A Hybrid Classification Scheme Using 2D-SWT and SVM for the Detection of Acute Lymphoblastic Leukemia

Sonali Mishra, Banshidhar Majhi, and Pankaj Kumar Sa

Abstract—**Acute lymphocytic leukemia (ALL) is a heterogeneous disease that differs considerably in their cellular and molecular characteristics and also affects a larger proportion of world population Advanced and specific techniques are available for classifying leukemia types however they are exceptionally costly and not accessible to many doctor's facilities in developing nations. Image processing can be a way to detect the disease more precisely and conjointly takes a trifle time. This paper presents a hybrid scheme for identification and classification of ALL. The suggested scheme utilizes 2D-SWT to extricate the texture features from the blood smear. Later on, the extracted features are fed to SVM classifier to get the classification results. The experimental results for leukemia classification show that the suggested method outperforms other standard classifiers regarding accuracy. The accuracy is found to be 99.56% with the help of SVM-R classifier.**

Index Terms—**Leukemia, CAD system, 2D-SWT, Support vector machine.**

I. Introduction

Cancer is a kind of malady that makes the cells of the body change its attributes and cause the abnormal development of cells. In 2017, there is a total of 1,688,780 cases of cancer diagnosed, and records of deaths are 600,920 in the US alone [1]. Leukemia is a form of liquid cancer caused due to the unnatural expansion in the production of immature leukocytes in the blood. Acute lymphoblastic leukemia (ALL) is caused due to the over production of immature white blood cells called lymphoblasts (also known as blast cells), which inhibit the production of normal white blood cells (WBCs). The term 'acute' signifies the rate of progression of the disease to the other parts of the body such as lymph nodes, liver, and spleen, etc. and if left untreated the disease can pose a danger to life a few months [2].

A standout amongst other diagnostic methods is the professional perception of the blood smear for the detection of the disease. This approach requires only an image of blood slide for analysis. The primary identification proof is performed by the by the Complete Blood Count (CBC) [3]. Bone marrow biopsy is proposed to the patient if CBC is by all accounts unusual. The image of the blood smear is caught with the microscope and is considered for morphological examination which is accomplished by a hematologist. The analysis thus gets restricted to the limits of knowledge and expertise of the medical staff performing the diagnosis. The accuracy is non-standard, measures of accuracy vary from one operator to the other. Hence, an automated approach towards analysis of the blood smear images for leukemia detection is inevitable. Machine learning techniques are now being incorporated towards the development of a Computer Aided Detection (CAD) System which would perform the task of diagnosis with minimal efforts and better efficiency.

The proposed methodology infers a Stationary Wavelet Transform based feature extraction technique from the nucleus and cytoplasm sub-images. The extracted features are subsequently fed to the SVM to characterize the blood cells as normal or abnormal. The simulation for the proposed scheme is done in MATLAB by using the image processing toolbox. The input database consists of 108 sample image taken from healthy and infected patients [4]. In the case of an infected blood sample, there is a large presence of immature lymphocytes (known as lymphoblasts). The number of lymphoblasts is a powerful indicator of the disease. Thus detection and counting of lymphoblast is the most trusted way of diagnosis. The implementation occurs in four stages. The first stage identifies the lymphocytes based on its physical characteristics and separates it from the rest of the blood sample, the second stage is the separation of grouped and individual lymphocytes, later on separating the nucleus and cytoplasm. The third phase deals with extracting features using an appropriate feature extraction technique. The final phase deals with the classification of the cells using a classifier.

The structure of the paper is illustrated as follows. Section II outlines the related work. Section III describes the automated detection of leukemia. Section IV presents the experimental results. The conclusion of the proposed scheme is given in Section V.

II. Related Work

The authors in [5] have made use of Lab color conversion and K-means clustering for obtaining the segmented nucleus and cytoplasm followed by morphological feature extraction. Particle Swarm Optimization (PSO), Radial basis functional network (RBFN) and Back-propagation neural network (BPN) were then applied for classification of which PSO gave the maximum accuracy of 95%. The authors in [6] have specified a model for segmentation of the RBCs. The segmentation model involves the color conversion of the RGB image to YCbCr to overcome the general issue regarding illumination. It is then followed by masking and use of morphological operators such as binary erosion, etc. to obtain the segmented

Manuscript received August 9, 2017; revised November 5, 2017.

The authors are with the National Institute of Technology, Rourkela, India, 769008 (e-mail: smishra.nitrkl@gmail.com, bmajhi@nitrkl.ac.in, pankajksa@nitrkl.ac.in).

doi: 10.18178/ijmlc.2017.7.6.650

WBC nucleus. It is used as a mask for isolating the WBC from the image. Segmented WBC is then passed through the watershed algorithm to separate overlapping cells. The authors in [7] have proposed a working model for segmentation of WBCs. Mathematical morphological operations were performed on the image, and then watershed transform was applied using image forest transform (IFT) to isolate the WBC nuclei from the background image. After nuclei segmentation, a series of morphological openings and size distribution data is then used for cytoplasm extraction from the remaining background image. Shitong *et al.* [8] have suggested a method for WBC detection which analyses two existing methods, i.e., TSMM and NDA. NDA unifies the benefits of TSMM and the fuzzy logic technique and defeats their shortcomings. With this new algorithm, almost all leukocytes were detected successfully, and the outline of each detected cell is closely perfect. Neelam *et al.* [9] have suggested a model for segmentation of WBCs which is a two-step process performed on the HSV-equivalent of the image. The image segmentation is performed using K-Means clustering followed by parameter refinement by EM-algorithm. Shape, texture and color features were then extracted to be applied to a wide variety of classifiers of which the best classification accuracy was yielded using Neural networks of 97%, then 94% using SVM. For segmentation of ALL blast cells from pathological blood stain, the authors in [10] have used image color transformation coupled with mathematical morphology followed by the watershed algorithm. The proposed method is capable of successfully segmenting 9 various subtypes of ALL blast cells with accuracy results of approximately 96%. Zhang *et al.* [11] have used wavelet transform to extract features which were reduced using Principal Component Analysis (PCA). The maximum classification accuracy for the radial basis kernel is found to be 99.38%. Rota *et al.* [12] have comparatively implemented three different approaches for detection of lymphocytes, using Flow Cytometry data samples. The first built a completely distinctive baseline using the Support Vector Machine, the second is a two-phase technique, the first phase being an unsupervised feature learning using a Stacked Auto-Encoder Neural Network, the second involves utilizing the formerly trained Neural Network to perform inference on the classification of data. The third approach is entirely generative and is developed by the Gaussian Mixture Model estimation and Bayes decision. The generative model gave better accuracy results comparatively. Mohapatra *et al.* [13] have used *k*-means clustering for lymphocyte identification followed by shadowed c-means clustering for nucleus and cytoplasm separation. Morphological features were then extracted and reduced using *t*-test feature selection. An ensemble of classifiers was then employed for classification to give an accuracy of 94.73%. Moreover, in [14], [15], authors have extracted different texture features to classify the blood cells.

III. Proposed Methodology

A CAD system can be a valuable tool for the detection of ALL. The overall methodology is exhibited in Fig. 1.

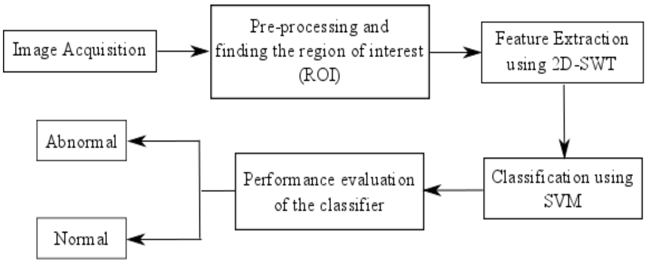


Image Acquisition
Pre-processing and finding the region of interest (ROI)
Feature Extraction using 2D-SWT
Classification using SVM
Performance evaluation of the classifier
Abnormal
Normal

Fig. 1. Proposed block diagram for the classification of ALL.

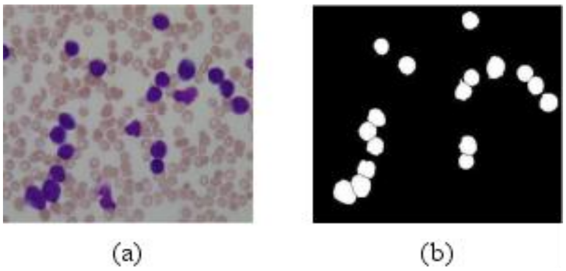
A. Pre-processing

The automatic diagnostic system of leukemia is a vital process, which can be used to help the specialists in diagnosis and evaluation process. The blood cell segmentation is an initial and important stage in developing such system. This paper confers a segmentation method to extract the blast cells in the blood. During the experiment, it is witnessed that, simply transforming the input image to the gray scale image will not give us required segmentation result due to the almost same contrast of the WBC and RBC in the image. We have used contrast enhancement to adjust the contrast of the image. The intensity transformation is applied to each component of the input RGB image and is done by the subsequent equation.

$$y(i,j)=\frac{255}{\text{high - low}}\times(x(i,j)-\text{low})$$

where, Here $y(i, j)$ is the transformed intensity, low and high are the minimum and maximum intensity to be changed of the current color component, $x(i, j)$ is the input intensity.

The adjusted color components merged to form a new color image containing only WBC from the image. To segment the touching cells, we have used the marker-controlled watershed algorithm [16]. The detection of WBCs from a microscopic image is shown in Fig. 2. After detection of WBC, the bounding box technique is used to isolate single leukocyte image which is then fed for the nucleus-cytoplasm separation.



(a). Microscopic image, (b). Detected WBCs

Fig. 2. Detection of WBCs in a sample microscopic image.

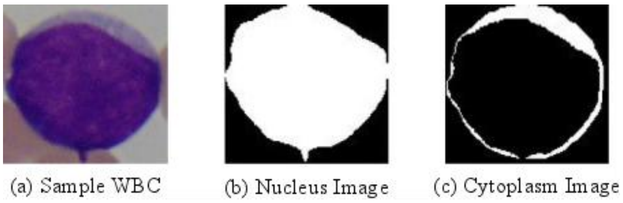


Fig. 3. Separation of nucleus and cytoplasm from A WBC.

The nuclei of WBCs have a more robust distinction than the cytoplasm and can be separated by amalgamating the binary image captured from the green component and also the a* part [17] from the LAB color space. A subtraction operation is then conducted to extract the cytoplasm. This operation is done between the leukocyte image and the image containing the nucleus. The procedural portrayal of this level is shown in Fig. 3.

B. *Feature Extraction*

The discrete wavelet change (DWT) is an effective scientific usage of the wavelet transform (WT) that is valuable for investigating the image at various ranges or resolutions. The WT provides the time-frequency localization of a signal, which is useful for classification. Unlike DWT, Stationary Wavelet Transform preserves the time invariance property. At level 1, the SWT convolutes the signal with a low-pass and high-pass filter, respectively to produces the approximation coefficients and detail coefficients. Unlike classical DWT, SWT only up samples the signal to produce the approximation and detailed coefficients which are same as the signal length at a particular level.

As the disintegrations of SWT are not devastated, it needs a larger space to store the features and is computationally costlier to use the features directly. Hence, energy and entropy have been calculated from the approximation and detailed coefficients and used as the primary features. Wavelet energy and entropy are the two fundamental feature descriptors which have been effectively utilized as a part of numerous applications. Fig. 4 demonstrates the second level 2D-SWT decomposition of a nucleus and cytoplasm image.

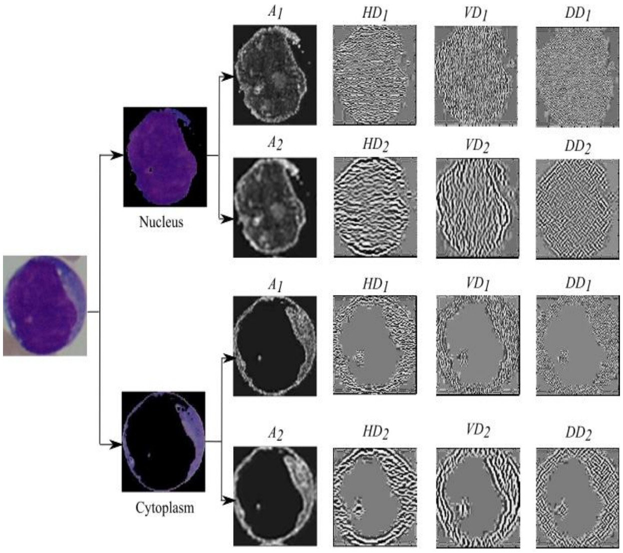


Fig. 4. A 2-level SWT decomposition of a nucleus and cytoplasm image (A: Approximation coefficient, HD: Horizontal detail, VD: Vertical detail, DD: High Pass detail)

C. *Classification*

This Section describes the basic work on support vector machines (SVM) for classification problems. It is a binary learning machine which is used for creating a hyperplane to be used as a 2-class classifier. In SVM, the input data patterns are transformed into a high dimensional space using a non-linear mapping and is given by,

$$y_k = \varphi(x_k)$$

where y_k is the output data pattern, k=1, 2, 3, …, n,
x_k is the input data pattern, patterns
It then searches for a linear discriminant function

$$y = a^t y + b$$

in the feature space. The SVM solution is obtained through maximizing the margin between the separating hyperplane and the data, where the margin is defined as, $\frac{|y|}{\|a\|}$. The basic goal of SVM is to find the weight vector a that maximizes b and minimizes $\|a\|$ so that,

$$\frac{|y|}{\|a\|} \geq b,$$

with the help of an objective function.

The subsequent investigation compares the SVM classifier with other familiar classifiers, including Na ïve Bayes, *k*-NN, MLP, and Random Forest. Also, four famous kernels, including, linear, polynomial, and sigmoid were inspected in this analysis. The pseudo code of the suggested methodology is exhibited in Algorithm 1.

```
Algorithm 1: Pseudo Code for the Proposed Algorithm
Require: Microscopic blood images,
    X: total number of images collected from ALL-IDB1
Ensure: Classification of cells as normal or abnormal.
-----------------------------------------------------------------
Step 1: Marker-based watershed segmentation of images to
separate leukocytes
Loop on i= 1 to X
    1. Read the input image
    2. Preprocess the image to detect all the leukocytes
    3. Apply Marker-based watershed algorithm to separate
       the grouped leukocytes
    4. Crop the image using bounding box technique to
       extract all the individual leukocytes
    5. Separate nucleus and cytoplasm and resize the image
       to 128×128.
End Loop
Step 2: 2D-SWT based feature extraction
Loop on j= 1 to M, where M is the cropped image of X
    1. Read the nucleus and cytoplasm image of size
       128×128
    2. Apply 2D-SWT (upto 2nd level) to extract the texture
       features
    3. Calculate the energy and entropy values for each
       co-efficient and store it in a matrix of size M×N,
       where N is the number of features
End Loop
Step 3: Classification using Support Vector Machine
    1. Create a dataset S(m_j, y_j), y_i∈Y (Output class)
    2. Train the classifier
    3. Calculate the performance measures to draw a
       conclusion.
```

IV. Experimental Results and Evaluation

The tests were performed on a personal computer with an internal memory of 4 GB, running under Windows 10 OS using Matlab toolbox. The performance of the suggested method is examined with other existing methods regarding true positive rate (TPR), true negative rate (TNR), and accuracy. The performance measures are presented in Table I.

Specimen blood samples are collected from a public database ALL-IDB [4]. The proposed method is trained and tested with the database ALL-IDB1. This database consists of 108 original blood sample images, which is divided into two parts where first part consists of the image that is not suffering from leukemia (lymphoblast are absent) and the second part consists of images suffering from leukemia (lymphoblast are present). These images were taken with an Olympus C2500L camera and have a resolution of 1712×1368. Performance analysis is necessary within the machine-driven arrangement, which was carried out in this paper to estimate the ability of the classifiers for ALL detection in blood images. A 5-fold cross validation procedure is applied to solve the over fitting problem.

TABLE I: CONFUSION MATRIX

	Positive	**Negative**	**Performance Measures**
Positive	TP	FP	PPV= TP/TP+FP
Negative	FN	TN	NPV= TN/TN+FN
Performance Measures	**TPR= TP/TP+FN**	**TNR= TN/TN+FP**	**Accuracy= (TP+TN)/ TP+FP+TN+FN**

The suggested scheme applies 2D-SWT to extract texture features from the nucleus and cytoplasm sub-images. Haar wavelet is utilized to calculate the energy and entropy from level-2 decompositions which produce $8 \times 2 = 16$ features. So, the total number of features extracted are 32 (nucleus and cytoplasm). Calculating energy and entropy value helps in reducing the feature size from $128 \times 128 \times 8$ to 32. The energy and entropy values of a normal and abnormal nucleus and cytoplasm image are recorded in Table II. Finally, the result of the feature extraction gives us feature matrix of size 695×32.

TABLE II: ENERGY AND ENTROPY VALUES FOR LEVEL-2 NUCLEUS AND CYTOPLASM SUB-IMAGE

	Normal		Abnormal	
	Nucleus	Cytoplasm	Nucleus	Cytoplasm
Average Energy	9.2177e+08	5.1462e+08	1.5656e+09	1.0299e+09
Average Entropy	1.17767	1.31853	1.07309	1.28768

The 32 features with an output class level were given to the SVM classifier. In Fig. 5, the classification accuracies of SVM classifier with the extracted features from the nucleus and cytoplasm sub-images are represented. Table III, Table IV, and Table V presents the fold-wise result of the 5-fold cross validation procedure for the determination of accuracy, sensitivity, and specificity for different classifiers. The optimum has been achieved with an accuracy of 99.56% using the SVM classifier. The sensitivity and specificity obtained for the same is 100% and 99.35% respectively. We have also evaluated the nucleus and cytoplasm features separately to compare the performance of the classifier. For accurate classification of samples, both nucleus and cytoplasm features are required. The individual accuracies obtained by the nucleus and cytoplasm features are 97.56% and 96.11% respectively. The performance of the proposed scheme was studied with the varied range of features, and the accuracies are shown in Fig. 5. It can be observed from the figure that, combining nucleus and cytoplasm features can comparatively increase the classification accuracy.

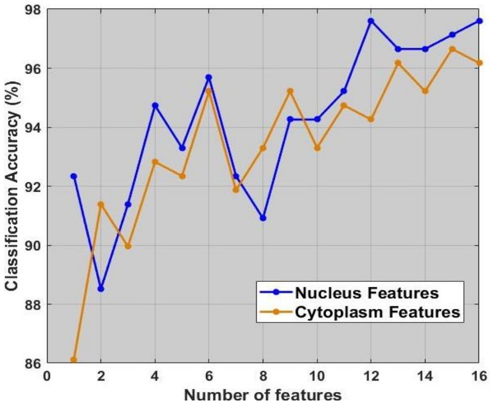


Fig. 5. Accuracy rate with the increase in number of features.

TABLE III: COMPARISON OF ACCURACY OF VARIOUS CLASSIFIERS OVER 5-FOLD

Classifier	fold					Avg. Accuracy (%)
	1	2	3	4	5	
NB	70.50	63.30	72.66	68.34	72.66	69.44
k-NN	93.52	92.80	98.56	97.12	93.52	95.10
MLP	98.56	96.40	97.12	97.12	97.12	97.26
RF	100.0	100.0	98.56	98.56	98.56	99.13
SVM-R	**100.0**	**100.0**	**100.0**	**99.28**	**98.56**	**99.56**

TABLE IV: COMPARISON OF SENSITIVITY (TPR) OF VARIOUS CLASSIFIERS OVER 5-FOLD

Classifier	fold					Avg. TPR(%)
	1	2	3	4	5	
NB	89.13	91.30	84.78	93.47	91.30	89.99
k-NN	92.85	92.78	100.0	98.90	94.68	95.84
MLP	95.65	89.13	95.56	93.47	91.30	93.02
RF	100.0	100.0	95.65	97.82	100.0	98.69
SVM-R	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

TABLE V: COMPARISON OF SPECIFICITY (TNR) OF VARIOUS CLASSIFIERS OVER 5-FOLD

Classifier	fold					Avg. TNR(%)
	1	2	3	4	5	
NB	61.29	49.46	66.66	55.91	63.44	59.35
k-NN	95.12	92.85	95.83	93.75	91.11	93.73
MLP	100.0	100.0	97.84	98.42	100.0	99.73
RF	100.0	100.0	100.0	98.92	97.84	99.35
SVM-R	**100.0**	**100.0**	**100.0**	**98.92**	**97.84**	**99.35**

V. CONCLUSION

Detection of ALL is required to evaluate the diagnosis and might resolutely modification the treatment set up of the suspected patients with cancer of the blood. In conventional leukemia subtyping, hematologists externally portray lymphoblasts display in the blood smear, and this assessment procedure is regularly moderate, subjective in nature and error prone. Therefore, in this paper, we have suggested a quantitative methodology utilizing image processing and machine learning techniques to classify blast cells. 2D-SWT is harnessed to extract the texture features from the nucleus and cytoplasm sub-images. The extracted features are then given to the SVM classifier to get the classification results. To validate the proposed scheme, it is being evaluated with some other standard classifiers. The results show that the proposed scheme outperforms others which achieves an accuracy of 99.56% with 32 features. The achieved sensitivity and specificity of the classifier is 100.00% and 99.35% respectively.

REFERENCES

[1] Cancer Facts and Figures. (2017). [Online]. Available: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer facts-figures-2017.html

[2] S. Mishra, B. Majhi, and P. K. Sa, "A survey on automated diagnosis on the detection of Leukemia: A hematological disorder," in *Proc. 3rd IEEE International Conference on Recent Advances in Information Technology*, 2016, pp. 460-466.

[3] R. D. Labati, V. Puiri, and F. Scotti, "ALL-IDB: The acute lymphoblastic leukemia image database for image processing," in *Proc. IEEE International Conference on Image Processing (ICIP), Brussels*, Sept. 2011.

[4] *ALL-IDB Dataset for ALL Classification.* Available: http://crema.di.unimi.it/~fscotti/all/

[5] G. Prabu and H. H. Inbarani, "PSO for acute lymphoblastic Leukemia classification in blood microscopic images," *History*, vol. 12, no. 30, pp. 138-145, 2015.

[6] J. M. Sharif, M. F. Miswan, M. A. Ngadi, M. S. H. Salam, and M. M. B. A. Jamil, "Red blood cell segmentation using masking and watershed algorithm: A preliminary study," in *Proc. IEEE International Conference on Biomedical Engineering (ICoBE)*, 2012, pp. 258-262.

[7] L. B. Dorini, R. Minetto, and N. J. Leite, "White blood cell segmentation using morphological operators and scale-space analysis," in *Proc. XX IEEE Brazilian Symposium on Computer Graphics and Image Processing*, October 2007, pp. 294-304.

[8] W. Shitong and W. Min, "A new detection algorithm (NDA) based on fuzzy cellular neural networks for white blood cell detection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 5-10, 2006.

[9] N. Sinha, and A. G. Ramakrishnan, "Automation of differential blood count," in *Proc. IEEE Conference on Convergent Technologies for the Asia-Pacific Region (TENCON 2003)*, October 2003, vol. 2, pp. 547-551.

[10] H. T. Madhloom, S. A. Kareem, and H. Ariffin, "Computer-aided acute leukemia blast cells segmentation in peripheral blood images," *Journal of Vibroengineering*, vol. 17, no. 8, pp. 4517-4532, 2015.

[11] Y. Zhang and L. Wu, "An MR brain images classifier via principal component analysis and kernel support vector machine," *Progress in electromagnetics research*, vol. 130, pp. 369-388, 2012.

[12] P. Rota, S. Groeneveld-Krentz, and M. Reiter, "On automated flow cytometric analysis for MRD estimation of acute lymphoblastic leukaemia: a comparison among different approaches," in *Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, November 2015, pp. 438-441.

[13] 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, no. 7-8, pp. 1887-1904, 2014.

[14] S. Mishra, L. Sharma, B. Majhi, and P. K. Sa, "Microscopic image classification using DCT for the detection of acute lymphoblastic Leukemia (ALL)," in *Proc. International Conference on Computer Vision and Image Processing*, 2016, pp. 171-180.

[15] S. Mishra, B. Majhi, P. K. Sa, and F. Siddiqui, "GLRLM based feature extraction for Acute Lymphoblastic Leukemia (ALL) detection," in *Proc. 5th International Conference on Advanced Computing, Networking, and Informatics (ICACNI)*, Springer, 2017.

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

[17] L. Putzu, G. Caocci, and C. Di Ruberto, "Leucocyte classification for leukemia detection using image processing techniques," *Artificial Intelligence in Medicine*, vol. 62, no. 3, pp. 179-191, 2014.

Sonali Mishra is currently pursuing her Ph.D. in the Department of Computer Science and Engineering, National Institute of Technology (NIT), Rourkela, India, under the guidance of Prof. Banshidhar Majhi. She received her M.Tech. degree in 2014 from the Department of Computer Science and Engineering, VSSUT Burla University, India. She has received her B.Tech. degree in 2012 from Department of Computer Science and Engineering, BPUT University, India. Her research interests include Wireless Sensor Networks, Medical Image Processing, and Pattern Recognition. Currently she is working on the designing of an automated system for the detection of Acute Lymphoblastic Leukemia.

Banshidhar Majhi received his PhD degree from Sambalpur University, Odisha, India, in 2001. He is currently working as a professor in the Department of Computer Science and Engineering at National Institute of Technology, Rourkela, India. His fields of interests include image processing, data compression, cryptography and security, parallel computing, soft computing, and biometrics. He is a professional member of MIEEE, FIETE, LMCSI, and FIE. He has authored more than hundred papers in journals and conferences of international repute.

Pankaj K. Sa earned his PhD in computer science from NIT Rourkela in the year 2010. He is currently an assistant professor in the CSE Department of NIT Rourkela. His research interest includes computer vision, biometrics, visual surveillance. He is a member of CSI and IEEE. He has coauthored a number of research articles in various journals, conferences, and book chapters. He has co-investigated some R&D projects funded by SERB, PXE, DeitY, ISRO. He has been conferred with various prestigious awards and honors. Apart from research and teaching, he is also actively involved with the automation of NIT Rourkela, where he conceptualizes and engineers the automation process.