Automated Detection of White Blood Cells Cancer Diseases

[1]ALINA AHMED, [2]AYESHA FATIMA
[1]PG Scholar, Dept of ES, Nawab Shah Alam Khan College of Engineering and Technology, Hyderabad, TS, India.
[2]Assistant Professor, Dept of ES, Nawab Shah Alam Khan College of Engineering and Technology, Hyderabad, TS, India.

Abstract: Automated diagnosis of white blood cells cancer diseases such as Leukemia and Myeloma is a challenging biomedical research topic. Our approach presents for the first time a new state of the art application that assists in diagnosing the white blood cells diseases. We divide these diseases into two categories; each category includes similar symptoms diseases that may confuse in diagnosing. Based on the doctor's selection, one of two approaches is implemented. Each approach is applied on one of the two diseases category by computing different features. Finally, Random Forest classifier is applied for final decision. The proposed approach aims to early discovery of white blood cells cancer, reduce the misdiagnosis cases in addition to improve the system learning methodology. Moreover, allowing the experts only to have the final tuning on the result obtained from the system. The proposed approach achieved an accuracy of 93% in the first category and 95% in the second category.

Keywords: Random Forest Classifier.

I. INTRODUCTION

As indicated by the WHO [1], Cancer is consider as the second driving reason for death on the planet and killed 8.8 million individuals in 2015, credited to the demise of about one of every six passings around the world. As per WHO [1] Low-pay and center salary nations represented practically 70% of disease passings. In addition, Egypt is ranked nineteenth among 176 nations worldwide in the rate of leukemia death. The discovery of these diseases in the early days greatly influences the time frame of treatment. Moreover, a portion of the sicknesses' sub-types are truly befuddling to the specialists. These days, there is an incredible propensity for demonstrative pathology to intensely depend on mechanized frameworks which can help in the analysis [2].

Picture handling is worried with advanced pictures to extricate helpful data. It is engaged with various different themes, for example, design arrive use 3, Character acknowledgment 4, 5 coin acknowledgment 6, Medical imaging 7,8. Medicinal imaging utilizes data extricated at advanced picture to improve the indicative of various illnesses. White platelets malignant growth infections; Leukemia and Myeloma, compromise individuals' life these days. Leukemia is discovered while the bone marrow produces strange white platelets, which don't work legitimately 9. It might be either intense or ceaseless. Intense Myeloid Leukemia (AML) is sub ordered to (M0, M1, M2, M3, M4, M5, M6, M7). Intense lymphoblastic leukemia (ALL) (L1, L2, L3) is sorted. Mystery Extensively, thresholding is a standout amongst the most wellknown, clear systems utilized for picture division assignments by and large and for restorative picture division undertakings specifically. 10 is another sort of malignant growth that creates Plasma cells from bone marrow cells.

This paper presented out of the blue a mixture robotized framework to encourage the analysis of various white platelets malignancy sicknesses; Subtypes of leukemia (AML, ALL) and myeloma. There are two main approaches to dispensing with perplexity while separating a portion of the subtypes. For each methodology, another mix of proportion highlights is led. Since it is invariant to scaling, we considered taking proportion highlights. At long last, we endeavored to make the framework as a specialist framework through enhancing the learning proficiency by enabling the framework to gain from the misdiagnosed input tests in which the specialists rename the sickness with the new name. Segment II talks about comparable research related to the paper.

II. RELATED WORKS

Few explorations are proposed to separate between:
1) (AML) and (ALL) maladies: Karthikeyan&Poornima proposed a methodology for identifying Leukemia in blood at beginning periods. They utilized versatile middle channel for clamor expulsion and versatile Histogram-Equalization differentiation upgrade preprocessing stage. They connected kmeans and Fuzzy c-implies grouping for division. They processed measurable, textural & geometrical highlights connected (Support Vector Machine) for grouping.

The methodology accomplished 90% of Fuzzy c-implies and 83% of k-implies utilizing Fuzzy Logic: Dataset for intelligence, control and information [12]. Another exploration by Mohapatra et al. [13] proposed a segmentation based on Fuzzy's blood image for automated detection of leukemia. They connected specific middle separation pursued in pre-processing by unsharp concealment. They used an enhanced variant of the fluffy grouping method in division. Gustafson Kessel bunching [14]. The closest neighbor characterization in the shading space of L*a*b* (L* for dainty, a* for redness greenness pivot, and b* for yellowness blueness hub) [15] is pursued. The processed highlights are two new shape highlights; the dimension of Hausdorff and

the signature form. Bolster Vector Machine (SVM) is used for characterization and on a database of 108 blood smear images of size 512 x 512 pixels, 93 percent have been achieved.

2) A methodology by Agaian et al. [16] proposed a basic procedure that naturally recognizes AML in blood spreads and fragments AML. The division was carried out by the K-Means grouping calculation in the CIELAB Color space. Highlights of the Hausdorff dimension were figured using the container tallying technique and the local binary pattern (LBP). The grouping achieved 98 percent accuracy using the American Society of Hematology (ASH) support vector machine (SVM) for the Leukemia dataset [17]. This dataset contained 80 pictures40 of patients with AML and 40 of patients without AML. The image measure used to group them was 184 x 138 pixels. Another framework for the discovery of acute lymphoblastic leukemia using the Watershed Transformation Technique proposed by Bhattacharjee and Saini [18]. They linked improvement in complexity and alteration in quality to upgrade images before division. They used the calculation of the watershed in division, secluding the platelet and the cell nucleus. They processed highlights of region, border, circularity and form factor. For characterization, Gaussian Mixture Models (GMM) and Binary Search Tree (BST) have been connected. GMM achieved 93 percent while BSTe achieved 86 percent. They linked their methodology to 150 pictures of lymphocyte cells (30 typical cells and 120 impact cells) from data sets ALL-IDB1 and ALLIDB2 [17].

3) Detection of sub-types (AML): Another methodology proposed by Sarrafzadeh et al. [19] focused primarily on separate sub-types of M2, M3 and M5 in order to evaluate the technique presented. The methodology was linked to the shading space of L*a*b*. The division is carried out by grouping K-implies to isolate leukocytes from other blood segments. Surface and shape highlights are separated so that they can be ordered using DDL. The data set of the Medical Image and Signal Processing Research Center (MISP) achieved 97.53 percent accuracy [20]. They used a dataset of 27 small images of three AML sub-kinds; 9 AML-M2, 10 AML-M3 and 8 AML-M5.

III. PROPOSED SYSTEM

Our proposed framework consists of two methodologies that contribute. One separates M5 Acute Myeloid Leukemia (AML), L1 and L2 Acute Lymphoblastic Leukemia (ALL), while the other separates the remaining sub-classifications. Due to comparable visual highlights that really befuddle specialists and can cause misclassification, we have assembled our affected diseases into two separate sets. The framework experiences one of the two proposed methodologies in accordance with the specialist wish for the information blood test. An alternative arrangement of highlights is calculated by approach. We used managed grouping when we prepared our framework with information about the sub-type tests of the diseases. The info-blood test is then characterized by the coordination similarities between it and the information prepared. The falling recognition process by two methodologies increases the overall frame productivity of the separation of diseases expressed in the area of investigation. The framework review has been published Fig.1.

A. Preprocessing & Segmentation

This phase is applied on our training and testing images. The purpose of white blood cell segmentation is to clearly extract relevant object; whole cell, from its relative background. Furthermore, we separate the cell into nucleus and cytoplasm.

1. Preprocessing: In this stage we prepare the blood sample image for the segmentation process by converting our input images from RGB color space to YCbCr space [21]. Our choice to the YCBCR color space (Y: Luminance, CB: Blue Value, CR: Red Value) was due to the reddish and bluish colors of our blood samples. After converting images to YCbCr space, the extracted Cb and Cr coefficients are used for cell segmentation process. Sample input image before and after conversion is shown in fig. 2.

2. Cell Segmentation: The purpose of this stage is to segment the whole cell from the relative background as shown in Fig.3. The extracted Cb and Cr coefficients from our training images during preprocessing stage are now used to build a Gaussian Distribution [22] as shown in equations 1, 2, 3, 4.

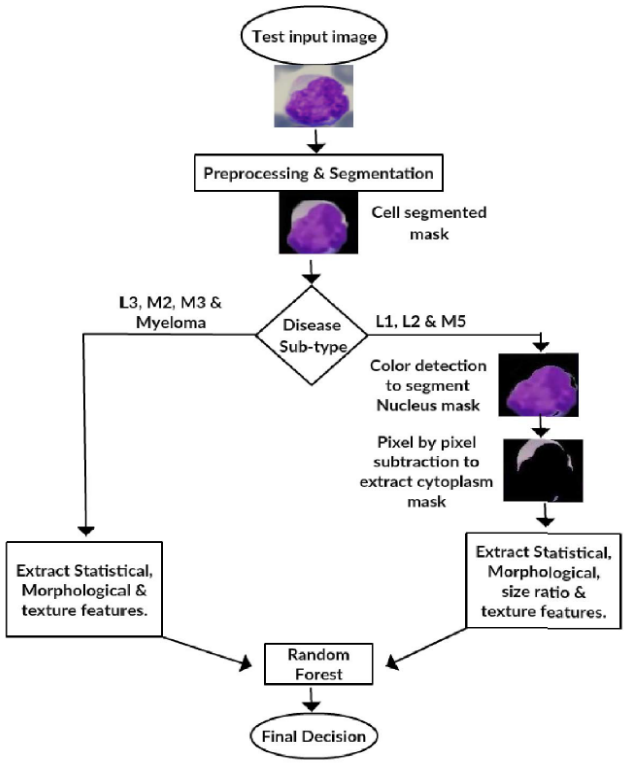
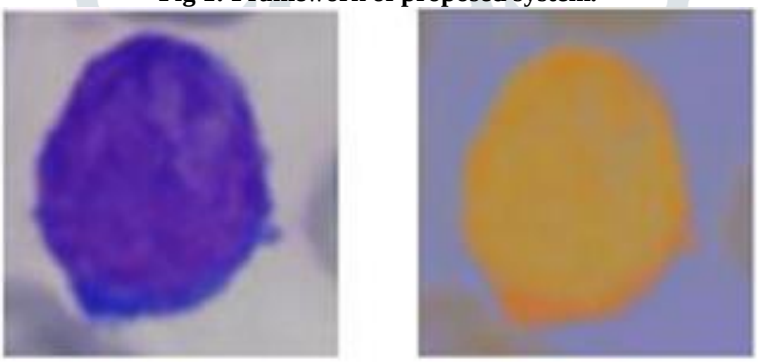


Fig 1: Framework of proposed system.



(a) **(b)**

Fig. 2: Cell input images of AML-M2 before and after Preprocessing (a) RGB Color Space, (b) YCBCR Color Space

$$bmean = mean(cb) \quad (1)$$

Where cb is the row vector containing all Cb coefficients obtained from our training images, and bmean is the blue mean of this vector.

$$rmean = mean(cr) \quad (2)$$

Where cr is the row vector containing all Cb coefficients obtained from our training images, and rmean is the red mean of this vector.

$$brcov = cov(cb, cr) \quad (3)$$

Where brcov is the co-variance of the two row vectors cb and cr. The result is a 2x2 matrix.

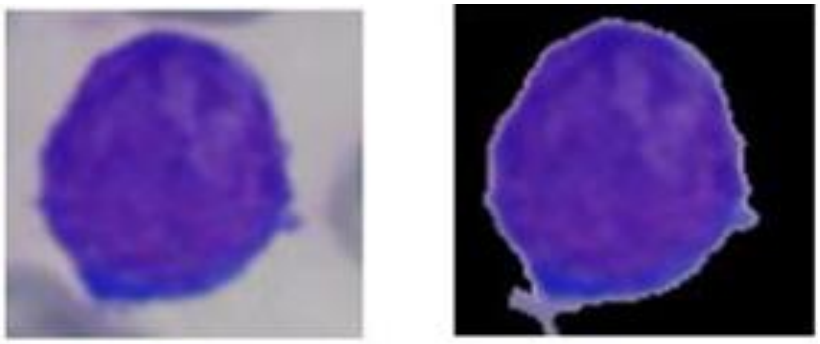
$$magCov = (brcov(1, 1) * brcov(2, 2) - brcov(2, 1) * brcov(1, 2)) \quad (4)$$

Where magCov is the magnitude of the brcov. In the testing phase the Guassian distribution is applied on the input test image to accomplish the segmentation stage as shown in equations 5, 6.

$$x = [(cb - bmean), (cr - rmean)] \quad (5)$$

$$f(x) = \frac{e^{-0.5 * x * brcov^{-1} * x'}}{2 * \pi * magCov} \quad (6)$$

This Gaussian distribution is applied to test images in the YCbCr space to extract our valuable pixels that are most probable included to our regions of interest ROI. After applying our defined distribution, normalization is applied followed by adaptive threshold algorithm [23].



(a) (b)

Fig. 3: (a) Cell original RGB image of AML-M2 before segmentation, (b) Cell (a) after Segmentation.

3. Nucleus & Cytoplasm Segmentation:

The outcome from cell division is a cover containing just the cell. Shading identification system [24] is connected on the cell cover with indicated scope of hues to fragment core veil [25]. By straightforward pixel to pixel subtraction of these two covers we can without much of a stretch concentrate an exact veil for the cytoplasm.

B. Feature Extraction

This stage is connected to fragmented images due to the pre-processing and division process. The separation between numerous types of ALL, AML and Myeloma requires distinctive visual similarities to be repaid. We figured morphological, factual, measurement proportions [26] and surface highlights [27] in our proposed methodology. As the running methodology indicates, it determines the distinctive arrangement of the highlights. The primary methodology relates to subcategories L1, L2, M5, while the alternative methodology relates to L3, M2, M3 and Myeloma. One of the two methodologies will be pursued in view of the choice of specialists. We considered proportional highlights because they are invariant to scaling as quickly examined in the accompanying subsections.

1. Morphological Features: This highlights speak to state of the cell and its measurement [26]. The determined highlights are region to border proportion, circularity, stretching, major to minor hub length proportion, degree and robustness.

Area to perimeter ratio: This is the ratio between the real number of pixels in the region of interest (ROI) and the separation between each pair of pixels on the outskirts of the ROI.

$$\text{AreaToPerimeterRatio} = \frac{\text{Area}}{\text{Perimeter}} \tag{7}$$

Circularity: This feature measures the complexity of the perimeter of the circular object.

Elongation: It is the ratio between length of the smallest rectangle containing the ROI and width of the smallest rectangle containing the ROI. It is also known as the growth in one direction of the ROI.

$$\text{Elongation} = \frac{\text{LSR}}{\text{WSR}} \tag{8}$$

Where, LSR is the length of the smallest rectangle containing the ROI and WSR is the width of the smallest rectangle containing the ROI.

Major to minor axis length ratio: It is the ratio between the major axis of the ellipse containing the ROI and the minor axis of the ellipse containing the ROI.

$$\text{MajorToMinorAxisLengthRatio} = \frac{\text{MajorAxisLength}}{\text{MinorAxisLength}} \tag{9}$$

Extent: It is the proportion of ROI area to the area of its bounding rectangle.

$$\text{Extent} = \frac{\text{Area}}{(\text{Width} * \text{Length})} \tag{10}$$

Solidity: It is the proportion of ROI area to area of its convex hull.

$$\text{Solidity} = \frac{\text{Area}}{\text{Convexarea}} \tag{11}$$

2. Statistical Features: These features also concern cell shape information but from different perspective [26]. The calculated features are the following

Mode: It is defined as most frequent value of the pixels intensity of the ROI.

Mean: It is the average value of the pixels intensity of the ROI.

Standard deviation: Standard deviation is a value represent how much pixels intensity differ from the mean of pixels intensities of the ROI.

Variance: Variance value of the pixels intensity of the ROI.

Sum: Sum of the pixels intensities of the ROI.
Gradient: Angles' gradient is calculated by Canny edge detection.

3. Size ratio features: Additional set of features are extracted based on the Segmentation of nucleus and cytoplasm as introduced in [26].
Nucleus cytoplasm area: It is the ratio of the area between the nucleus and the cytoplasm.

$$\text{NucleusToCytoplasmArea} = \frac{\text{NucleusArea}}{\text{CytoplasmArea}} \quad (12)$$

Nucleus cell area: It is the ratio of the area between the nucleus and the cell.

$$\text{NucleusToCellArea} = \frac{\text{NucleusArea}}{\text{CellArea}} \quad (13)$$

Nucleus cell perimeter: It is the ratio of the perimeter between the nucleus and the cell.

$$\text{NucleusToCellPerimeter} = \frac{\text{NucleusP erimeter}}{\text{CellPerimeter}} \quad (14)$$

4. Texture Features: These highlights relate to cell subtleties such as openings and granules. We performed the highlights of Haralicks[27]. There are many 14 surface highlights determined by the dark dimension co-event framework using 4 nearness headings. These highlights are accurate second minute, differentiate, connect, change, reverse distinguishing minute, total normal, total fluctuation, aggregate entropy, entropy, contrast entropy, contrast difference, ratio of relationship 1, measurement of relationship 2 and the highest coefficient of connection.

C. Classification

Random forest algorithm [28], [29] is a supervised classification algorithm that constructs a forest with several decision trees. Highest accuracy results are achieved with the higher number of trees. Random forest algorithm achieved successes in medical field [30] as its one of the most powerful algorithms that is widely used in different applications. It has many advantages as it can be used in different classification problems such as banking, stock market and E-commerce, it can be used for both classification and regression and it performs feature selection to only extracts the crucial features.

In our proposed method, Random Forest Classifier is used for the two main categories of the system. Random Forest classifier is the best classifier that is able to differentiate between different types and the one which gives us the highest accuracy as stated in our experiments table I and II. Also, the architecture that this classifier is based on fits our problem as we have three parent disease classes including ALL, AML and Myeloma and each one has many sub-classes as their subtypes.

IV. CONCLUSION

In this paper, we propose the structure, improvement and evaluation of a computerized framework for the accurate identification of malignant growth diseases in white platelets. It recognizes leukemia (ALL and AML) and myeloma types and sub-types. It is a recognition framework linked to small blood pictures obtained at this point that carries out preprocessing, division, extraction and characterisation. The proposed arrangement changes over pictures to the shading space of YCBCR and builds 53 CB and CR estimates of Gaussian transport. In order to prepare the classifier, factual, surface, estimate proportion and morphological highlights are processed. In contrast to existing frameworks, our framework has the ability to obtain classified Miss tests to improve the frame's future accuracy. Irregular forest classification is the best classification that can distinguish between different types and the one that gives us the best accuracy. In identifying and grouping types and subtypes, the framework achieved 94.3 percent accuracy. As our next stage, we want to identify more types of malignant growth disorders in white platelets to create a general framework for diseases of white platelets.

V. EXPERIMENTAL RESULTS

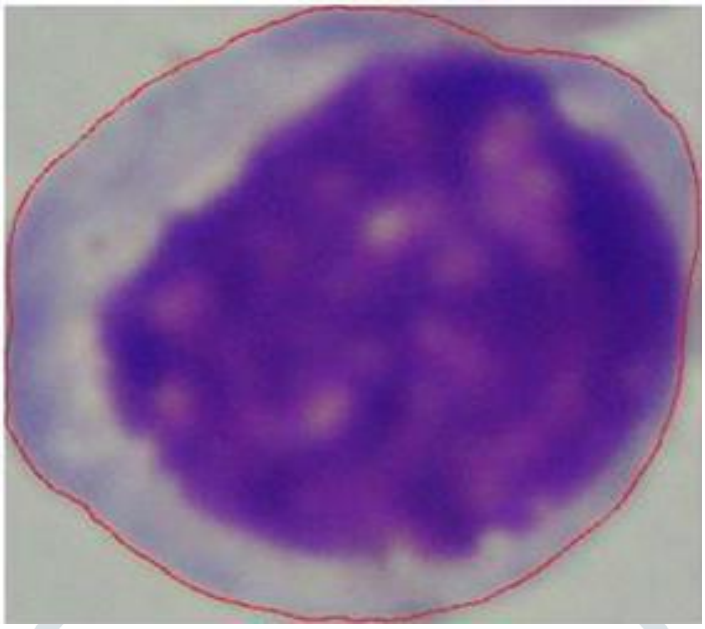


Fig 4: Input colour image.

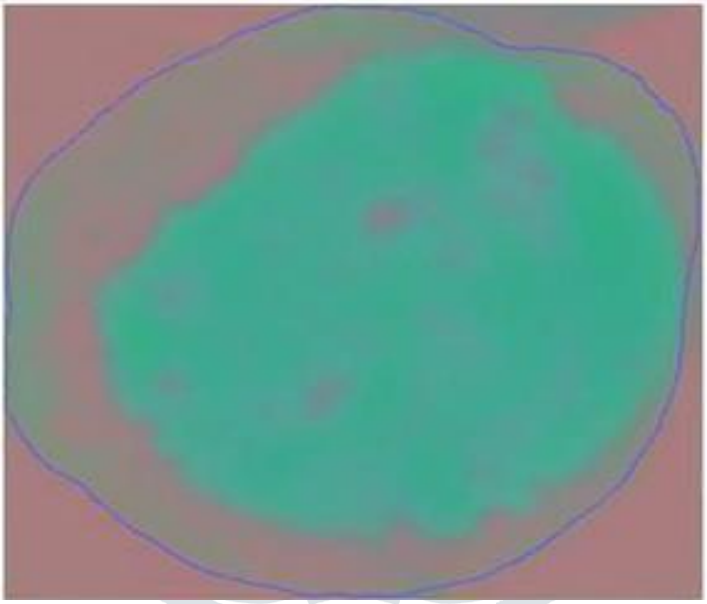


Fig 5: RGB-2-YCBCR Converted image.

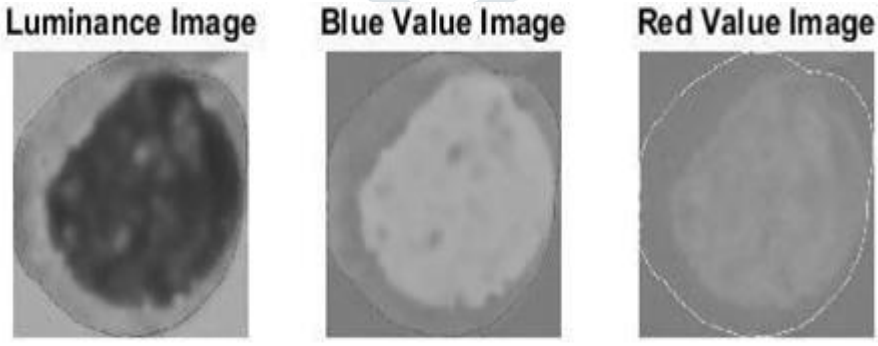


Fig 6: Luminance Image, Blue value image, Red value image.

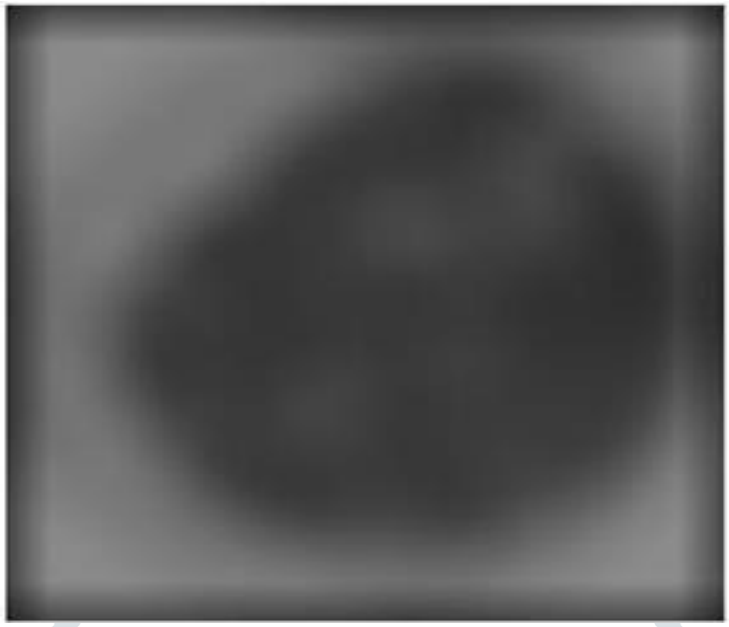


Fig 7: Filtered image.

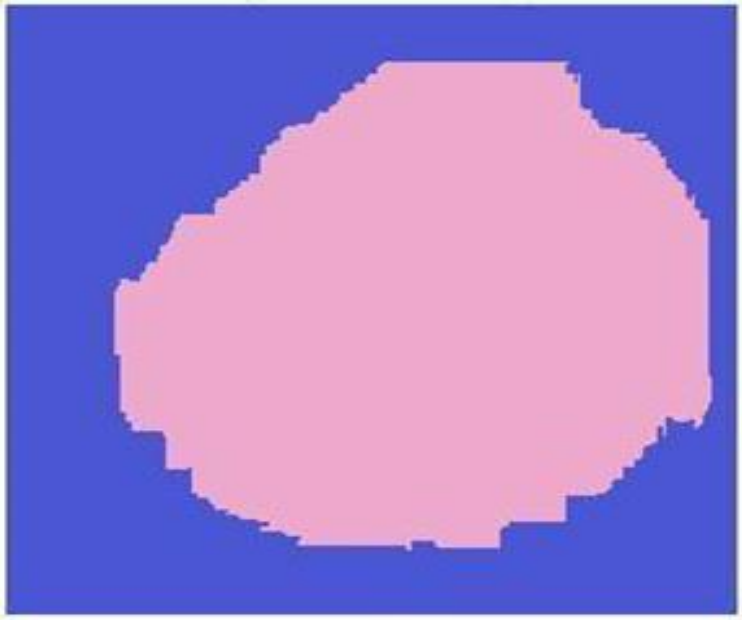


Fig 8: Segmented Image.

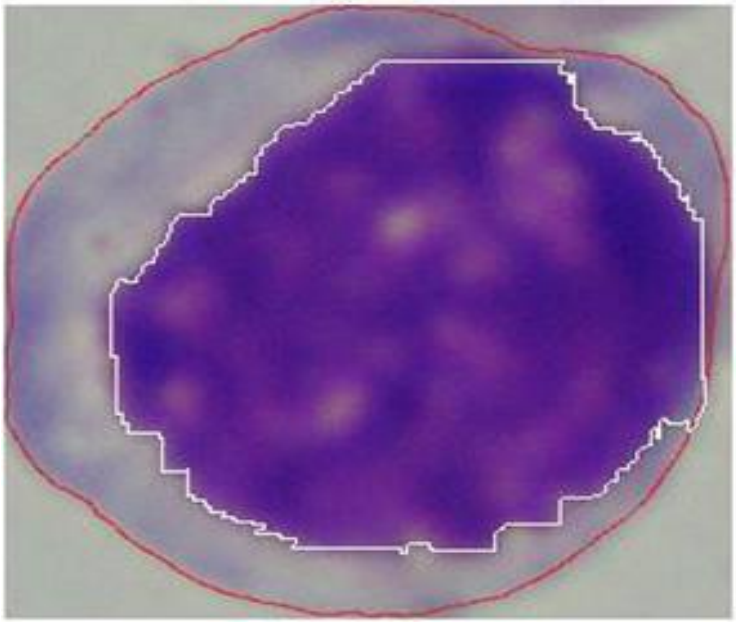


Fig 9: Output Segmented image.

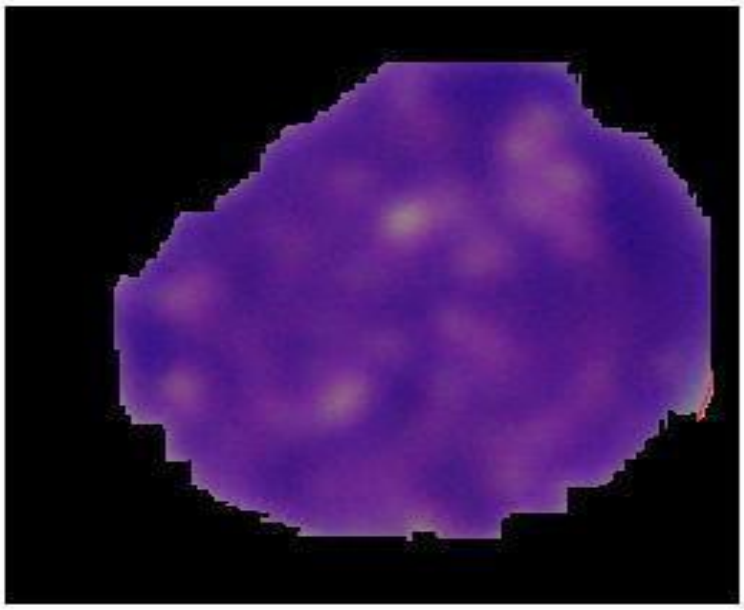


Fig 10: Output Segmented image2.

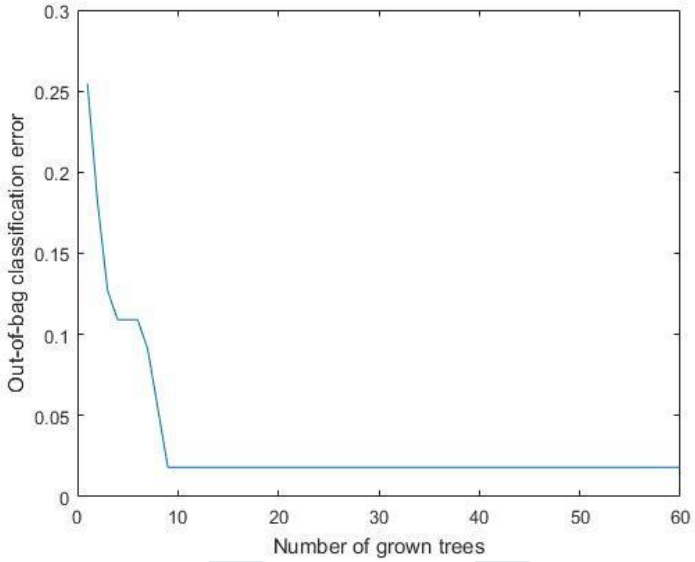


Fig 11: Classification of errors for Segmented image.

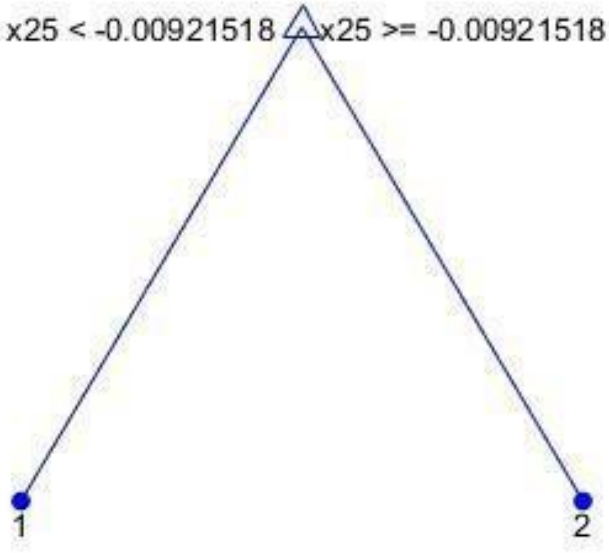


Fig 12: Decision tree.

```
Decision tree for classification
1  if x25<-0.00921518 then node 2 elseif x25>=-0.00921518 then node 3
2  class = 1
3  class = 2

-----------------------------------------------
ANALYSIS - CLASSIFIERS
Sensitivity:                96.9697%
Specificity:                100%
Accuracy:                   98.1818%
>>
```

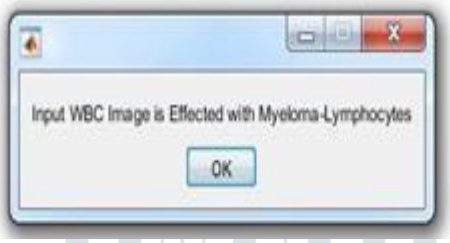


Fig 13: Analysis classifiers.

VI. REFERENCES

[1] K. M. Sensing, Cancer. URL http://www.who.int/mediacentre/factsheets/fs297/en.html
[2] J. B. Henry, J. P. AuBuchon, Clinical diagnosis and management by laboratory methods, Archives of Pathology and Laboratory Medicine 121 (9) (1997) 1016.
[3] E. Sarhan, E. Khalifa, A. M. Nabil, Post classification using cellular automata for landsat images in developing countries, in: International Conference on Image Information Processing (ICIIP), Shimla, India, IEEE, 2011, pp. 1–4.
[4] A. AbdelRaouf, C. A. Higgins, T. Pridmore, M. I. Khalil, Arabic character recognition using a haar cascade classifier approach (hcc), Pattern Analysis and Applications 19 (2) (2016) 411–426.
[5] A. AbdelRaouf, C. A. Higgins, T. P. Pridmore, M. I. Khalil, Arabic corpus enhancement using a new lexicon/stemming algorithm., in: 2nd International Conference on Pattern Recognition Applications and methods (ICPRAM), Barcelona, Spain, 2013, pp. 435–440.
[6] T. M. Ghanem, M. N. Moustafa, H. I. Shahein, Gabor wavelet based automatic coin classsification, in: 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, IEEE, 2009, pp. 305–308.
[7] S. Sameh, M. A. Azim, A. AbdelRaouf, Narrowed coronary artery detection and classification using angiographic scans, in: 12th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, IEEE, 2017, pp. 73–79.
[8] S. Moataz, M. A. Azim, A. AbdelRaouf, Automated fish signals fusion detection for chronic myeloid leukemia diagnosis, in: 26th International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt, IEEE, 2016.
[9] M. F. for Medical Education, Research, Leukemia. URL https://www.mayoclinic.org/diseases-conditions /leukemia/symptoms-causes/syc-20374373.html
[10] What is multiple myeloma? URL https://www.cancer.org/cancer/multiple-myeloma/about/what-is-multiple-myeloma.html
[11] T. Karthikeyan, N. Poornima, Microscopic image segmentation using fuzzy c means for leukemia diagnosis, Leukemia 4 (1).
[12] J. Yen, R. Langari, Fuzzy logic: intelligence, control, and information, Vol. 1, Prentice Hall Upper Saddle River, NJ, 1999.
[13] S. Mohapatra, S. S. Samanta, D. Patra, S. Satpathi, Fuzzy based blood image segmentation for automated leukemia detection, in: Devices and Communications (ICDeCom), 2011 International Conference on, IEEE, 2011, pp. 1–5.
[14] D. Graves, W. Pedrycz, Fuzzy c-means, gustafson-kesselfcm, and kernel-based fcm: A comparative study, in: Analysis and Design of Intelligent Systems using Soft Computing Techniques, Springer, 2007, pp. 140–149.
[15] K. M. Sensing, Identifying color differences using l*a*b* or l*c*h*coordinates. URL https://sensing.konicaminolta.us/blog/identifying-color-differences-using-l-a-b-or-l-c-h-coordinates/
[16] S. Agaian, M. Madhukar, A. T. Chronopoulos, Automated screening system for acute myelogenous leukemia detection in blood microscopic images, IEEE Systems journal 8 (3) (2014) 995–1004.
[17] R. D. Labati, V. Piuri, F. Scotti, All-idb: The acute lymphoblastic leukemia image database for image processing, in: Image processing (ICIP), 2011 18th IEEE international conference on, IEEE, 2011, pp. 2045–2048.
[18] R. Bhattacharjee, L. M. Saini, Detection of acute lymphoblastic leukemia using watershed transformation technique, in: Signal Processing, Computing and Control (ISPCC), 2015 International Conference on, IEEE, 2015, pp. 383–386.
[19] O. Sarrafzadeh, H. Rabbani, A. M. Dehnavi, A. Talebi, Detecting different sub-types of acute myelogenous leukemia using dictionary learning and sparse representation, in: Image Processing (ICIP), 2015 IEEE International Conference on, IEEE, 2015, pp. 3339–3343.
[20] Medical image and signal processing research center. URL http://misp.mui.ac.ir/en

[21] T. Acharya, P.-S.Tsai, JPEG2000 standard for image compression concepts, algorithms and VLSI architectures, Wiley-Interscience, 2005. [22] K. Krishnamoorthy, Handbook of statistical distributions with applications, CRC Press, 2016.
[23] A. Elmoataz, Image and signal processing 5th International Conference, ICISP 2012, Agadir, Morocco, June 28-30, 2012. Proceedings, Springer, 2012.
[24] R. Lukac, K. Plataniotis, Color Image Processing: Methods and Applications, Image Processing Series, CRC Press, 2006.
[25] E. A. Mohammed, B. H. Far, C. Naugler, M. M. Mohamed, Application of support vector machine and k-means clustering algorithms for robust chronic lymphocytic leukemia color cell segmentation, in: e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on, IEEE, 2013, pp. 622–626.
[26] C. Reta, L. Altamirano, J. A. Gonzalez, R. Diaz-Hernandez, H. Peregrina, I. Olmos, J. E. Alonso, R. Lobato, Segmentation and classification of bone marrow cells images using contextual information for medical diagnosis of acute leukemias, PloS one 10 (6) (2015) e0130805.
[27] R. M. Haralick, Statistical and structural approaches to texture, Proceedings of the IEEE 67 (5) (1979) 786–804.
[28] A. Liaw, M. Wiener, Classification and regression by random forest, R News 2 (3) (2002) 18–22.
[29] Y. Pavlov, Random Forests, VSP, 2000.
[30] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F. A. Hamprecht, A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC bioinformatics 10 (1) (2009) 213.
[31] G. Sites, Bone marrow microscopic data - drhosseinrabbani. URL https://sites.google.com/site/hosseinrabbani khorasgani/datasets-1.
[32] I. Steinwart, A. Christmann, Support Vector Machines, Information Science and Statistics, Springer New York, 2008.
[33] M. Kirk, Thoughtful Machine Learning: A Test-Driven Approach, O'Reilly Media, 2014.
[34] D. Kleinbaum, M. Klein, Logistic Regression: A Self-Learning Text, Statistics for Biology and Health, Springer New York, 2010.

Author's Profile:

Ms. Alina Ahmed has completed herB.Tech(ECE) from Shadan College Of Engineering and Technology, Moinabad, RR District.JNTU Hyderabad. Presently, She is pursuing her Masters in Embedded System from Nawab Shah Alam Khan College Of Engineering and Technology, Hyderabad, JNTU Hyderabad, TS. India.

Ms. Ayesha Fatima has completed B.Tech (ECE) from Vignan Institute of technology and science, JNTUH University, Hyderabad and M.Tech (Embedded Systems) from Royal Institute Of Technology and Science JNTU University, Hyderabad. Currently she is working as an Assistant Professor of ECE Department in Nawab Shah Alam Khan College Of Engineering and Technology, Hyderabad, TS. India.