei morphometry," *In 2010 IEEE Students Technology Symposium (TechSym), IEEE*, pp. 59–66, 2010.

[5] Q. Liao and Y. Deng, "An accurate segmentation method for white blood cell images," *In Proceedings IEEE International Symposium on Biomedical Imaging, IEEE*, pp. 245–248, 2002.

[6] J. Angulo and G. Flandrin, "Microscopic image analysis using mathematical morphology:Application to haematological cytology," *Science, technology and Education of Microscopy: An overview*, vol.1, pp. 304–312, 2003.

[7] M. Madhukar, S. Agaian, and A. T. Chronopoulos, " New decision support tool for acute lymphoblastic leukemia classification'" *In Image Processing: Algorithms and Systems X; and Parallel Processing for Imaging Applications II*, vol. 8295, pp. 829518, 2012.

[8] S. Mishra, B. Majhi, P. K. Sa, and L. Sharma, "Gray level co-occurrence matrix and random forest based acute lymphocytic leukemia detection," *Biomedical Signal Processing and Control*, vol. 33, pp. 272–280, 2017.

[9] A. Rahman and M. M. Hasan, "Automatic Detection of White Blood Cells from Microscopic Images for Malignancy Classification of Acute Lymphoblastic Leukemia," *In 2018 International Conference on Innovation in Engineering and Technology (ICIET), IEEE*, pp. 1–6, Dec., 2018.

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

[11] R. C. Gonzalez and R. E. Woods, "Digital image processing," *Pearson*, 2018.

[12] R. GeethaRamani, and L. Balasubramanian, "Retinal blood vessel segmentation employing image processing and data mining techniques for computerized retinal image analysis," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 102–118, 2016.

[13] S. Arslan, E. Ozyurek, and C. GunduzDemir, "A color and shape based algorithm for segmentation of white blood cells in peripheral blood and bone marrow images," *Cytometry Part A*, vol. 85, no. 6, pp. 480–490, 2014.

[14] A. K. Mohanty, S. Beberta, and S. K. Lenka, "Classifying benign and malignant mass using GLCM and GLRLM based texture features from mammogram," *International Journal of Engineering Research and Applications*, vol. 1, no. 3, pp. 687–693, 2011.

[15] F. Alaei, A. Alaei, U. Pal, and M. Blumenstein, "A comparative study of different texture features for document image retrieval," *Expert Systems with Applications*, vol. 121, pp. 97-114, 2019.

[16] R. M. Haralick and K. Shanmugam, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, vol. 6, pp. 610–621, 1973.

[17] X. Tang, "Texture information in run-length matrices," *IEEE Transactions on Image Processing*, vol. 7, no. 11, pp. 1602–1609, 1998.

[18] J. Laosai and K. Chamnongthai, "Acute leukemia classification by using SVM and K-Means clustering," *In 2014 International Electrical Engineering Congress (iEECON), IEEE*, pp. 1–4, 2014.

[19] A. Gautam, V. Bhateja, A. Tiwari, and S. C. Satapathy, "An improved mammogram classification approach using back propagation neural network," *In Data Engineering and Intelligent Computing*, pp. 369–376, 2018.

[20] S. Mohapatra, D. Patra, and S. Satpathy, "An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images," *Neural Computing and Applications*, vol. 24, pp. 1887–1904, 2014.

[21] L. Putzu, G. Caocci, and C. D. Ruberto, "Leucocyte classification for leukaemia detection using image processing techniques," *Artificial intelligence in medicine*, vol. 62, pp. 179–191, 2014.

[22] S. Mishra, B. Majhi, and P. K. Sa, "Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection," *Biomedical Signal Processing and Control*, vol. 47, pp. 303–311, 2019.

[23] L. H. Vogado, R. M. Veras, F. H. Araujo, R. R. Silva, and K. R. Aires, "Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 415–422, 2018.

[24] G. H. Park, H. H. Cho, and M. R. Choi, "A contrast enhancement method using dynamic range separate histogram equalization," *IEEE Trans. Consumer Electronics*, vol. 54, pp. 1981–1987, 2008.

[25] G. Yadav, S. Maheshwari,and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," *In 2014 Int. Conf. Advances in Computing, Communications and Informatics (ICACCI), IEEE*, pp. 2392–2397), 2014.