



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 11, November 2025

Blood Cancer Detection using Deep Learning

Dhivya.M, A.Nooral Shaba, N.Sivakumar

M.E-CSE, M.P.Nachimuthu M.Jaganathan Engineering College, Chennimalai, Tamil Nadu, India

Assistant Professor, Department of CSE, M.P.Nachimuthu M.Jaganathan Engineering College, Chennimalai,
Tamil Nadu, India

Associate Professor, Department of CSE, M.P.Nachimuthu M.Jaganathan Engineering College, Chennimalai,
Tamil Nadu, India

ABSTRACT: Blood cancer, particularly leukemia, poses a significant global health challenge, requiring early and accurate diagnosis for effective treatment and improved patient outcomes. Traditional diagnostic methods, primarily manual microscopic analysis of blood smears, are time-consuming, labor-intensive, and prone to human error due to the subtle morphological variations in abnormal cells. To address these limitations, this study proposes an automated and efficient system for blood cancer detection utilizing advanced deep learning (DL) techniques. The proposed methodology leverages Convolutional Neural Networks (CNNs) to analyze microscopic images of blood cells. The system architecture involves several key stages: image acquisition, preprocessing (including normalization and data augmentation to enhance model generalization), automated feature extraction, and final classification into benign or malignant categories. We evaluated the performance of a custom CNN model on publicly available datasets (e.g., ALL-IDB) and compared its efficacy against other state-of-the-art architectures like ResNet and MobileNetV2. Experimental results demonstrate that the deeplearning approach achieved high diagnostic accuracy, with our model attaining an accuracy of over 98% in differentiating cancerous from healthy cells, outperforming conventional machine learning methods and existing diagnostic approaches. This system offers a reliable, fast, and objective diagnostic aid that can significantly reduce the workload on healthcare professionals and facilitate timely intervention, particularly in resource-constrained settings. Future work will focus on improving model interpretability, clinical validation with more diverse datasets, and optimizing computational efficiency for real-world deployment.

I. INTRODUCTION

Blood cells are generated from the stem cell which is present in the bone marrow. Generally, Blood is made up of 55% of plasma and 45% components of red blood cells (which carries oxygen and brings back carbon dioxide away from tissues), WBC (which fights against infection) and blood platelets (prevents clotting of blood and controls bleeding). In the Human body system, the cells grow and multiply afterwards the old cells get destroyed and the new cell takes its place. But in cancer the old cells do not die, so the new cells do not get enough places to live which affect the functioning of the blood marrow which results in the increasing number of damaged white blood cells that are immature cells that destroy other cells and affect the body's ability to fight against the infection. These immature cells are known as leukemia cells.

Leukemia takes place owing to the absence of normal blood cells and it is identified by blood tests and blood marrow biopsy. In major, it is classified into four types. They are ALL, AML, CLL and CML. Leukemia belongs to a large set of tumors which infects the blood, bone marrow and lymphoid system. Here the main purpose is to detect the abnormal cells present in the blood using different image processing techniques. Input images are taken from the online database and pre-processing is done to remove unwanted noises and it will also increase the intensity of the image to get accurate output. After pre-processing, segmentation is done to segment the cells present in the image. Watershed segmentation is used to segment the overlapped cells present in the image. It is used to segment the combined images or objects touching each other and gives accurate results. After that K-means clustering is done. It helps to separate the background image and the specific area in the image. It is used to find the abnormal cells or blast cells present in the image. By changing the K value, we can get the accurate result. Feature extraction reduction process in which the original image is taken as raw data and it is divided into many valuable data forms. Here, SVM and KNN classifier is used and calculated the accurate result. The main objective of our paper is to increase the accuracy.

II. LITERATURE SURVEY

Melanie MacGregor., et al 2021, in this paper they detected bladder cancer cells using a device. This follows the principle of photodynamic diagnostic (PDD). With the help of a micro fluidic-based platform with a whole cell of immuno-capture technology to detect the bladder cancer cells in patients urine. The design and working principle of this diagnostic approach are collection of sample urine stored in a micro fluid chamber with anti- EpCAM antibodies, a membrane with biomarker specific for the cancer cells[1]. Then Hexaminolevulinate hydrochloride(HAL) optimization and finally detection and diagnosis of bladder cancer cells. This diagnostic device design can be done by a coating of polyoxazoline-based plasma polymer. In the initial test, all the cancer cells were detected and the specificity result was 80%. Future work involves the enhancement of devices to find the low grade cancer cells also.

Milan Tripathi., et al 2021, in this project images are classified using convolutional neural networks. This project includes computer vision and deep learning. Inputs are collected from the computer and stored in it. Then the images are subjected to image processing [2]. Then computer vision and deep learning are processed to get the final result. Classifiers are used to classify the image set obtained from the computer as input. Major elements used in this operation are input images from the computer, feature detector and feature map. Feature map is obtained by feature detector from input images. The matrix representation of input images is created and multiplied element wise using feature detectors to get an accurate feature map. Adam optimizer is used to change the neural network characteristics. Based on DenseNet model, the result of training and testing accuracy is 99.25% and 100% respectively.

Mugunthan., et al. 2021, sigmoidal function is improved using the Extreme Learning Machine(ELM) domain. In general, most of the basic regression problems are solved using least square minimization [3]. But in this, ELM is used for its less training time of weighted vectors. The input weights are selected randomly by a singular column rank matrix to calculate the training procedure with an orthogonal projection for better stability and speed of the process. This work can be continued in future with the help of IoT devices.

Dr.Samuel Manoharan .,el al 2020, in this paper, the author has detected the lung cancer using neural network classification. It is mainly used to reduce the false positive identification including enhancement filter [4]. It is used to detect the early stage of cancer by calculating the probability of malignant and segmented lung nodule. Input image is taken and pre processing is done which is erosion and dilation of the image. The input image is enhanced and unwanted noises are removed from the image and it is converted to gray scale. After that superimpose and distance transform is done for the input image. Gradient, Superimpose and marker control is done for the input image. After that segmentation is also done with graph cut, Level set and region splitting segmentation etc., In Graph set; original segmented image can be detected. In level set segmentation, the region and shapes are segmented. In region splitting segmentation, fixed parameters are not used instead of that dynamic methods are used. Here they have combined thresholding and region growing is used to extract the cavity region with high accuracy. Then, they have combined the thresholding and morphological algorithm, thresholding along with ball algorithm is used to segment the thorax area of the input. Proposed algorithm is used to extract the nodules from the input image; segmentation is also done to get the accurate result and early detection of lung cancer. At last performance is calculated using Baye's theorem.

Astha Ratley., et al 2020, in this paper they selected some microscopic images as input and pre-processing is done. In this process, unwanted noises are removed with the help of filters like wiener and median filter. After the removal of noise thresholding takes place, here the certain range of pixels are considered as an image and remaining pixels are considered to be a noise[5]. Then feature extraction takes place in which the features of the images are extracted from the binary image. In morphological features, shape, area, perimeter and edge features are extracted. Textural features are mainly used to identify the particular area of interest in an image. It extracts the internal details of the input image like sharpness and smoothness. GLCM features come under texture features and some of the GLCM features are correlation, contrast and entropy of the pixels. For the classification of disease, machine learning algorithms are used and they are ANN, SVM and LDA. The classification accuracy is about 97%.

III. PROPOSED METHODS

The aim of the work is to compare the image processing techniques to get the accurate result for detection of leukemia with good accuracy. The exact count of cells determines the early detection of leukemia. In the field of medicine,

Complete Blood Count (CBC) plays a vital role. It determines the normal and sufficient or required blood count for humans. Microscopic images which are taken from the database which is used as an input image for this leukemia detection process. These images undergo pre-processing techniques to make the images free from noises. After pre-processing, thresholding is done to enhance the contrast of an image. The particular cells present in the input image are separated using a segmentation process. In feature extraction, many features are calculated to find the differences among infected and non-infected blood cells. Finally classification techniques like SVM and KNN are used to classify the type of cancer, if cancer is detected.

The overview of the work is given as a flow. First we have to get the input image. After that pre-processing is carried out this removes the unwanted noises present in the image. Image Segmentation and feature extraction is done. Finally, classification takes place which gives the accuracy, sensitivity, error rate and specificity with the help of confusion matrices.

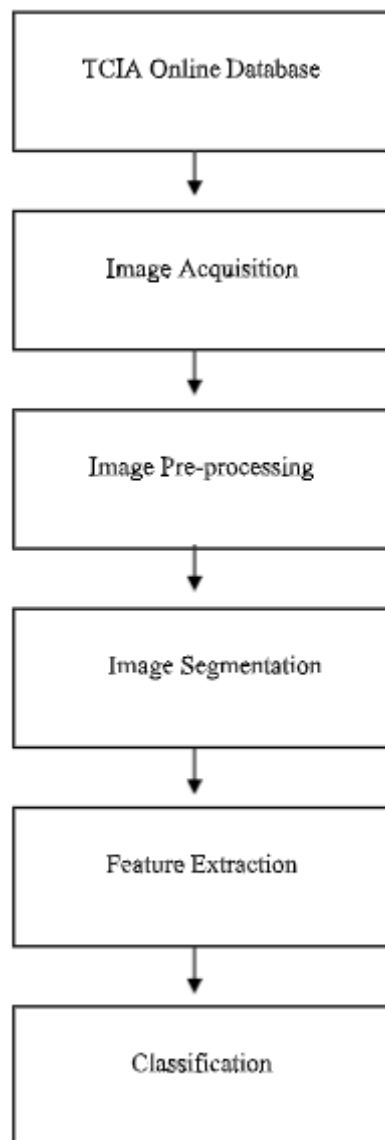


Figure.1 System Architecture

Image Acquisition

Blood cells are captured from a microscope. These microscopic images are taken as input images. In this work, the input images are taken from the TCIA online database. These images are separated as a training set and experiment set.

Pre-processing

Pre-processing is done to improve the quality of the image. It is also used to adjust the contrast of the image. The unwanted noises present in the input images are removed with the help of median filter. Histogram equalization is used to adjust the contrast of the image.

1. Median filter

Median filter is a powerful non-linear filter technique and it is effective in the presence of salt and pepper noise and impulse noise. This filter is used to protect the edges of the image during the removal of noise. It is used to remove the intensity variation from one pixel to nearby pixel. This filter is used to replace the centre pixel value with the median value to the nearby pixel. Here, MAX filter is good to remove the pepper noise and the MIN filter is good for salt noise.

2. Histogram equalization

The histogram represents the number of pixels which contains each and every value of color. In a grayscale image, it is basically represented by grey values. In RGB images, it is represented in every color component which is nothing but red, green and blue color. Basically, the histogram is represented between 0's and 1's to avoid dependency on the quantization level or number of pixels. Here, it is used to adjust the contrast of the image. This method helps to enhance and improve the dark background and contrast of the image.

C. Segmentation

Image segmentation is a process of partitioning the original image into many parts or a particular segment based on the characteristics of pixels on the image. It helps to obtain the uniform histogram for the input images.

1. Thresholding**i. Otsu's Thresholding**

This method is used in the clustering images. Here the histogram of the image is taken and this method is performed only when the histogram is bimodal. It usually reduces the within class variance and at the same time it increases the between class variance. Here, it is compared with probability of classes, means and compare sigma with maximum value.

Algorithm for Otsu's Thresholding:

1. Input image is taken and converted into gray scale image.
2. Calculate foreground and background variances for a single threshold.
3. Histogram is taken for the gray scale image.
4. Otsu's thresholding takes place only when the histogram is bimodal.

ii. HSV equalization Thresholding:

HSV Thresholding is done using Hue, Saturation and Value in the input image. It is very simple and easy to implement the RGB coloured image into HSV. Hue is used to identify the particular area with the colour. Saturation is used to verify the intensity of colour. Value is used to measure the brightness whether that particular pixel emits low or high light. In this method, it separates the background image, WBC cells and other cells present in the images using different colours. It is used to segregate the intensity of the image from the colour information. Most of the information of WBC cells is collected by H component.

iii. Edge detection

Edge detection is used to identify the edges of the objects within images. It is mostly used for image segmentation and extraction of information. Mostly in images like blood cells, the shape details of an image are enclosed in edges. Sobel filter is used to detect the edges in an image and then by enhancing those particular areas of image which will increase edges and sharpens the image. Then the image will become clearer.

2. Watershed Segmentation

This technique is used to segment the overlapped cells present in the image. It is used to segment the combined images or objects touching each other and gives accurate results. In each region, the variations in intensity have least gradient values. Watershed segmentation is not directly to the image. It is performed in the gradient of the image. Here marker controlled watershed segmentation is used and it is very flexible for segmentation of objects which is along with closed objects. It is used to control over segmentation based on markers. Internal markers are used to get the object and external markers are used to find the boundary of the object. At last, the boundaries are arranged on the desired ridges, so objects are separated from their neighbour objects.

3. K-means clustering

In clustering technique, the pixel and cluster center present in the image are differentiated based on the location, density and other factors like K-means clustering which is used to divide the images based on the distance. It helps to separate the background image and the specific area in the image. It is used to find the abnormal cells or blast cells present in the image. By changing the K value, we can get the accurate result. This clustering method is fast, simple and it is effective for large datasets.

Steps involved in K-Means algorithm:

1. Choose k number of clusters.
2. Select random K values.
3. Assign each and every data to the centroid to form a cluster.
4. Calculate and keep the new centroid of every cluster.
5. Assign each data to the new closest centroid. If any changes take place, go to step 4 and repeat the process, or else the function is completed to show the result.

D. Feature extraction

Feature extraction is a reduction process in which the original image is taken as raw data and it is divided into many valuable data forms. In other words, a large amount of data is divided into a large number of variables. Feature extraction is done after pre-processing and segmentation. In this work we find area, perimeter, Centroid, Bounding Box, Major Axis Length, Minor Axis Length as geometrical features. Statistical features like convex hull, convex image, convex area, circularity, Euler number, solidity and eccentricity. Textural features like color, filled image, filled area and extrema are found using functions.

E. Classification

1. SVM

In general, a Support Vector Machine is a linear classification. Mostly it is used for binary classification and it is also used for multiclass classification. A separate dataset is created for SVM classification and then it is classified. From the dataset, a hyper plane is created to divide the data into the same space of the same class. To train SVM, most relevant features are collected and tabulated for a new dataset. From the input dataset, a hyper plane is formed as a barrier for the rest of values to separate themselves as separate classes. Linear kernel is used to linearly separate the particular dataset. It is the most commonly used kernel because it is fast when compared with the other kernel. Finally in this work, the accuracy of 97% is achieved using SVM classifier. $k(x, x_i) = \sum(x * x_i)$ -----(1) Linear kernel is used only when the data is linearly separable. It is the most common kernel used and it is very faster kernel when compared with other. Here large number of features is used. In code, we have changed the iteration and x parameters to increase the accuracy.

2. K Nearest Neighbour

K-nearest neighbour algorithm comes under supervised machine learning. It analyses the function and it is also used to solve both regression problems and classification. It gives new data points accordingly to the k number or the nearest points. It divides the image based on the distance. There is a difference between the pixel and the cluster centre on the image based on the density or location. It is simple, fast and it is effective for large datasets. It is a lazy learner and performing classification or statistics by just memorizing its training data. It doesn't undergo any training for classification.

If $k = 1$, then it will be assigned to a class of that single nearest neighbour. Finally in this work, the accuracy of 91% is achieved using KNN classifier.

These are the steps followed in the KNN algorithm.

1. Select the K value.
2. To calculate the distance we use the Euclidean distance formula.
 $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
3. Take the K nearest neighbours as per the calculated Euclidean distance.
4. Among these k neighbours, count the number of the data points
5. Assign the new data points to that category for which the number of the neighbour maximum and the function is ready.

IV. RESULTS AND DISCUSSION

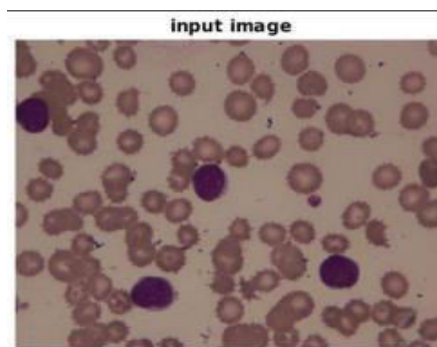


Figure.2 Input Image

The input is taken from the TCIA online database. In that database, we have chosen one image as input for this process.

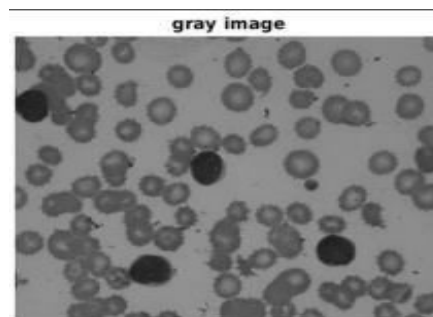


Figure.3 Gray image

The input image is converted into gray image. Gray scale image consists of different shades of grey colour. RGB image is converted into gray scale image. The intensity values of the input image are converted to 8 bit integer in 256 different ways of gray shades.

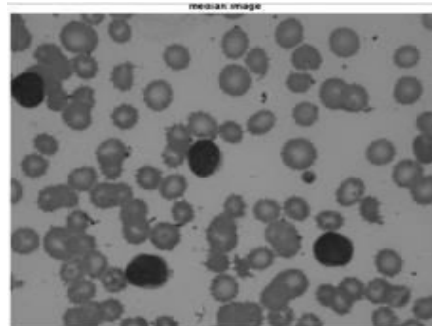


Figure.4 Median image

In the median filter, if the window size increases the edges of the image get blurred and also have the ability to suppress the noise.

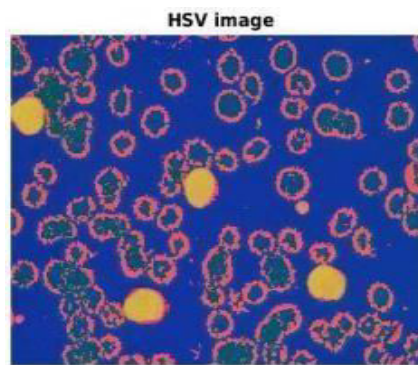


Figure.5 Histogram equalization

Histogram equalization is used to increase the contrast of the given image. In this image, it enhances and improves the dark background and contrast of the image.

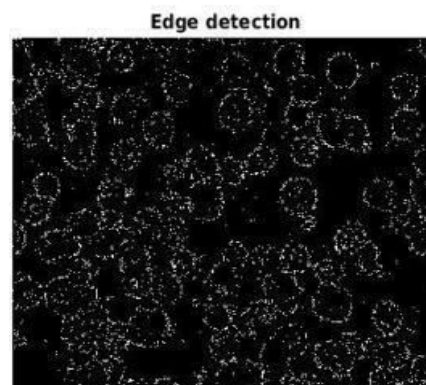


Figure.6 Edge detection

Edge detection is used to show the boundaries of the cells in the input image which are detected. Sobel filter is used to detect the edges in an image and then by enhancing those particular areas of image which will increase edges and sharpens the image. Then the image will become clearer.

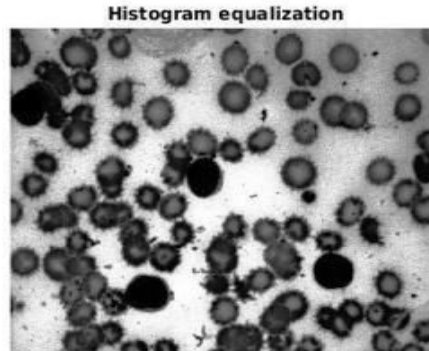


Fig.7 HSV thresholding image

In HSV thresholding, it separates the background image, WBC cells and other cells present in the images using different colours. It is used to segregate the intensity of the image from the colour information.

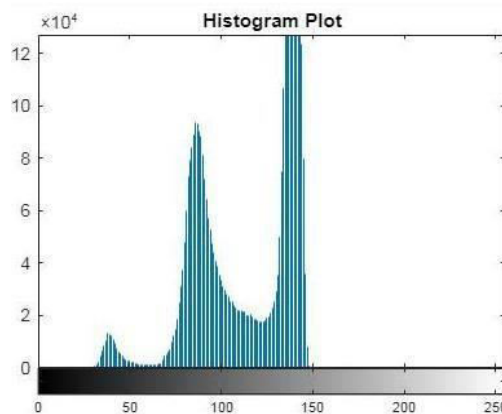


Figure.8 Histogram plot

The above figure represents the histogram for the input image. It is bimodal, so it works perfect for the Otsu's thresholding.

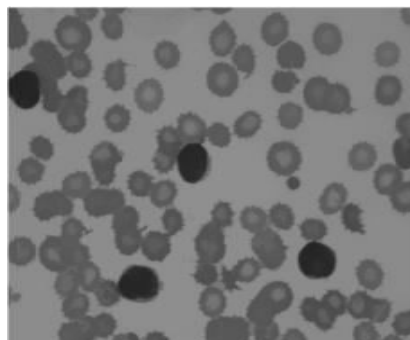


Figure.9 Otsu's thresholding image

Otsu's thresholding gives a single intensity value to separate the pixel into foreground or background. It iterates through all possible threshold values in the foreground as well as in the background in finding the threshold value where the sum of foreground and background is less. It works well if the histogram is bimodal.

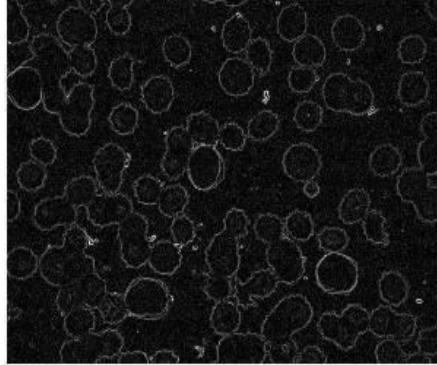


Figure.10 Gradient magnitude

Gradient magnitude is used to find the continuous edges present in the image. It changes the direction of image intensity. This process is repeated in every position of the image which will detect the edges present in the image.

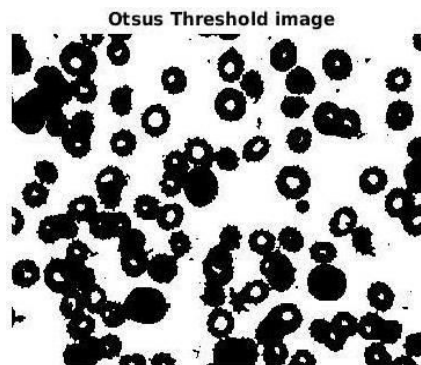


Figure.11 opening image

Opening images are found using erosion and dilation, which is used to find the foreground images clearly.

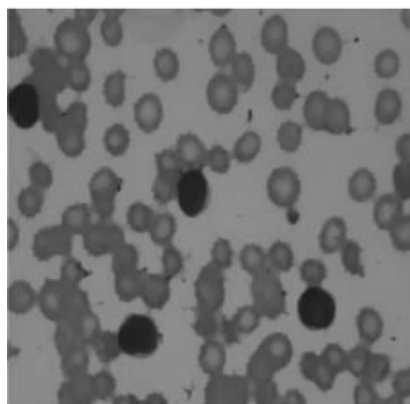


Figure.12 Closing image

Closing images are found using erosion and dilation, which is used to find the foreground images clearly.

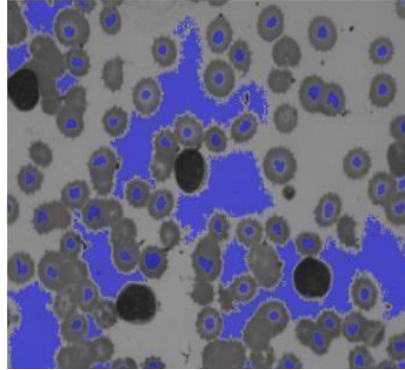


Fig.13 Reginal maxima superimposed on original image

In these reginal maxima, the cells shadows are not marked. The foreground cell images are marked and shown.

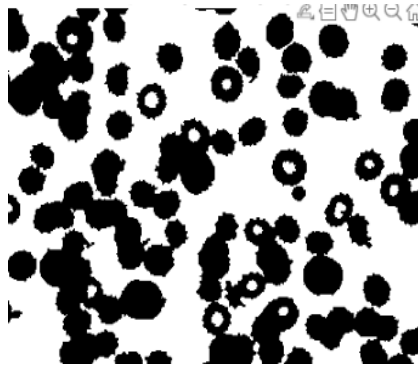


Figure.14 Threshold image

In threshold image, the cells are shown in black colour and the background is represented in white colour.

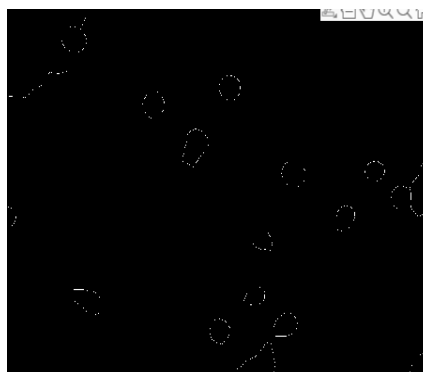


Figure.15 Watershed Ridge Lines

In watershed ridge, the edges are separated by lines. The edges of the abnormal cell are shown. So it is very easy to detect the cells present in the image.

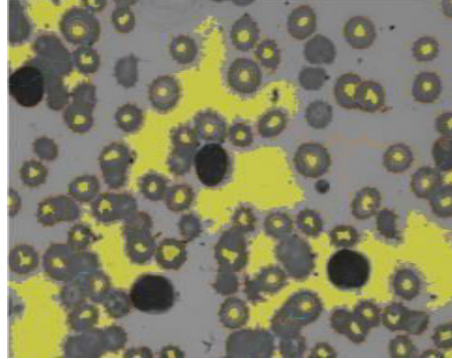


Figure.16 Object Boundaries Superimposed on Original Image

In this image, the original image is superimposed in the marker to differentiate the foreground, background and to identify the segmented part.

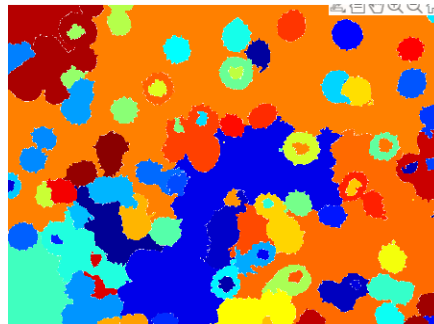


Figure.17 Coloured Watershed Label Matrix

Here, Coloured matrix is used to differentiate the cells in the input image.

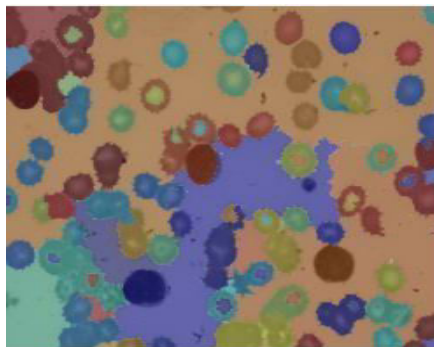


Figure.18 Coloured Labels Superimposed

Finally, the overlapped cells are separated from the image using watershed segmentation.

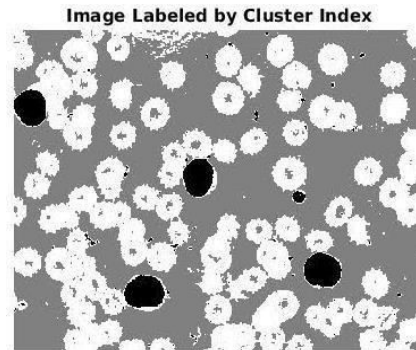


Fig.19. Image labelled cluster index

In the above image, the clusters of cell are formed using a distance which is used to separate the cells as two clusters. The distance calculated from the centre of the clusters. Here, K value is 3.

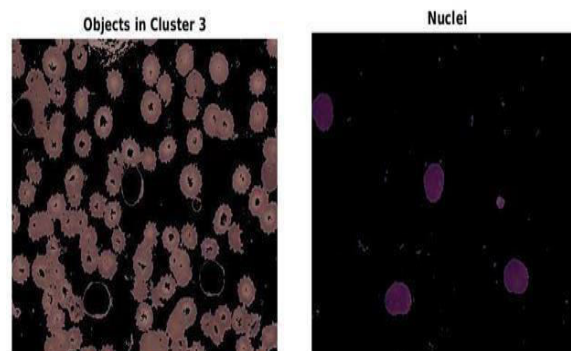


Figure.20 Objects of cluster 1 and 2

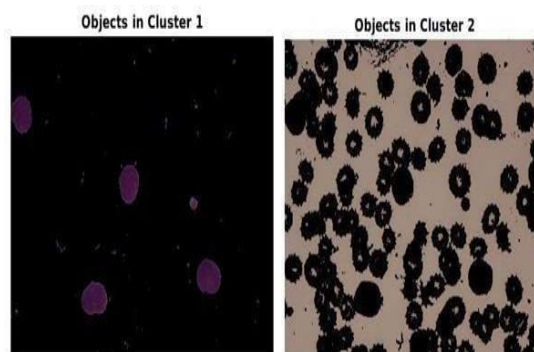


Figure.21 Objects of cluster 3 and extracted nuclei

The distance is calculated from the centre of clusters. Depending on distance from the clusters we have to reassign it to the nearby clusters. The cluster present in the images is shown and the nuclei are separated with the help of K means clustering.

FEATURES OF ALL BLOOD SMEAR					
Cell Areas	100	170	208	224	228
Cell Centroid	0.1790	0.6994	0.8657	1.2744	1.4859
Cell Perimeter	69.7420	76.7390	100.7720	117.8680	227.2750
Cell Major Axis Length	26.9182	31.1668	38.0289	44.4887	58.0322
Cell Minor Axis Length	11.7701	14.3850	16.8971	16.9449	18.8692
Cell Eccentricity	0.7309	0.7833	0.8421	0.8997	0.9502
Cell Bounding Box	0.0400	0.0420	0.5065	0.6895	1.5285
Cell Orientation	57.4022	-89.9144	-4.7495	8.7161	40.8836
Cell Convex Area	108	188	240	312	633
Cell Euler Number	100	170	208	224	228
Cell Extrema	0.1790	0.6994	0.8657	1.2744	1.4859
Cell Equiv Diameter	69.7420	76.7390	100.7720	117.8680	227.2750
Cell Solidity	26.9182	31.1668	38.0289	44.4887	58.0322
Cell Circularity	11.7701	14.3850	16.8971	16.9449	18.8692
Cell Filled Area	0.7309	0.7833	0.8421	0.8997	0.9502
Cell Extent	0.0400	0.0420	0.5065	0.6895	1.5285

Table.1 Feature extraction

The above table consists of features of the cell. Feature extraction is a reduction process in which the original image is taken as raw data and it is divided into many valuable data forms. Feature extraction is done after pre-processing and segmentation process which is used extract the exact features present in the cell. With the help of feature extraction, we can differentiate the normal and abnormal cells by the features of the cell.

CLASSIFICATION	ACCURACY
Linear Kernel	97%
KNN	91%

Table.2 Accuracy for classification

The accuracy of the classifiers is shown above by simulating the python code.

	LINEAR KERNEL		KNN	
	Disease Positive	Disease Negative	Disease Positive	Disease Negative
Test Positive	20	1	20	1
Test Negative	0	14	2	12

Table.3. Confusion matrix

This is the confusion matrix we got by simulating the code. The accuracy, specificity, F1 score, sensitivity and error rate for this confusion matrix were calculated by using the following formulas manually.

$$\text{Accuracy} = \frac{TN + TP}{TP + FN + FP + TN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

$$\text{Error rate} = \frac{FN + FP}{TP + FN + FP + TN} \quad (4)$$

$$F1 = \frac{2TP}{2TP + FN + FP} \quad (5)$$

Here, TP = amount of true positives

TN = amount of true negatives

FP = number of False positive

FN = number of False negative

Classification	Accuracy	Sensitivity	Specificity	F1	Error rate
Linear kernel	97.14	100	93.33	97.56	2.85
KNN	91.42	90.90	92.30	93.02	8.5

Table.4 Performance measures

Accuracy defines the correct output of the model prediction. It determines the accuracy of the classification problems. Error rate is also called as misclassification rate. It is used to define the model gives wrong predictions. Sensitivity is used to calculate the number of correct positive predictions divided by total number of positive. F score is used to evaluate the sensitivity and precision at a time.

V. CONCLUSION

In this paper, the blood smears are taken as input from the TCIA (The Cancer Imaging Archive) database. In pre-processing, median filter is used to enhance the image without salt and pepper noise and Histogram equalisation is used to increase contrast. Image segmentation is used to segment the images. Here, watershed segmentation and K means clustering is used to extract the cells and nuclei from the input images. Feature extraction plays a major role in determining the type which consists of a lot of information present in the blast cells. Classifier is used to detect leukemia from the information obtained from blast cells. Performance parameters like accuracy, specificity, F1 scores, sensitivity and error rate are calculated using a confusion matrix. This work can be further carried out detect the subtypes (ALL, AML, CLL, CML) of leukemia.

REFERENCES

- [1]. Melanie MacGregor, Hanieh Safizadeh Shirazi, Kit Man Chan, Kola Ostrikov, Kym McNicholas, Alex Jay, Michael Chong, Alexander H. Staudacher, Thomas D. Michl, Aigerim Zhalgasbaikyzy, Michael P. Brown, Moein Navvab Kashani, Adam Di Fiore, Alex Grochowski, Stephen Robb, Simon Belcher, Jordan Li, Jonathan M. Gleadle, Krasimir Vasilev, "Cancer cell detection device for the diagnosis of bladder cancer from urine", in Biosensors and Bioelectronics, Volume 171, 2021.
- [2]. Milan Tripathi, "Analysis of Convolutional Neural Network based Image Classification Techniques", in Journal of Innovative Image Processing, 2021.
- [3]. Mugunthan, S. R., and T. Vijayakumar, "Design of Improved Version of Sigmoidal Function with Biases for Classification Task in ELM Domain", in Journal of Soft Computing Paradigm, 2021.
- [4]. Dr. J. Samuel Manoharan, Dr. V. Parthasaradi, K. Suganya. "A Spatio-Frequency Domain Anisotropic Filtering for Contrast Enhancement of Histopathological Images", in Annals of the Romanian Society for Cell Biology, May 2021.
- [5]. Astha Ratley, Mrs. Pooja and Mrs. Jasmine Minj Patre, "Leukemia Disease Detection and Classification Using Machine Learning Approaches", in International Conference on Power, Control and Computing Technologies, 2020.
- [6]. Alexandra Bodzas, Jan Zidek and Pavel Kodytek, "Automated Detection of Acute Lymphoblastic Leukemia from microscopic Images Based on Human Visual Perception", in Department of Cybernetics and Biomedical Engineering, research paper, published on 28 August 2020.
- [7]. Siddhika Arunachalam, "Applications of Machine learning and Image processing techniques in the detection of leukemia", in International Journal of Scientific Research in Computer science and Engineering, vol.8, Issue April (2020).



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details