

**LEUKEMIA DISEASE DETECTION AND
CLASSIFICATION USING MACHINE
LEARNING APPROCHES**

*A Project report submitted in partial fulfillment of the requirements for
the award of the degree of*

**BACHELOR OF TECHNOLOGY
IN
ELECTRONICS AND COMMUNICATION ENGINEERING**

Submitted by

K.SAIKIRAN(317126512086)

**A.PASYANTHI PADMA MALINI
(317126512061)**

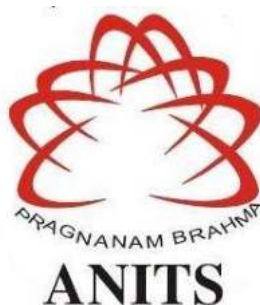
**M.SATYA SAI DURGA PRASAD
(317126512091)**

**A.SURYA VENKATA RAMYA
RAGHUVU(317126512062)**

Under the guidance of

Mr.BIBEKANANDA JENA(Ph.D)

Assistant Professor



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

**ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES
(UGC AUTONOMOUS)**

*(Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade)
Sangivalasa, bheemili mandal, visakhapatnam dist.(A.P)2020-2021*

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES
(UGC AUTONOMOUS)**

*(Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A'
Grade)*

Sangivalasa, Bheemili mandal, Visakhapatnam dist. (A.P)



This is to certify that the project report entitled "LEUKEMIA DISEASE DETECTION AND CLASSIFICATION USING MACHINE LEARNING APPROCHES" submitted by K.SAIKIRAN (317126512086), A.PASYANTHI PADMA MALINI (317126512061), M.SATYA SAI DURGA PRASAD (317126512091), A.SURYA VENKATA RAMYA RAGHUVU (317126512062) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Electronics & Communication Engineering of Andhra University, Visakhapatnam is a record of bonafide work carried out under my guidance and supervision.

Project Guide

B.Jena(PH.D)

Asst.Professor

Department of E.C.E

ANITS

**Assistant Professor
Department of E.C.E.**

Anil Neerukonda

**Institute of Technology & Sciences
Sangivalasa, Visakhapatnam-531 162**

Head of the Department


Dr. V.Rajyalakshmi

Professor&HOD

Department of E.C.E

ANITS

Head of the Department

Department of E C E

**Anil Neerukonda Institute of Technology & Sciences
Sangivalasa - 531 162**

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our project guide B.Jena ,Assistant Proffesor, Department of Electronics and Communication Engineering, ANITS, for his guidance with unsurpassed knowledge and immense encouragement. We are grateful to **Dr. V. Rajyalakshmi**, Head of the Department, Electronics and Communication Engineering, for providing us with the required facilities for the completion of the project work.

We are very much thankful to the **Principal and Management, ANITS, Sangivalasa**, for their encouragement and cooperation to carry out this work.

We express our thanks to all **teaching faculty** of Department of ECE, whose suggestions during reviews helped us in accomplishment of our project. We would like to thank **all non-teaching staff** of the Department of ECE, ANITS for providing great assistance in accomplishment of our project.

We would like to thank our parents, friends, and classmates for their encouragement throughout our project period. At last but not the least, we thank everyone for supporting us directly or indirectly in completing this project successfully.

PROJECT STUDENTS

K.Saikiran(317126512086),

A.Pasyanthi Padma Malini(317126512061),

M.Satya Sai Durga Prasad(317126512091),

A.Surya Venkata Ramya Raghuvu(317126512062)

ABSTRACT

Acute lymphoblastic leukemia(ALL) generally occurs in children these are distinguished by large number of lymphoid blasts in the blood. These generally occurs in the age group of 3-7 in children. The differentiating nature of these cells could lead to wrong diagnosis. Due to wrong diagnosis, it may lead to other disorders. Prudent examination of these cells by microscopic is the only way for the correct diagnosis. There are some techniques which are used for specific leukemia detection such as fluorescence in situ hybridization (FISH), immunophenotyping, cytogenetic analysis and cytochemistry. The diagnosis is costly and it often consumes more time so the need of automated leukemia detection has risen. An efficient and low cost technique is more often to save time and to start the treatment as per the diagnosis where we use image analysis for quantitative examination of stained blood microscopic images for leukemia detection. We can partition other blood components from leukocytes by using the fuzzy based two stage colour segmentation. The shape, texture features, novel shape features such as Hausdorff dimension and contour signature are often used for leukemia final destination. Total of 108 images were considered for feature extraction and Support Vector Machine(SVM) classification was employed.

In India fewer than 1 million cases of leukemia are reported every year. Some of the common symptoms are skin rashes, bleeding, feeling tired, fever and increased risk of infections. It is very important to detect leukemia at early stages. Traditional methods (such as microscopic analyses of blood smears) of detecting Leukemia are time consuming, not cost effective and totally dependent on medical personnel. To overcome these drawbacks we propose an automation algorithm using image processing for the detection and classification of Leukemia using processing tool MATLAB. In this process inputs are the microscopic images, and these images are processed using image processing techniques such as Image enhancement, segmentation, feature extraction and classification.

Leukemia means blood cancer which is featured by the uncontrolled and abnormal production of white blood cells (leukocytes) by the bone marrow in the blood. Analyzing microscopic blood cell images, diseases can be identified and diagnosed early. Hematologist are using technique of image processing to analyze, detect and identify leukemia types in patients recently. Detection through images is fast and cheap method as there is no special

need of equipment for lab testing. We have focused on the changes in the geometry of cells and statistical parameters like mean and standard deviation which separates white blood cells from other blood components using processing tools like MATLAB and LabVIEW. Images processing steps like image enhancement, image segmentation and feature extraction are applied on microscopic images.

CONTENTS

LIST OF SYMBOLS	viii
LIST OF FIGURES	ix
LIST OF TABLES	x
LIST OF ABBREVIATIONS	xi
CHAPTER 1 INTRODUCTION	12
1.1 Project Objective	12
1.2 Project Outline	13
CHAPTER 2 METHODS	15
2.1 Introduction	15
2.2 Methodology	16
2.3 Image acquisition	17
2.4 Preprocessing	19
2.4.1 Read image	19
2.4.2 Resize image	20
2.4.3 Remove noise	21
2.4.4 Segmentation	21
2.4.5 Morphology	22
2.5 Color Conversion	23
2.6 Image segmentation	24
2.6.1 Clustering	25
2.6.2 Classification of Image Clustering	26
2.6.3 Different Clustering Methods	26
2.6.4 K-Means Clustering	27
2.6.5 Fuzzy Clustering	27
2.6.6 Fuzzy C-Means Clustering	28
2.6.7 Algorithm steps for fuzzy c-means clustering	29
2.7 Sub Imaging	30
2.7.1 Bounding Box	31
2.7.2 Types of Boxes	31
CHAPTER 3 FEATURE EXTRACTION	33

3.1 Introduction	33
3.2 Fractal Dimension	35
3.3 Contour Signature	36
3.4 Shape Features	37
3.4.1 Area	38
3.4.2 Perimeter	38
3.4.3 Compactness	38
3.4.4 Solidity	38
3.4.5 Eccentricity	38
3.4.6 Formfactor	39
3.4.7 Elongation	39
3.4.8 Nuclear-cytoplasmic ratio	39
3.5 Color Feature Extraction	39
3.6 Texture features	40
3.6.1 Homogeneity	40
3.6.2 Energy	40
3.6.3 Correlation	41
3.6.4 Entropy	41
3.7 Classification	41
3.7.1 Support Vector Machine	42
CHAPTER 4 THE MATLAB	46
4.1 Introduction	46
4.2 Features of Matlab	46
4.3 Uses of Matlab	46
4.4 Local Environment Setup	47
4.5 The M files	50
4.6 Getting Help	50
4.7 Matlab Using Image Processing	51
4.7.1 Basic Image Import,Processing and Export	51
CHAPTER 5 EXPERIMENTAL RESULTS	55
5.1 Introduction	55

5.2 Blood Smear Image dataset	55
5.3 Morphological features of ALL blast cells	56
5.4 Image Processing	56
CHAPTER 6 ANALYSIS	61
CONCLUSIONS AND FUTURE WORK	62
REFERENCES	63

LIST OF SYMBOLS

k	Iteration step
β	Termination criterion
J	Objective function
N	Number of Squares
$N(s)$	Number of occupied boxes
a	Major axis
b	Minor axis
x	X-coordinate of pixel
y	Y-coordinate of pixel
R_{max}	Maximum distance
R_{min}	Minimum distance
i	Row element of cooccurrence matrix
j	column element of cooccurrence matrix
HD	Hausdorff dimension
σ^2	Variance
\bar{x}	Mean of x coordinate
\bar{y}	Mean of y coordinate

LIST OF FIGURES

Figure no	Title	Page no
Fig. 2.1	Read Image	20
Fig. 2.2	Resizing an Image	20
Fig. 2.3	Noisy image and Enhanced Image	21
Fig. 2.4	Original image and Segmented image	22
Fig. 2.5	Morphology of an Image	23
Fig. 2.6	Conversion of RGB to LAB	24
Fig. 2.7	Original image and Segmented image	25
Fig. 2.8	Cluster image	26
Fig. 2.9	Classification of Image Clustering	26
Fig. 2.10	FCM Segmented Image	29
Fig. 2.11	Before Bounding Box and After Bounding Box	32
Fig. 2.12	Separated nucleus subimages from bounding box image	32
Fig. 3.1	Leukemia and Non Leukemia cells	35
Fig. 3.2	Segmented outputs of Leukemia and Non Leukemia cells	35
Fig. 3.3	Support Vector Machine	46
Fig. 4.1	Setup Windows	47
Fig. 4.2	Default layout window	48
Fig. 4.3	Current folder window	48
Fig. 4.4	Command Window	49
Fig. 4.5	Workspace Window	49
Fig. 4.6	Command History	49
Fig.4.7	Image1	51
Fig.4.8	Histogram of Image 1	52
Fig.4.9	Image 2	53
Fig.4.10	Histogram of Image 2	54
Fig.5.1	Blood Smear Image Dataset	56

Fig.5.2	Original Image	57
Fig.5.3	Segmented Output	58
Fig.5.4	Separated nucleus subimages using bounding box technique	58
Fig.5.5	Nucleus image and Edge image	59
Fig.5.6	Box counting algorithm results	59

LIST OF TABLES

Table no	Title	Page no
Table 5.1	Results of Shape Features	60
Table 5.2	Results of Texture features	60
Table 7.1	Results of Features	62

LIST OF ABBREVIATIONS

ALL	Acute lymphoblastic leukemia
CAD	Computer-aided detection
WBC	White blood cells
3D	Three dimensional
SVM	Support Vector Machine
FCM	Fuzzy C-Means
MRF	Markov Arbitrary Field
OBB	Oriented Bounding Box
FDH	Fixed-direction hull
CH	Convex hull
HD	Hausdorff Dimension
ANN	Artificial Neural Network
RF	Random Forest
KNN	K-Nearest Neighbour
MLP	Multilayer Perception
PCA	Principal Component Analysis

CHAPTER 1

INTRODUCTION

Acute lymphoblastic leukemia (ALL):

Acute lymphoblastic leukemia (ALL), also called acute lymphoblastic leukemia and acute lymphoid leukemia, is a blood cancer that results when abnormal white blood cells (leukemia cells) accumulate in the bone marrow. ALL progresses rapidly, replacing healthy cells that produce functional lymphocytes with leukemia cells that can't mature properly. The leukemia cells are carried in the bloodstream to other organs and tissues, including the brain, liver, lymph nodes and testes, where they continue to grow and divide. The growing, dividing and spreading of these leukemia cells may result in a number of possible symptoms. ALL is typically associated with having more B lymphatic cells than T cells. B and T cells play active roles in preventing the body from infections and germs and destroying cells that have already become infected. B cells particularly help prevent germs from infecting the body while T cells destroy the infected cells.

1.1 Project Objective:

The need of the leukemia detection by using image processing is because the above diagnosis time consuming and costly. As it is difficult to differentiate the leukaemia cells from white blood cells by using clinical methods as it changes with respect to time. By this method we can easily differentiate the cells from white blood cells in less time and most effective way just by considering the microscopic images.

The diagnosis of ALL requires a broad spectrum of information derived from several modalities, including morphology, cell phenotyping, cytochemistry, cytogenetics, and molecular genetics. Despite technological advances in medicine, morphology remains the frontline hematological diagnostic technique. The observation of excessive leukemic cell buildup and morphological anomalies in cellular structures during the visual examination of peripheral blood smears arouses the first suspicion of leukemia. Because manual microscopic examination is a time-consuming process that requires a considerable amount of experience and is prone to humane error, such an automated inspection is needed, which would standardize .

To minimize human intervention and overcome the above mentioned limitations, several computerized methods have been explored. Most of these methods utilize conventional image processing and machine learning techniques, which involve mainly segmentation, feature extraction, and classification methods. Especially the segmentation and feature extraction phases are considered the most significant and challenging task. The main reason lies in the large variety of blood smear images, taken under different conditions, and the potential morphological differences between blast cells. Although some of these proposed methods were found to be faster and more cost effective than manual examination, their impact and accuracy remain insufficient. Whereas, achieved a detection speed of 14 to 100 milliseconds by utilizing convolution neural networks and GPU, most proposed methods produce false-negative errors and achieve overall accuracy.

Here task is to detect immature cell using different image processing techniques and count total number of cells. So we need to use the technology that identifies different types of blood cells within short duration of time in emergency. Furthermore it is vital to study in detail how to differentiate different cell and recognize it as immature cell and according to it, detect the leukemia. Acute and chronic also have two types. lymphoblastic and myelolastic that both are due to immature blast of lymphoid and myeloid cell respectively.

1.2 Project outline:

A group of hematological neoplasia which generally affects blood, bone marrow and lymph nodes is known as Leukemia. The abnormal increment of white blood cells in bone marrow which is not responding to cell growth inhibitors is characterized as leukemia. It suppresses the hematopoiesis and thereafter anemia, thrombocytopenia and neutropenia which results due to leukemia. Due to abnormal rise in the count of WBC can also accumulate in various sites such as meninges, gonads, thymus, liver, spleen and lymph nodes. They also flow into the blood stream due to rise of lymphoid blast. The symptoms of leukemia indicates excess amount of lymphoid blast cells. For classification and identification of blast cells the hematologists examines the blood smear under the microscope. Acute and chronic can be classified by pathologically the leukemia.

Acute lymphoblastic leukemia(ALL) only considered and the objective is to classify the lymphocyte as a normal or lymphoblast in the present paper. Standard leukemia diagnosis technique regardless of advanced techniques is the examination of microscopic blood cells. The examination of leukemia is done under the microscope the results are standard when the operator is experienced and tiredness which results in inconsistent and subjective reports. The most economical way of leukemia diagnosis is examining under the microscope. So we need an effective cost technique and accurate method to detect the leukemia which gives the accurate result without involvement of the operator fatigue.

Many automatic segmentation and leukemia detection has been proposed over these years. The most of the techniques were based on local image information. By using the HSV color model a two step segmentation is used in [4]. Much work has been there in order to meet the real clinical methods due to the complex nature of leukemia cells. There are similar researches on segmentation and detection in literature. It solely depends on the automated segmentation and detection of leukemia. We proposed the automation method in the present paper which is for examining the blood smear which can supplement to the physician for better diagnosis and treatment. The method proposed in the present paper is to separate the other blood components from leukocytes and extract the lymphocytes. From that extracted lymphocytes fractal features, shape features and other texture features are extracted. For cell nucleus boundary roughness measurement two new features were proposed for leukemia detection. Based on extracted features the images are classified into healthy and leukemia by Support vector machine(SVM).

Its cure rate and prognosis depends mainly on the early detection and diagnosis of the disease. So detection of leukemia using image processing is a better and an easy way because images are cheap and do not require expensive testing and lab equipment and gives faster output which overcomes the disadvantages in manual testing. Leukemia is a cancer of early blood-forming cells, most frequently of the white blood cells although some leukemia begins on other blood cell types. Leukemia can be described as fast-growing (acute) or slow growing (chronic). The different types of leukemia have varied outlooks and treatment options.

Chapter 2

Methods

2.1 Introduction:

The Process for identification of leukocyte in microscopic image starts with preprocessing, segmentation, feature extraction and classification. The microscopic image consists of Red Blood Cells, White Blood Cells and Platelets. The method we have chosen is based on color image segmentation and our aim is to separate White Blood Cells from the nucleus and cytoplasm. Acute Leukemia is taken and the cytoplasm is paltry so we depend on nucleus and vital characteristics are drawn out.

Digital histopathology has witnessed a lot of improvement in the recent years. With the new technological advancement, much effective methods have been proposed for automated microscopic image analysis. Due to this development, computer-aided detection (CAD) is becoming a reliable method for ALL detection. CAD system for ALL detection can be divided into four phases, namely, preprocessing, segmentation, feature extraction, and classification. This section provides a detailed survey of different techniques and methods that have been proposed, developed, and used in the automatic detection of acute lymphocytic leukaemia cell.

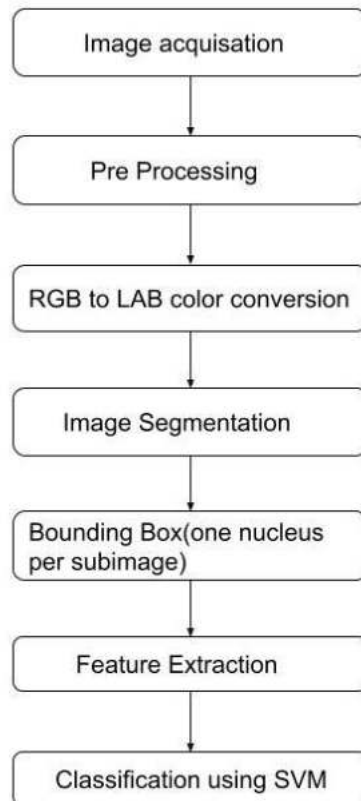
This method is an incomparable method that has been used to detect acute lymphoblastic leukemia (ALL) from the microscopic images of peripheral blood smear. The block diagram of the proposed methodology is shown in Fig. 1. In the proposed method, color threshold has been used to separate white blood cells (WBCs) or lymphocytes (which are one kind of WBC) from the microscopic images of blood smear. At first only the white blood cells have been segregated and has been removed the other blood cells (like RBCs, PLT) because acute lymphoblastic leukemia (ALL) is produced by the immature lymphocyte cells (called lymphoblast) which is a type of WBC. So, the several operation has applied only the WBCs which are affected by ALL. The method has been used in this work is simple, easy to understand, easy to calculate, fast and more accurate. The system has developed in MATLAB.

The main goal of this work is to develop a fully automated system for ALL detection that can be applied to complete blood smear images containing multiple WBCs. The

solution presented in this paper is based on conventional image processing techniques and comprises four main stages, which are described in the following sub chapters.

This inspection is performed to find white blood cells that are not normal as an indication of cancer cells. Over the decades, the inspection was usually done by experienced operators, who will usually do two analysis: classification and counting of cells (which are now carried out by cytometer). The morphological analysis does not usually require a blood sample because it can be done through a single image. Therefore, this analysis does not require matching done due to the expensive cost, accuracy is usually the same for different images, and the system is remote screening. In computer vision technique, the stages of detecting leukemia using microscopic image processing consists of five stages: image acquisition, pre-processing, image segmentation, feature extraction and detection of leukemia cells.

2.2 Methodology:



2.3 Image Acquisition:

Until the early 1990s, most image acquisition in video microscopy applications

was typically done with an analog video camera, often simply closed circuit TV cameras. While this required the use of a frame grabber to digitize the images, video cameras provided images at full video frame rate (25-30 frames per second) allowing live video recording and processing. While the advent of solid state detectors yielded several advantages, the real-time video camera was actually superior in many respects.

Image acquisition in image processing can be broadly defined as the action of retrieving an image from some source, usually a hardware-based source, so it can be passed through whatever processes need to occur afterward. Performing image acquisition in image processing is always the first step in the workflow sequence because, without an image, no processing is possible. The image that is acquired is completely unprocessed and is the result of whatever hardware was used to generate it, which can be very important in some fields to have a consistent baseline from which to work. One of the ultimate goals of this process is to have a source of input that operates within such controlled and measured guidelines that the same image can, if necessary, be nearly perfectly reproduced under the same conditions so anomalous factors are easier to locate and eliminate.

Depending on the field of work, a major factor involved in image acquisition in image processing sometimes is the initial setup and long-term maintenance of the hardware used to capture the images. The actual hardware device can be anything from a desktop scanner to a massive optical telescope. If the hardware is not properly configured and aligned, then visual artifacts can be produced that can complicate the image processing. Improperly setup hardware also may provide images that are of such low quality that they cannot be salvaged even with extensive processing. All of these elements are vital to certain areas, such as comparative image processing, which looks for specific differences between image sets.

One of the forms of image acquisition in image processing is known as real-time image acquisition. This usually involves retrieving images from a source that is automatically capturing images. Real-time image acquisition creates a stream of files that can be automatically processed, queued for later work, or stitched into a single media format. One common technology that is used with real-time image processing is known as background image acquisition, which

describes both software and hardware that can quickly preserve the images flooding into a system.

For best results, one must select an appropriate sensor for a given application. Because microscope images have an intrinsic limiting resolution, it often makes little sense to use a noisy, high resolution detector for image acquisition. A more modest detector, with larger pixels, can often produce much higher quality images because of reduced noise.

Microscope image processing is a broad term that covers the use of digital image processing techniques to process, analyze and present images obtained from a microscope. Such processing is now common place in a number of diverse fields such as medicine, biological research, cancer research, etc. A number of manufacturers of microscopes now specifically design in features that allow the microscopes to interface to an image processing system. For best results, one must select an appropriate sensor for a given application. Because microscope images have an intrinsic limiting resolution, it often makes little sense to use a noisy, high resolution detector for image acquisition. A more modest detector, with larger pixels, can often produce much higher quality images because of reduced noise. This is especially important in low-light applications such as fluorescence microscopy.

There are some advanced methods of image acquisition in image processing that actually use customized hardware. Three-dimensional (3D) image acquisition is one of these methods. This can require the use of two or more cameras that have been aligned at precisely describes points around a target, forming a sequence of images that can be aligned to create a 3D or stereoscopic scene, or to measure distances. Some satellites use 3D image acquisition techniques to build accurate models of different surfaces.

Moreover, one must also consider the temporal resolution requirements of the application. A lower resolution detector will often have a significantly higher acquisition rate, permitting the observation of faster events. Conversely, if the observed object is motionless, one may wish to acquire images at the highest possible spatial resolution without regard to the time required to acquire a single image.

2.4 Preprocessing:

Image pre-processing is the term for operations on images at the lowest level of abstraction. These operations do not increase image information content but they decrease it if entropy is an information measure. The aim of pre-processing is an improvement of the image data that suppresses undesired distortions or enhances some image features relevant for further processing and analysis task. Image preprocessing use the redundancy in images. Neighboring pixels corresponding to one real object have the same or similar brightness value. If a distorted pixel can be picked out from the image, it can be restored as an average value of neighboring pixels. Image pre-processing methods can be classified into categories according to the size of the pixel neighborhood that is used for the calculation of a new pixel brightness.

The steps to be taken in preprocessing are:

1. Read image
2. Resize image
3. Remove noise (Denoise)
4. Segmentation
5. Morphology (smoothing edges)

2.4.1 Read Image:

In this step, we store the path to our image dataset into a variable then we created a function to load folders containing images into arrays.

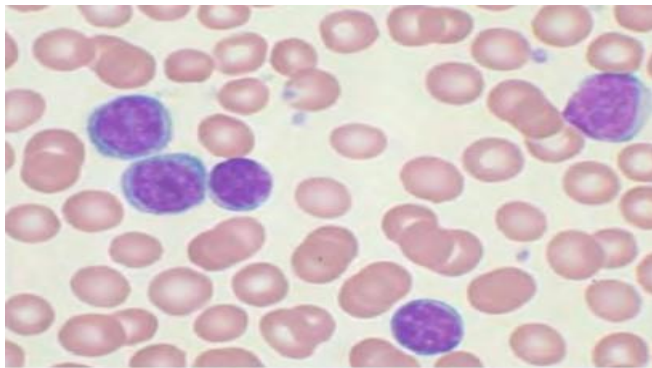


Fig 2.1 Read image

2.4.2 Resize Image:

Image resizing is necessary when you need to increase or decrease the total number of pixels, whereas remapping can occur when you are correcting for lens distortion or rotating an image. Zooming refers to increase the quantity of pixels, so that when you zoom an image, you will see more detail.

Many compact digital microscopes can perform both an optical and a digital zoom. A microscope performs an optical zoom by moving the zoom lens so that it increases the magnification of light. However, a digital zoom degrades quality by simply interpolating the image. Even though the image with digital zoom contains the same number of pixels, the detail is clearly far less than with optical zoom.

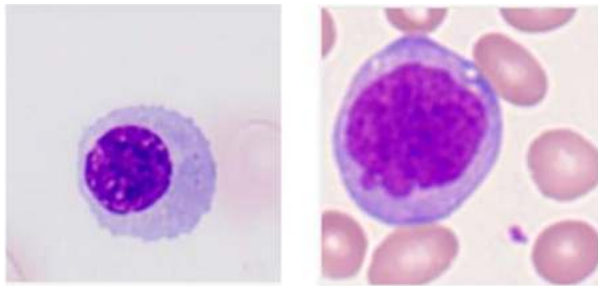


Fig 2.2 Resizing an image

2.4.3 Remove Noise:

Noise reduction is the process of removing noise from a signal. Noise reduction techniques exist for audio and images. Noise reduction algorithms may distort the signal to some degree. All signal processing devices, both analog and digital, have traits that make them susceptible to noise. Noise can be random or white noise with an even frequency distribution, or frequency-dependent noise introduced by a device's mechanism or signal processing algorithms.

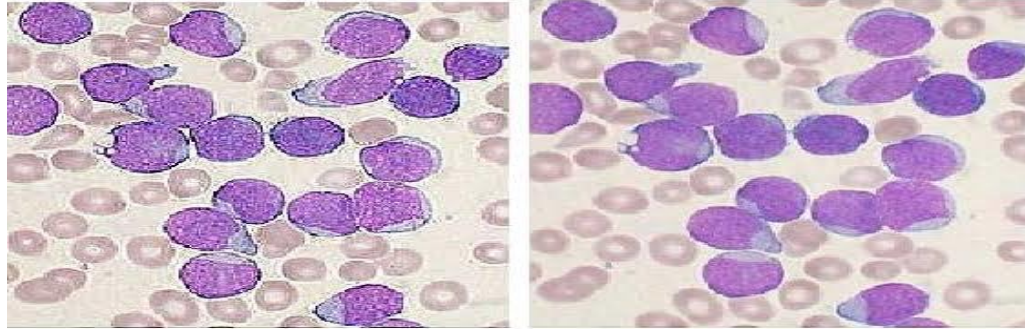


Fig 2.3 Noisy image and Enhanced image

Median Filtering:

The median filter is a non-linear digital filtering technique, often used to remove noise from an image or signal. Such noise reduction is a typical pre-processing step to improve the results of later processing. Median filtering is very widely used in digital image processing because, under certain conditions, it preserves edges while removing noise, also having applications in signal processing.

The main idea of the median filter is to run through the signal entry by entry, replacing each entry with the median of neighboring entries. Median filtering is one kind of smoothing technique, as is linear Gaussian filtering. All smoothing techniques are effective at removing noise in smooth patches or smooth regions of a signal, but adversely affect edges. Often though, at the same time as reducing the noise in a signal, it is important to preserve the edges. Because of this, median filtering is very widely used in digital image processing.

2.4.4 Segmentation:

In digital image processing and computer vision, image segmentation is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in a region are

similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristics.

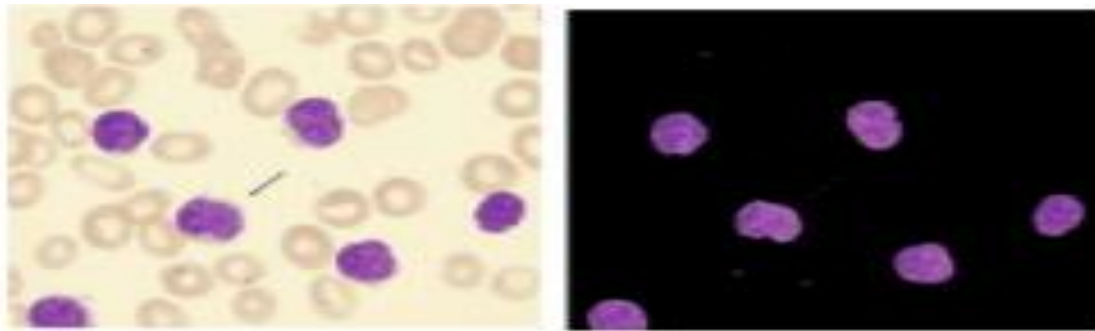


Fig 2.4 Original image and Segmented image

2.4.5 Morphology:

Morphological image processing is a collection of non-linear operations related to the shape or morphology of features in an image. Morphological operations rely only on the relative ordering of pixel values, not on their numerical values, and therefore are especially suited to the processing of binary images. Morphological operations can also be applied to greyscale images such that their light transfer functions are unknown and therefore their absolute pixel value

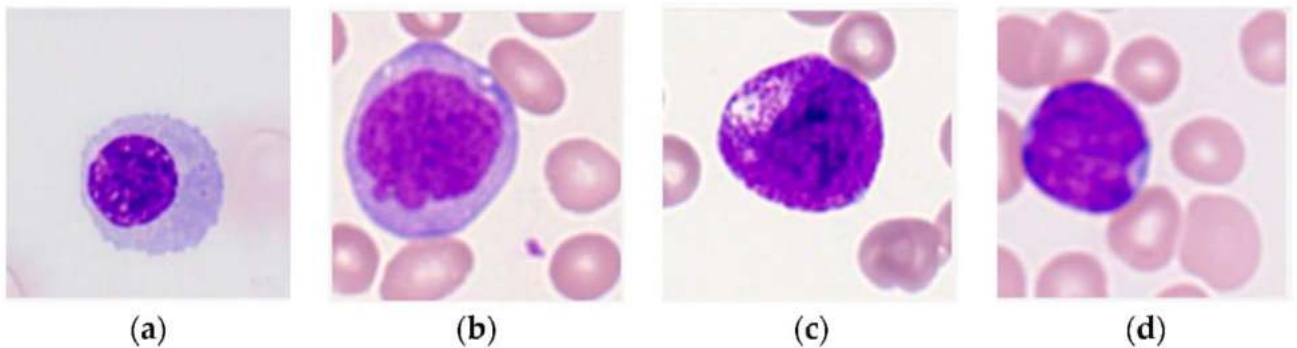


Fig 2.5 Morphology of an image

2.5 Color Conversion:

The images taken from the digital microscope are in RGB color space which are difficult to segment. So the RGB color images are converted into LAB color space which is used to reduce color dimension from three to two. The $L^*a^*b^*$ consists of L^* known as luminosity layer and a^*, b^* are known as chromaticity layer. Here a^* and b^* represent color dimension. The

$L^*a^*b^*$ color space has less color dimension so this is mostly used in color based clustering.

An RGB color space is any additive color space based on the RGB color model. A particular RGB color space is defined by the three chromaticities of the red, green, and blue additive primaries, and can produce any chromaticity that is the triangle defined by those primary colors. The complete specification of an RGB color space also requires a white point chromaticity and a gamma correction curve.

A Lab color space is a color-opponent space, with dimension L for lightness and a and b for the color-opponent dimensions. The nonlinear relations for L^* , a^* , and b^* are intended to mimic the nonlinear response of the eye. Furthermore uniform changes of components in the $L^*a^*b^*$ color space aim to correspond to uniform changes in perceived color, so the relative perceptual differences between any two colors in $L^*a^*b^*$ can be approximated by treating each color as a point in a three dimensional space L^* , a^* , b^* and taking the Euclidean distance between them.

This section describes $L^*a^*b^*$ colorspace that is used to aspire the perceptual uniformity of human vision. A color space is a method to specify, create and visualize color. As humans, color may be defined by its attributes of brightness, hue and colorfulness. A computer may describe a color using the amounts of red, green and blue phosphor emission required to match a color. A color is specified using co-ordinates, or parameters. These parameters describe the position of the color within the color space being used.

$L^*a^*b^*$ color is designed to approximate human vision. It aspires to perceptual uniformity, and its L^* component closely matches human perception of lightness [2]. When a color is expressed in CIELAB, L^* defines lightness. The $L^*a^*b^*$ space consists of a luminosity-layer L^* , chromaticity-layer a^* indicating where color falls along the red-green axis, and chromaticity-layer b^* indicating where the color falls along the blue-yellow axis. All of the color information is in the a^* and b^* layers. This color space is better suited to many digital image manipulations than RGB color space, which is typically used in image editing programs. $L^*a^*b^*$ color space is useful for sharpening images and the removing artifacts in images or various images from digital cameras and satellite.

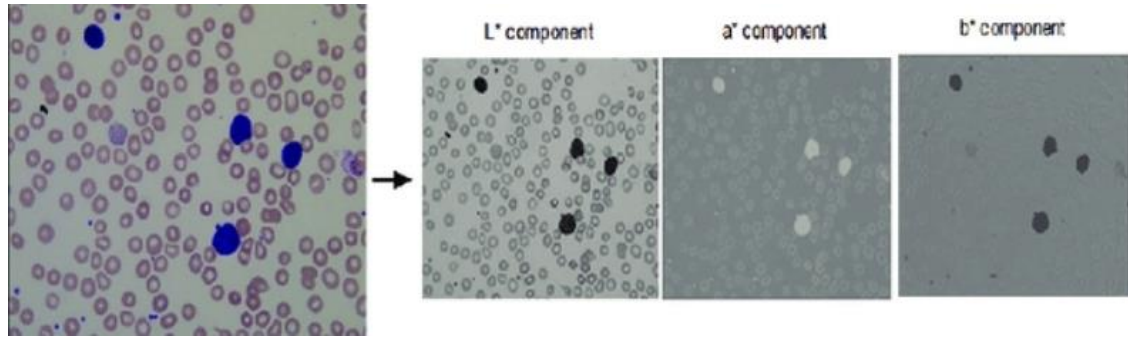


Fig 2.6 Conversion of RGB to LAB

2.6 Image Segmentation:

Image segmentation is a commonly used technique in digital image processing and analysis to partition an image into multiple parts or regions, often based on the characteristics of the pixels in the image. Image segmentation could involve separating foreground from background, or clustering regions of pixels based on similarities in color or shape.

Several algorithms and techniques for image segmentation have been developed over the years using domain-specific knowledge to effectively solve segmentation problems in that specific application area. These applications include medical imaging, automated driving, video surveillance, and machine vision. During medical diagnosis for cancer, pathologists stain body tissue with hematoxylin and eosin to distinguish between tissue types. They then use an image segmentation technique called clustering to identify those tissue types in their images.

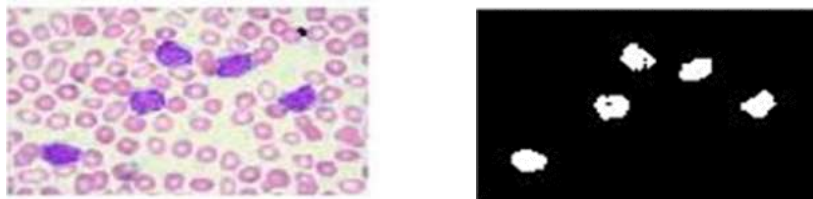


Fig 2.7 Original image and Segmented image

2.6.1 Clustering:

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same

group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. It is the undertaking of collection an arrangement of articles such that items in a similar gathering (called a cluster) are more comparative (in some sense or another) to each other than to those in dissimilar gatherings (groups). Bunch examination itself isn't one particular calculation, yet the general assignment to be unraveled. It can be accomplished by different calculations that vary essentially in their idea of what constitutes a group and how to effectively discover them. Prominent ideas of bunches incorporate gatherings with little separations among the group individuals, thick zones of the information space, interims or specific measurable circulations. Grouping can be detailed as a multi-target enhancement issue. Group examination all things considered isn't a programmed task, yet an iterative procedure of information revelation or intuitive multi-target progression that includes trial and disappointment. It is frequently important to change information pre processing and demonstrate parameters until the point when the outcome accomplishes the coveted properties.

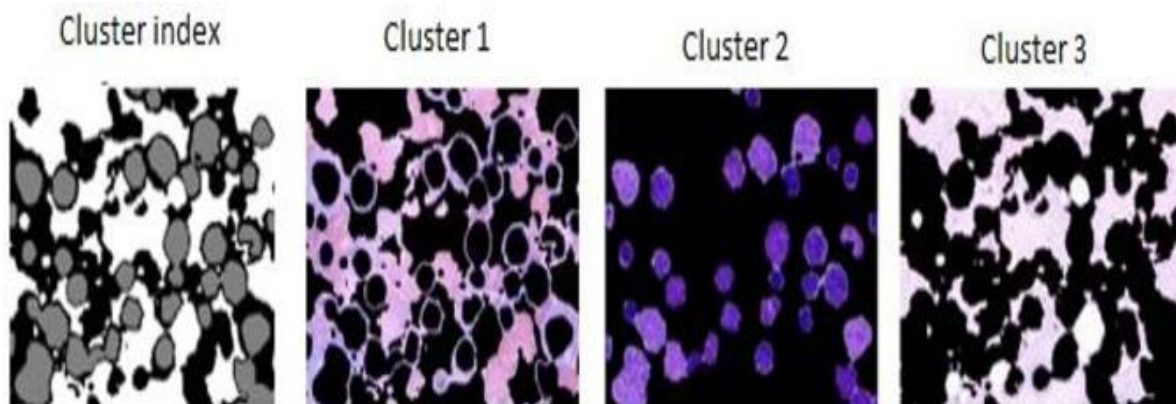


Fig 2.8 Cluster image

2.6.2 Classification of Image Clustering:

Image clustering identifies with content-based picture recovery frameworks. It empowers the usage of proficient recovery calculations and the production of an easy to use interface to the database. There are several clustering techniques:

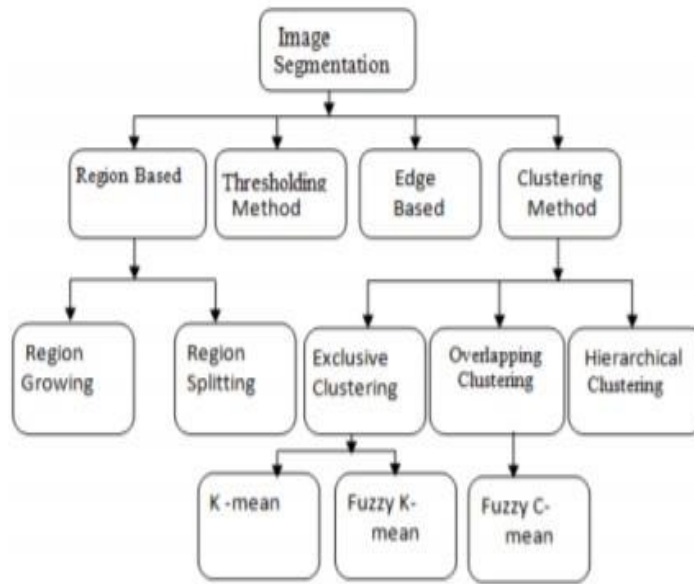


Fig 2.9 Classification of Image Clustering

2.6.3 Different Clustering Methods:

In order to perform image clustering, we initially need to pick a portrayal space and after that to utilize a proper separation measure (closeness measure), to coordinate amongst pictures and group focuses in the chose portrayal space. The picture grouping is then performed in an administered procedure, utilizing human intercession or in an unsupervised procedure, depending on the likeness between the pictures and the different bunch focuses.

2.6.4 K-Means clustering:

K-means is one of the simplest unsupervised learning algorithms that explain the well known clustering problem. The method takes after a straightforward and simple approach to characterize a given informational collection through a specific number of groups (expect k clusters) settled from the earlier. The primary thought is to characterize k centroids, one for each group. These centroids ought to be set shrewdly as a result of various area causes diverse outcome. In this way, the better decision is to put them however much as could reasonably be expected far from each other. The subsequent stage is to take each guide having a place toward a given informational index and associate it to the closest centroid. At the point when no point is waiting, the initial step is finished and an early group age is finished. Presently we need to refigure k new

centroids as bary centers of the packs coming to fruition due to the past progress. After we have these k new centroids, another coupling must be done between comparable educational list centers and the nearest new centroid. A circle has been made. In view of this circle we may see that the k centroids change their zone all around requested until the point that the moment that no more changes are done. By the day's end centroids don't move any more.

2.6.5 Fuzzy clustering:

Fuzzy clustering generalizes partition clustering methods by allowing an individual to be partially classified into more than one cluster. In regular clustering, each individual is a member of only one cluster. In fuzzy clustering, the membership is spread among all clusters. The m_{ik} can now be between zero and one, with the stipulation that the sum of their values is one. We call this a fuzzification of the cluster configuration. It has the advantage that it does not force every object into a specific cluster. It has the disadvantage that there is much more information to be interpreted.

2.6.6 Fuzzy C-Means Clustering:

Fuzzy C-Means (FCM) Clustering is the most wide spread clustering approach for image segmentation because of its robust characteristics for data classification.

FCM is one such soft segmentation technique applicable for medical images. The performance of this method to obtain an optimal solution depends on the initial positions of the centers of the clusters, the measure of membership degree for each data point, and so on. In the standard FCM, the centers are initialized randomly and the measure of membership only uses the gray feature. This leads to be quite time-consuming and be sensitive to noise. For years, many research efforts have been made towards effective FCM image segmentation approaches. In order to accelerate the segmentation process, some approaches focus on how to initialize the centers of required clusters. As noise is always emerged in the medical images, some FCM segmentation approaches which have less sensitivity to noise are also presented. Moreover, in order to overcome the defect that the FCM is easy to fall into local optimal solution, some scholars combine the FCM and other mathematical approaches.

The most admired algorithm in the fuzzy clustering is the Fuzzy C-Means (FCM) algorithm [S]. Fuzzy c-means (FCM) is a technique of clustering which allows one piece of data to belong to two or more clusters. Our investigation demonstrated that FCM has a noteworthy issue: a lot of capacity necessity. Keeping in mind the end goal to defeat this issue, we have built up an altered variant of FCM (from this point forward called MFCM) which utilizes a recursive strategy, instead of a clump system utilized as a part of FCM, to refresh group focuses [6]. We connected the MFCM to the picture pressure issue and showed that it can diminish the capacity essentially. In this work, we build up a picture partition calculation in light of the MFCM bunching calculation. The MFCM (or FCM) groups each picture pixel (test) without the utilization of spatial requirements. To enhance the division, we adjust the target capacity of the FCM to incorporate spatial requirements. We make utilization of the spatial requirements by accepting that the measurable model of picture bunches is the Markov Arbitrary Field (MRF). The MRF has stirred wide consideration as of late, which has ended up being an intense displaying device in a few parts of picture preparing

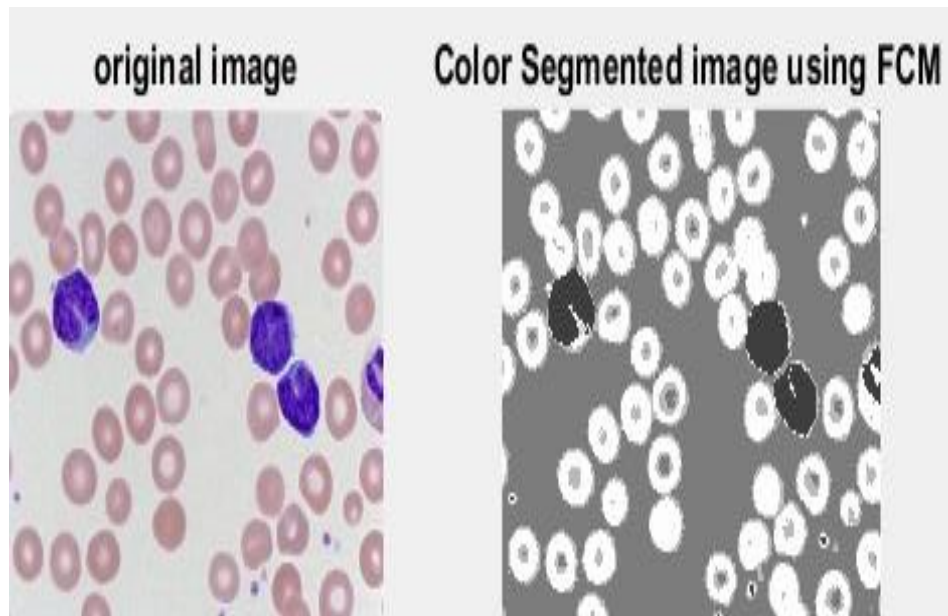


Fig 2.10 FCM segmented image

2.6.7 Algorithm steps for fuzzy c-means clustering:

Let $X = \{x_1, x_2, x_3 \dots, x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3 \dots, v_c\}$ be the set of centers.

1. Randomly select 'c' cluster centers.

2. Calculate the fuzzy membership μ_{ij} using:

$$\mu_{ij} = \text{Equation}$$

3. Compute the fuzzy centers v_j using:

$$v_j = \text{Equation}$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^c (d_{ij} / d_{ik})^{(2/m-1)}}$$

4. Repeat step two and three until the minimum J value is achieved or $||U(k+1) - U(k)|| < \beta$.

Where,

$$v_j = \frac{\sum_{i=1}^n (\mu_{ij})^m x_i}{\sum_{i=1}^n (\mu_{ij})^m}, \forall j = 1, 2, \dots, c$$

k is the iteration step.

β is the termination criterion between [0, 1].

r

$U = (\mu_{ij})_{n \times c}$ is the fuzzy membership matrix.

J is the objective function.

2.7 Sub Imaging:

Sub images containing single nucleus per image are essential for feature extraction and were obtained using bounding box technique. The cluster images are large, so for accurate leukemia detection each nucleus feature has to be extracted for classifying as a blast cell. For feature extraction sub images containing single nucleus are essential which are obtained using

bounding box technique. By using image morphology we select only the sub images which contain lymphocytes. The association of nucleus sub images of neutrophils, eosinophils and basophils would not be considered for feature extraction.

Another approach called subimage-by subimage or block-by-block process. In this method an image is divided into many subimages and each subimage is processed separately and then combined with the others. The size of the subimage is typically between to pixels.

As peripheral blood smear images are relatively larger usually, the cluster images are also large. But for accurate leukemia detection each nucleus feature has to be extracted individually for classifying it as a blast cell. Sub images containing single nucleus per image are essential for feature extraction and were obtained using bounding box [9] technique. Using image morphology we select only those sub images which contain lymphocytes. The nucleus sub images of neutrophils, eosinophils and basophils are not considered for feature extraction as they are not associated with lymphocytic leukemia.

2.7.1 Bounding Box:

In digital image processing, the bounding box is merely the coordinates of the rectangular border that fully encloses a digital image when it is placed over a page, a canvas, a screen or other similar bi-dimensional background. A bounding box is an imaginary rectangle that serves as a point of reference for object detection and creates a collision box for that object.

Data annotators draw these rectangles over images, outlining the object of interest within each image by defining its X and Y coordinates. This makes it easier for machine learning algorithms to find what they're looking for, determine collision paths, and conserves valuable computing resources. Compared to other image processing methods, this method can reduce costs and increase annotation efficiency.

The bounding boxes are typically used in training self-driving car vision models to identify different types of artifacts on the road, such as traffic signals, lane barriers, and pedestrians, among other items. Locate the presence of objects with a bounding box and types or classes of the located

objects in an image. An image with one or more objects, such as a photograph. One or more bounding boxes (e.g. defined by a point, width, and height), and a class label for each bounding box.

2.7.2 Types of Boxes:

There are five types of bounding boxes, i.e., a surrounding sphere (SS), an axis-aligned bounding box (AABB), an oriented bounding box (OBB), a fixed-direction hull (FDH), and a convex hull (CH). A type box is also known as a bounding box. Which photo shop tools allows users to select an area of an image using a specific color, or color range? ... Using the Image menu is the only way to modify the size of a canvas.

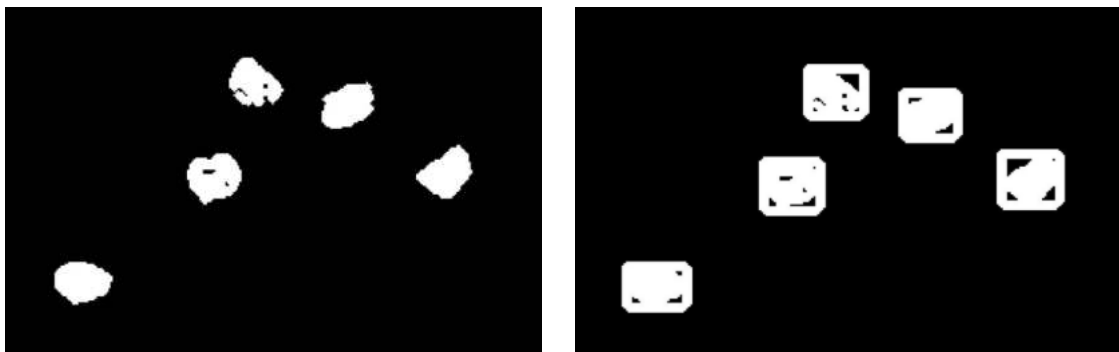


Fig 2.11 Before Bounding Box and After Bounding Box

The separated nucleus subimages using bounding box technique as:



Fig 2.12 Separated nucleus subimages from bounding box image

CHAPTER III

FEATURE EXTRACTION

3.1 Introduction:

Leukemia, simple called “Blood cancer” in which usually the number of WBC increase in the bone marrow and peripheral blood. These leukemic cells (usually immature) replace the other normal blood cells causes malfunction of the bone marrow as well as peripheral blood. A massive contribution has been done by researchers in the aim of leukemia detection and diagnosis using image processing over the past years. The researches have been different from the segmentation methods and classification methods they have used. This blog post discuss on the feature extraction techniques used in leukemia diagnosis using image processing. Feature extraction redefines a large set of redundant data into a set of reduced dimensions called features. Feature extraction helps to check the resulted values of the parameters with the standard values and differentiate between leukemic and healthy data. In feature extraction, the acquired data from the image is transformed and labelled to a particular set of features, which is going to be used for further classification. This stage used to extract and identify the features derived from the objects that were segmented from parts of the image or from the whole image. In other words, transforming the data that is obtained from the image into the set of features for pattern recognition is called feature extraction.

One of the important issues related to pattern recognition is choosing the relevant set of features extraction in order to extract the relevant information to perform the task and get accurate information. The features extraction have been used in many applications such as leukemia detection, character recognition, reading bank deposit slips, applications for credit cards, tax forms, data entry, check sorting and others . Many features can be extracted from the objects in the image, such as the shape features (e.g. area, perimeter, solidity and others), Texture Features (e.g. homogeneity, energy, angular second, entropy contrast, and others), Statistical Features (e.g. mean, skewness and, variance), Geometrical features (e.g. perimeter, area, compactnes and symmetry), color features and so forth . In acute leukemia detection, the features extraction stage plays an important role in determining the leukemia type because blast cells (ROI) have a lot of information that included characteristics of nucleus and cytoplasm . There are different features have been extracted in the current. The results of feature extraction stage will be useful for the

classification stage (next stage) The binary equivalent images produced by the segmentation technique of blood cell and cell nucleus are used to extract those morphological features The extracted features tells us the texture information derived from the segmented pattern and thereby help to reduce the dimension of the image to produce a result that is more informative and less redundant than the original image. In this phase ,we targeted to extract the descriptive information from an image. The correct selection of the feature is considered the second most challenging step in the field of automated identification of leukemic cells. In this work ,we implemented sixteen widely used features ,of which nine had shape features and seven had texture characteristics.

Feature extraction in image processing is a technique of redefining a large set of redundant data into a set of features (or feature vector) of reduced dimension. This transformation of the input data into the set of features is called feature extraction . In the present paper broadly three types of features are extracted i.e. fractal dimension, shape features including contour signature and texture. In addition color features are also extracted from the nucleus image. Feature extraction is a part of the dimensionality reduction process, in which, an initial set of the raw data is divided and reduced to more manageable groups. So when you want to process it will be easier. The most important characteristic of these large data sets is that they have a large number of variables. These variables require a lot of computing resources to process them. So Feature extraction helps to get the best feature from those big data sets by select and combine variables into features, thus, effectively reducing the amount of data. These features are easy to process, but still able to describe the actual data set with the accuracy and originality. The technique of extracting the features is useful when you have a large data set and need to reduce the number of resources without losing any important or relevant information. Feature extraction helps to reduce the amount of redundant data from the data set. In the end, the reduction of the data helps to build the model with less machine's efforts and also increase the speed of learning and generalization steps in the machine learning process. Some of the features that are subjected to explore are given below.

In machine learning, pattern recognition, and image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human.A Detailed Review of Feature Extraction in Image

Processing Systems. Abstract: Feature plays a very important role in the area of image processing. Before getting features, various image preprocessing techniques like binarization, thresholding, resizing, normalization etc. are applied on the sampled image. Feature detection generally concerns a low-level processing operation on an image. It examines every pixel to see if there is a feature present at that pixel. The types of image features include “edges,” “corners,” “blobs/regions,” and “ridges,” which will be stated in Sect.

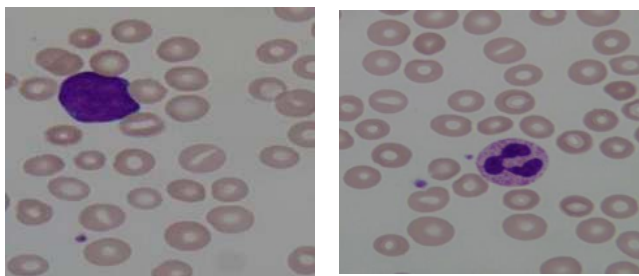


Fig 3.1 Lekemia and Non Leukemia cells



Fig 3.2 Segmented outputs of Leukemia and Non Leukemia cells

3.2 Fractal Dimension:

A fractal dimension is an index for characterizing fractal patterns or sets by quantifying their as a ratio of the complexity change in detail to the change in scale. Several types of fractal dimension can be measured theoretically and empirically. Fractal dimensions are used to characterize a broad spectrum of objects ranging from the abstract to practical phenomena, including turbulence, river networks, urban growth, human physiology, medicine, and market trends. The essential idea of fractional has a long history in mathematics that can be traced back to the 1600s, but the terms fractal dimension were coined by mathematician Benoit Mandelbrot in 1975.

Fractals have been used in medicine and science earlier for various quantitative measurement. Perimeter roughness of nucleus is a important measure that decides whether a particular nucleus represents a lymphoblast or a mature lymphocyte. Fractal geometry is a more convenient way to parameterize the cell boundary surface in comparison to euclidean geometry. Hausdorff dimension is an essential feature for fractal geometry and will be an essential quantitative measure for cell boundary roughness measurement. The procedure for Hausdorff Dimension measurement using box counting method is introduced below as an algorithm:

1. Each nucleus color (RGB) image is converted to gray and successively to binary image.
2. Nucleus edge boundary is extracted using Canny edge detection technique.
3. A grid of N squares is superimposed over the edges,while counting the edge occupied squares.
4. Step 3 is continued for an increasing number of squares.
5. The Hausdorff Dimension (HD) may then be defined as in (1).

$$HD = \frac{\log(N)}{\log(N(s))}$$

where, N is the number of squares in the superimposed grid and N(s) is the number of occupied squares or boxes (box count). Higher HD signifies higher degree of roughness.

3.3 Contour Signature:

It is defined as rough boundary is a significant feature for labeling a WBC nucleus as a blast cell. Along with the fractals contour signature method is also followed to measure the irregularity quantitatively. The nucleus boundary can be represented by a contour of dimension two. A better way of irregularity measurement of the contour is converting from coordinate based representation to distances from each contour point or edge pixels to a reference point. Since most nucleus have irregular shapes a convenient reference for the entire contour is the centroid or centre of mass. Euclidean distance measurement from the centroid to the contour points is described as follows:

1. Nucleus boundary pixel indices are obtained from the edge image which is obtained during HD measurement.
2. Centroid of the nucleus region is calculated using the edge pixels which represents a contour.

3. Euclidean distance is calculated from each boundary pixel to the centroid.
4. To measure the irregularity of the nucleus boundary variance (σ^2) of all the distances from the centroid obtained in step 3 is calculated.

$$\bar{x} = \frac{1}{M} \sum_{n=0}^{M-1} x(n), \bar{y} = \frac{1}{M} \sum_{n=0}^{M-1} y(n)$$

where (x,y) are the coordinates of the pixels along the contour and N is the total no of pixels on the contour.

3.4 Shape Features:

According to hematologist the shape of the nucleus is an essential feature for discrimination of blasts. Region and boundary based shape features are extracted for shape analysis of the nucleus. All the features are extracted from the binary equivalent image of the nucleus with none zero pixels representing the nucleus region. The quantitative evaluation of each nucleus is done using the extracted features under two classes i.e. region based and boundary based. Shape feature extraction technique is one of the key methods in feature extraction field. It can be classified into two groups, which include the region-based and contour-based methods. In the contour-based method, the shape features are calculating from the boundary of the shape only, while in the region-based method, the features are extracting from the whole region in the image. Shape features play an important role in acute leukemia cell detection The features are as follows:

3.4.1 Area:

The area was determined by counting the total number of nonzero pixels within the image region.

3.4.2 Perimeter:

It was measured by calculating distance between successive boundary pixels.

3.4.3 Compactness :

Compactness or roundedness is the measure of a nucleus The extent to which the shape is compact. Depending on the maturity and the type of the leucocytes, the shape of nucleus varies greatly. Mature leucocytes usually have more than two lobed nuclei with lobes connected by thin strands . In some special cases, the nucleus can have kidney bean shaped contours. Contrarily, leukemic cell nuclei are round in shape and exhibit higher overall

compactness than the nuclei of to mature cells. The compactness measure is denoted by the following formula .

$$compactness = \frac{perimeter^2}{area}$$

3.4.4 Solidity :

The ratio of actual area and convex hull area is known as solidity and is also an essential feature for blast cell classification. This measure is defined in .

$$solidity = \frac{area}{convex\ area}$$

3.4.5 Eccentricity :

This parameter is used to measure how much a shape of a nucleus deviates from being circular. It's an important feature since lymphocytes are more circular than the blast. To measure this a relation is defined in .

$$Eccentricity = \frac{\sqrt{a^2-b^2}}{a}$$

where a is the major axis and b is the minor axis of the equivalent ellipse representing the nucleus region.

3.4.6 Formfactor :

This is an dimensionless parameter which changes with surface irregularities and is defined as .

$$form\ factor = \frac{4 \times \pi \times area}{perimeter^2}$$

3.4.7 Elongation:

Abnormal bulging of the nucleus is also a feature which signifies towards leukemia. Hence nucleus bulging is measured in terms of a ratio called elongation.

$$elongation = \frac{R_{max}}{R_{min}}$$

3.4.8 Nuclear- cytoplasmic ratio:

It is defined as the ratio of the area of cell nucleus to the cytoplasm area. This measure is a very important feature for the assessment of the maturity of the cell

$$nuclear\ cytoplasmic\ ratio = \frac{area\ of\ the\ nucleus}{cytoplasm\ ratio}$$

3.5 Color Feature Extraction:

Since color is an important feature that human perceive while visualizing it is

considered for extraction from nucleus regions. Hence for each nucleus image the mean color values in RGB and HSV color spaces are obtained. It is defined based on a particular model or color space. There are a number of color spaces such as LUV, HSV, RGB and others that have been used to make extracting the features easier. Therefore, Color features are useful for extraction the information from the blood cells image for better classification.

3.6 Texture features:

Texture feature is a useful characterization for an image, where pixel properties are used to measure the color in the image while the group of pixels is used to measure a texture. Two techniques are used to extract the texture features based on the domain, which include the spectral texture feature extraction and spatial texture feature extraction. In the spectral texture feature extraction approach, an input image is transformed into the frequency domain, and the texture feature from the transformed image is then calculated. While in the spatial approach, the extracting features have been accomplished by computing the statistics of a pixel in the original image domain. In the blood smear images, the texture is an important feature that used to identify the blast cells by analyzing that features to get the ROI, which can help to obtain better classification.

Other descriptors used for the identification of blast cells are based on changes in the nuclear chromatin pattern reflecting DNA formation and on cytoplasmic changes. To capture the important information of the structural arrangement of the nucleus and the entire cell, two types of statistical measures were used. The first-order texture measures are based on the histogram of the greyscale image, e.g., the cytoplasm and the nucleus mean color, and the second-order statistical measures are derived from the gray level co-occurrence matrix , which carries information about the spatial relationships of the image pixels. Nucleus texture measurements were performed on gray scale version of the nucleus images. These features were computed from the co-occurrence matrices for each nucleus image. This includes

3.6.1 Homogeneity :

It is a measure of degree of variation. Homogeneity reflects the homogeneity of image textures and scaled the local changes of image texture. High values of homogeneity denote the absence of intra-regional

$$Homogeneity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$$

3.6.2 Energy :

It Is used to measure uniformity. Same values of all co-occurrence matrix resulted in small energy profiles; on the contrary, high energy might be expected in case of unequal values.

$$Energy = \sum_{i,j=0}^{N-1} (P_{ij})^2$$

3.6.3 Correlation :

This represents correlation between pixel values and its neighborhood. Correlation reflects the consistency of image texture.

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

3.6.4 Entropy :

Usually used to measure the randomness. Entropy reflects the non-uniformity and complexity of image texture.

$$Entropy = \sum_{i,j=0}^{N-1} -\ln(P_{ij})P_{ij}$$

3.7 Classification:

Classification is the task of assigning to the unknown test vector, a label from one of the known classes. Since the patterns are very close in the feature space, support vector machines (SVM) are employed for classification. SVM is a powerful tool for data classification based on hyper plane classifier . This classification is achieved by a separating surface (linear or non linear) in the input space of the data set. They are basically two class classifiers that optimize the margin between the classes . The classifier training algorithm is a procedure to find the support vectors. Relevant extracted features as described in Section II-F are used as input to the SVM.

The stage of classification is one of the most important stages in image processing and machine learning techniques, and it is an indemand field in this area. Classification is used to assign and classify a set of unclassified data. There are two types of classifiers that are known as supervised and unsupervised classifications. In the supervised classification, the set of possible

results or classes are known in advanced. While in the unsupervised classification, the set of classes are unknown in advance. Many methods can be used to form a classification of data that are known as the classifiers. These classifiers can be used to classify objects types, such as support vector machine (SVM), Artificial Neural Network (ANN), Random forest (RF), KNN (*K*-Nearest Neighbor), Naive Bayes (NB), Multilayer Perceptron (MLP),Hybrid and others, as summarized . Once the features are selected and extracted from the segmented image, the object's type is recognized and determined through this stage. Image classification is a image processing method which to distinguish between different categories of objectives according to the different features of images. It is widely used in pattern recognition and computer vision. Support Vector Machine (SVM) is a new machine learning method base on statistical learning theory, it has a rigorous mathematical foundation, built on the structural risk minimization criterion. We design an image classification algorithm based on SVM in this paper, use Gabor wavelet transformation to extract the image feature, use Principal Component Analysis (PCA) to reduce the dimension of feature matrix.

There are various approaches for image classification. Most of classifiers, such as maximum likelihood, minimum distance, neural network, decision tree, and support vector machine, are making a definitive decision about the land cover class and require a training sample. On the contrary, clustering based algorithm, e.g. K-mean, K-NN or ISODATA, are unsupervised classifier, and fuzzy-set classifier are soft classification providing more information and potentially a more accurate result. Besides, the knowledge based classification, using knowledge and rules from expert, or generating rules from observed data, is becoming attractive. We refer to D. Lu and Q. Weng for complete treatment of image classification approaches. In recent years, combine of multiple classifiers have received much attention. Some researchers combine NN classifier , SVM classifier or AdaBoost classifier for image classification. The aim of this paper is bring together two areas in which are Artificial Neural Network (ANN) and Support Vector Machine (SVM) applying for image classification.

3.7.1 Support vector machine (SVM):

In machine learning and image processing techniques, the support vector machine (SVM) also known as the support vector network . It is a supervised learning model that analyzes the data to use for both classification and regression tasks. The main objective of the SVM algorithm is to

find a hyper-plane in N number of features that classify the data points. The SVM classifier separates the classes based on the labeled training data. The main objective of the separate classes is to find the maximum margin, i.e., find the maximum distance between the classes. The maximizing distance provides a capability that can classify future data points with more confidence. Support Vector Machine (SVM) is a relatively simple Supervised Machine Learning Algorithm used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line.

In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. The SVM classifier is seeking to find the compromise between the complexities models according to the training data (limited sample information). This type of classifier is suitable for small samples of circumstances. The SVM classifier algorithm provides four types of the kernel that include Sigmoid, Radial Basis Function, Polynomial and Linear. The main reason for selecting SVM for detection of leukemia cell is the efficiently classify between the normal and abnormal cells. A support vector machine (SVM) is supervised algorithm used for many classification and regression problems, including signal processing medical applications, natural language processing, and speech and image recognition. The first use of SVMs in the medical field was based on cancer recognition, which was an image-based application. But the algorithm took its flight in the field when it was first used in the protein analysis tasks. We all know that human-based proteins are very delicate structures and are prone to too much noise as well as errors while using the algorithms for recognition. Another field there is the remote homology which uses SVMs to the fullest. This is where the analysis is dependent on how the protein sequences are modeled. Support vector machine algorithm is the most primarily and also the most well-known classifier. It is one of the best-known machine algorithms for effectively classifying data points and helps in the easy creation and separation of classes. They are beneficial when there are a larger number of variables. Such support vector machine example can be said for text clarification from a bag of words model. Non-linear kernels show promising and effective performance in most scenarios and often are head to head with other random forests. Also, they are useful particularly when the need is to classify data by rank or commonly known as ordinal classification and are widely implemented in

“learning to rank” algorithms.

The objective of the SVM algorithm is to find a hyperplane that, to the best degree possible, separates data points of one class from those of another class. “Best” is defined as the hyperplane with the largest margin between the two classes, represented by plus versus minus in the figure below. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points. Only for linearly separable problems can the algorithm find such a hyperplane, for most practical problems the algorithm maximizes the soft margin allowing a small number of misclassifications. Support vectors refer to a subset of the training observations that identify the location of the separating hyperplane. The standard SVM algorithm is formulated for binary classification problems, and multiclass problems are typically reduced to a series of binary ones. Digging deeper into the mathematical details, support vector machines fall under a class of machine learning algorithms called kernel methods where the features can be transformed using a kernel function. Kernel functions map the data to a different, often higher dimensional space with the expectation that the classes are easier to separate after this transformation, potentially simplifying a complex non-linear decision boundaries to linear ones in the higher dimensional, mapped feature space. In this process, the data doesn’t have to be explicitly transformed, which would be computationally expensive. This is commonly known as the kernel trick.

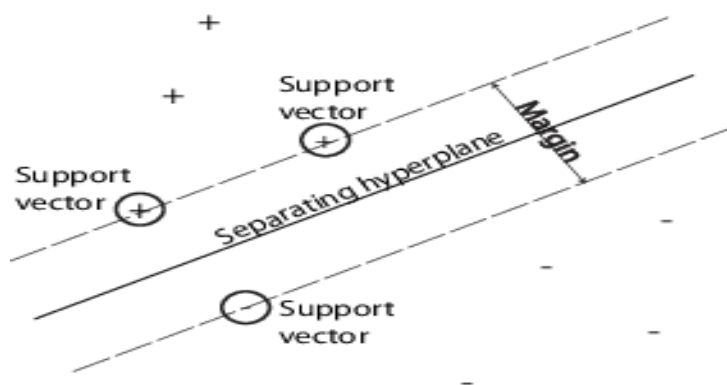


Fig 3.3 Support Vector Machine

- ▶ An SVM classifies data by finding the best hyperplane that separates all data points of one

class from those of the other class.

- ▶ The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points.
- ▶ The support vectors are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab. The following figure illustrates these definitions, with + indicating data points of type 1, and – indicating data points of type –1.

CHAPTER IV

THE MATLAB

4.1 Introduction:

MATLAB is a programming language developed by MathWorks. It started out as a matrix programming language where linear algebra programming was simple. It can be run both under interactive sessions and as a batch job. It allows matrix manipulations; plotting of functions and data; implementation of algorithms; creation of user interfaces; interfacing with programs written in other languages, including C, C++, Java, and FORTRAN; analyze data; develop algorithms; and create models and applications. It has numerous built-in commands and math functions that help you in mathematical calculations, generating plots, and performing numerical methods.

4.2 Features of MATLAB:

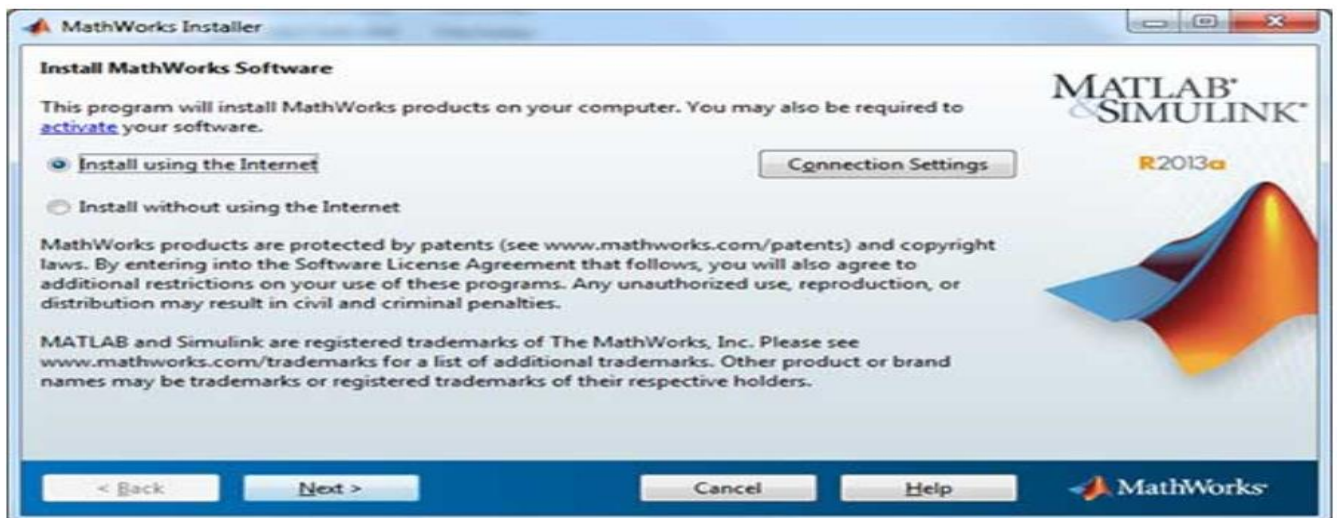
- It is a high-level language for numerical computation, visualization and application development.
- It also provides an interactive environment for iterative exploration, design and problem solving.
- It provides vast library of mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration and solving ordinary differential equations.
- It provides built-in graphics for visualizing data and tools for creating custom plots.
- MATLAB's programming interface gives development tools for improving code quality maintainability and maximizing performance.
- It provides tools for building applications with custom graphical interfaces.
- It provides functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET and Microsoft Excel.

4.3 Uses of MATLAB:

MATLAB is widely used as a computational tool in science and engineering encompassing the fields of physics, chemistry, math and all engineering streams. It is used in a range of applications including Signal Processing and Communications, Image and Video Processing, Control Systems Test and Measurement Computational Finance.

4.4 Local Environment Setup:

Setting up MATLAB environment is a matter of few clicks. The installer can be downloaded from [here](#). MathWorks provides the licensed product, a trial version and a student version as well. You need to log into the site and wait a little for their approval. After downloading the installer the software can be installed through few clicks.



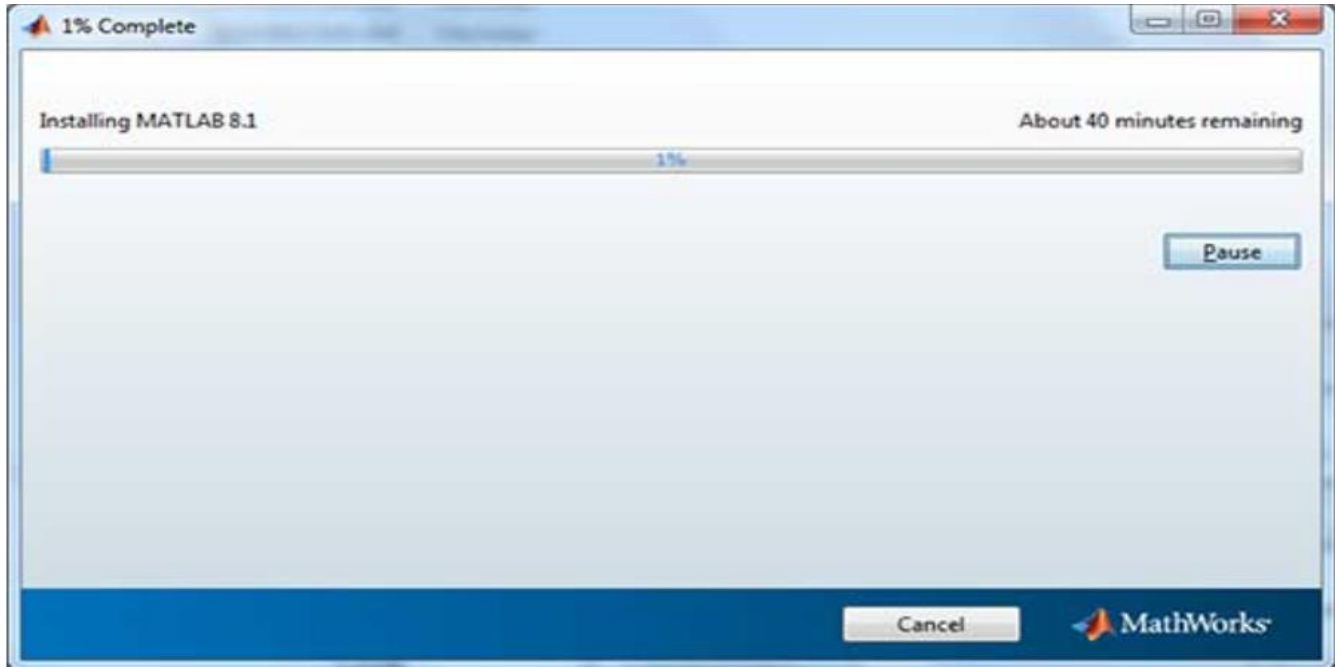


Fig 4.1 Setup windows

Understanding the MATLAB Environment

MATLAB development IDE can be launched from the icon created on the desktop. The main working window in MATLAB is called the desktop. When MATLAB is started, the desktop appears in its default layout –

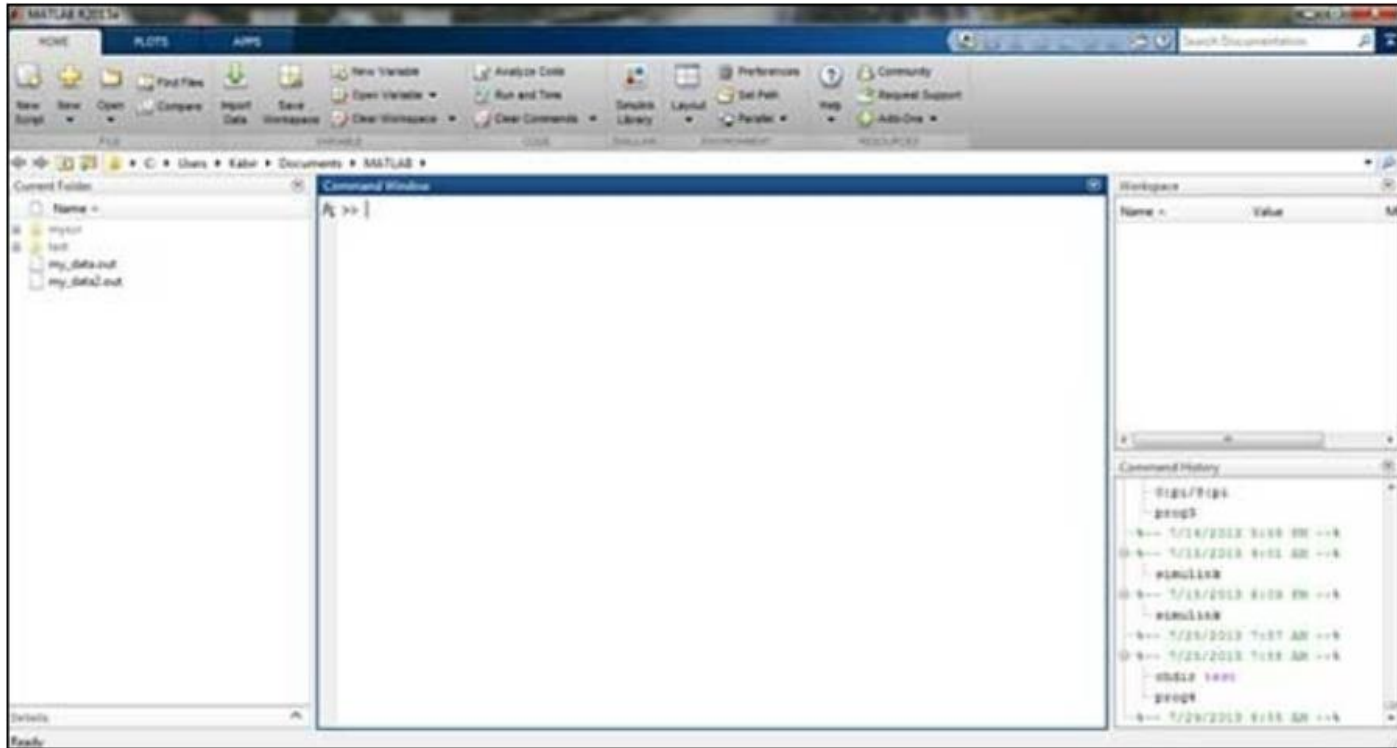


Fig 4.2 Default layout window

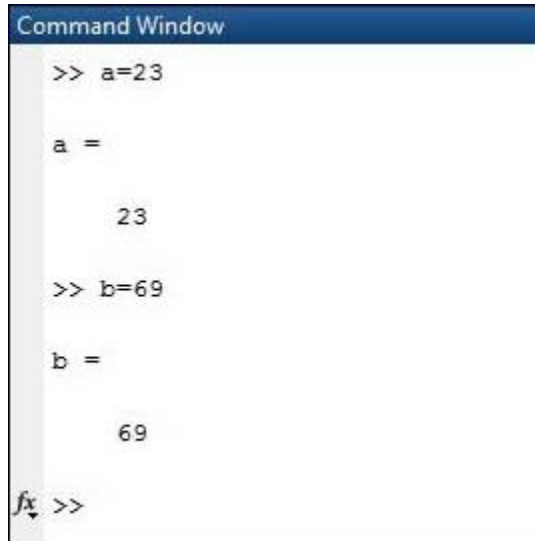
The desktop has the following panels –

Current Folder – This panel allows you to access the project folders and files.



Fig 4.3 Current folder window

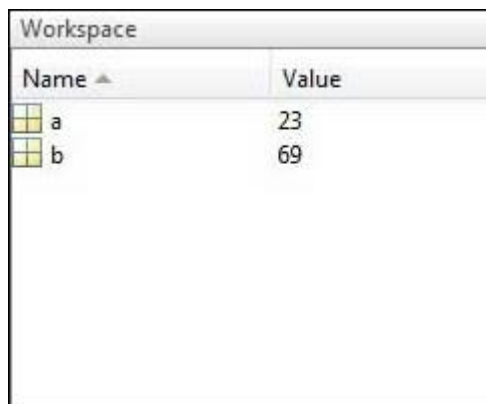
Command Window – This is the main area where commands can be entered at the command line. It is indicated by the command prompt (>>).



```
Command Window
>> a=23
a =
    23
>> b=69
b =
    69
fx >>
```

Fig 4.4 Command Window

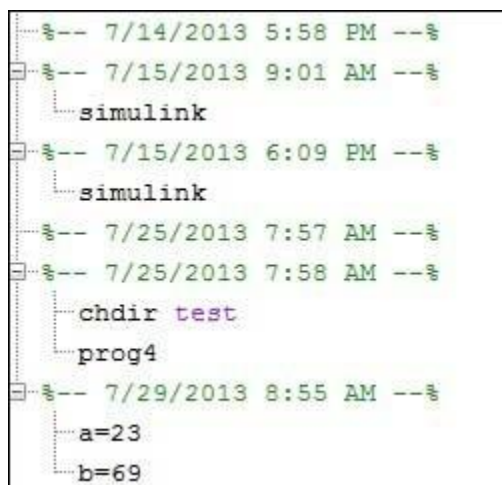
Workspace – The workspace shows all the variables created and/or imported from files.



Name ^	Value
a	23
b	69

Fig 4.5 Workspace Window

Command History – This panel shows or return commands that are entered at the command line.



```
--- 7/14/2013 5:58 PM ---
--- 7/15/2013 9:01 AM ---
simulink
--- 7/15/2013 6:09 PM ---
simulink
--- 7/25/2013 7:57 AM ---
--- 7/25/2013 7:58 AM ---
chdir test
prog4
--- 7/29/2013 8:55 AM ---
a=23
b=69
```

Fig 4.6 Command History

4.5 The M Files:

MATLAB allows writing two kinds of program files –

Scripts – script files are program files with **.m extension**. In these files, you write series of commands, which you want to execute together. Scripts do not accept inputs and do not return any outputs. They operate on data in the workspace.

Functions – functions files are also program files with **.m extension**. Functions can accept inputs and return outputs. Internal variables are local to the function.

You can use the MATLAB editor or any other text editor to create your **.mfiles**. In this section, we will discuss the script files. A script file contains multiple sequential lines of MATLAB commands and function calls. You can run a script by typing its name at the command line.

- Creating and Running Script File
- To create scripts files, you need to use a text editor. You can open the MATLAB editor in two ways –
- Using the command prompt
- Using the IDE
- If you are using the command prompt, type **edit** in the command prompt. This will open the editor. You can directly type **edit** and then the filename (with .m extension)
- If you are creating the file for first time, MATLAB prompts you to confirm it. Click Yes.

- After creating and saving the file, you can run it in two ways –
- Clicking the **Run** button on the editor window or
- Just typing the filename (without extension) in the command prompt: `>> prog1`
- MATLAB does not require any type declaration or dimension statements. Whenever MATLAB encounters a new variable name, it creates the variable and allocates appropriate memory space. If the variable already exists, then MATLAB replaces the original content with new content and allocates new storage space, where necessary.

4.6 Getting Help:

The arch way to get advice online is to use the MATLAB advice browser opened as abstract window either by beating on the catechism mark attribute() on the desktop toolbar, or by accounting advice browser at the alien in the command window. The advice

Browser is a web browser chip into the MATLAB desktop that displays a Hypertext Markup Language (HTML) document. The Advice Browser consists of two panes, the advice navigator pane, acclimated to accretion information, and the affectation pane, acclimated to appearance the information Self explanatory tabs added than navigator area are acclimated to accomplish a search.

4.7 Matlab Using Image Processing:

Image Processing Toolbox™ provides a comprehensive set of reference-standard algorithms and workflow apps for image processing, analysis, visualization, and algorithm development. ... You can interactively segment image data, compare image registration techniques, and batch-process large data sets.

4.7.1 Basic Image Import, Processing, and Export:

Step 1: Read and Display an Image: Read an image into the workspace, using the `imread` command. The example reads one of the sample images included with the toolbox, an image of a young girl in a file named `pout.tif`, and stores it in an array named `I`. `imread` infers from the file that the graphics file format is Tagged Image File Format (TIFF). `I = imread('pout.tif')`

Display the image, using the `imshow` function. You can also view an image in the Image Viewer app. The `imtool` function opens the Image Viewer app which presents an integrated environment for displaying images and performing some common image processing tasks. The Image Viewer app provides all the image display capabilities of `imshow` but also provides access to several other tools for navigating and exploring images, such as scroll bars, the Pixel Region tool, Image Information tool, and the Contrast Adjustment tool `imshow(I)`.



Fig 4.7 Image 1

Step 2: Check How the Image Appears in the Workspace: Check how the `imread` function stores the image data in the workspace, using the `whos` command. You can also check the variable in the Workspace Browser. The `imread` function returns the image data in the variable `I`, which is a 291-

by-240 element array of uint8 data.

whos I

Name	Size	Bytes	Class	Attributes
I	291x240	69840	uint8	

Step 3: Improve Image Contrast:View the distribution of image pixel intensities. The image pout.tif is a somewhat low contrast image. To see the distribution of intensities in the image, create a histogram by calling the imhist function. (Precede the call to imhist with the figure command so that the histogram does not overwrite the display of the image I in the current figure window.) Notice how the histogram indicates that the intensity range of the image is rather narrow. The range does not cover the potential range of [0, 255], and is missing the high and low values that would result in good contrast.

figure

imhist(I)

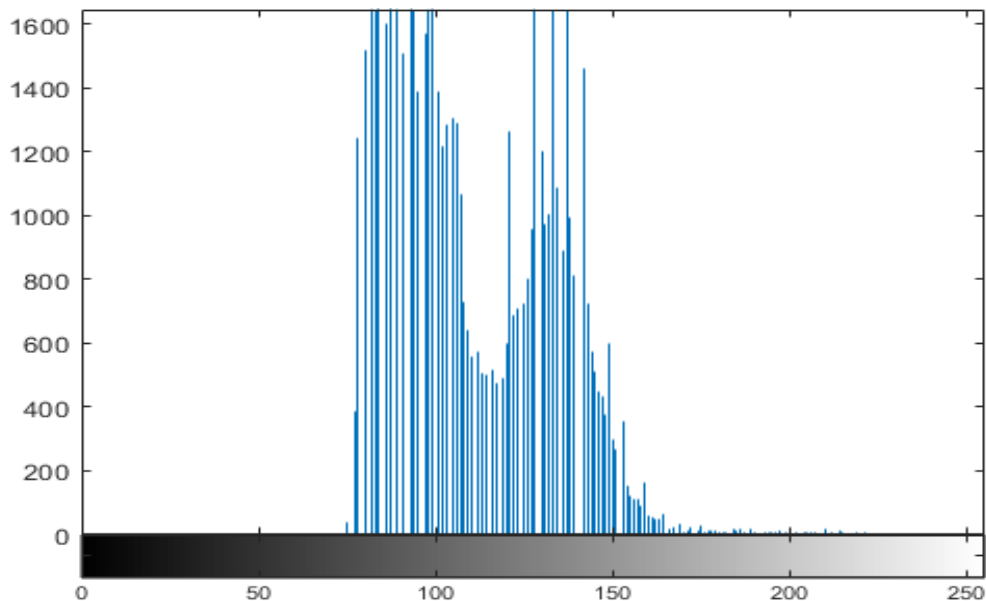


Fig 4.8 Histogram of Image1

Improve the contrast in an image, using the histeq function. Histogram equalization spreads the intensity values over the full range of the image. Display the image. (The toolbox includes several other functions that perform contrast adjustment, including imadjust and adapthisteq, and

interactive tools such as the Adjust Contrast tool, available in the Image Viewer.)

```
I2 = histeq(I);  
figure  
imshow(I2)
```



Fig 4.9 Image2

Call the `imhist` function again to create a histogram of the equalized image `I2`. If you compare the two histograms, you can see that the histogram of `I2` is more spread out over the entire range than the histogram of `I`.

```
figure  
imhist(I2)
```

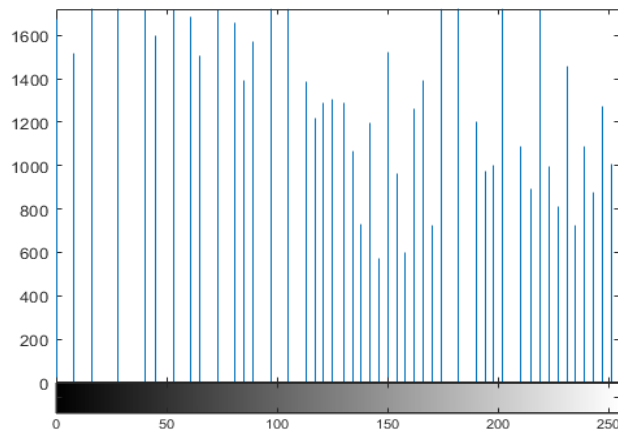


Fig 4.10 Histogram of Image2

Step 4: Write the Adjusted Image to a Disk File: Write the newly adjusted image I2 to a disk file, using the `imwrite` function. This example includes the filename extension `'.png'` in the file name, so the `imwrite` function writes the image to a file in Portable Network Graphics (PNG) format, but you can specify other formats. `imwrite (I2, 'pout2.png');`

Step 5: Check the Contents of the Newly Written File: View what `imwrite` wrote to the disk file, using the `imfinfo` function. The `imfinfo` function returns information about the image in the file, such as its format, size, width, and height.

CHAPTER V

EXPERIMENTAL RESULTS

5.1 Introduction:

The proposed technique has been applied on 108 peripheral blood smear images obtained from two places as mentioned earlier. The superiority of the scheme is demonstrated with the help of an experiment.

5.2 Blood Smear Image Dataset:

The proposed system was trained as well as tested on a local dataset, which was provided by the Kaggle website. The anonymized dataset consists of 18 microscopic blood smear images obtained from patients without pathological findings and 13 blood smear images from patients with diagnosed ALL. On average, six blood smear images with a resolution of $4,080 \times 3,072$ were captured per patient. Since WBCs are distributed unevenly, with a predominance of large cells on the border and smaller cells in the center of the blood smear, systematic data acquisition was required. This was carried out by the meander inspection pattern, which allowed microscopic images to be captured from different consecutive locations, particularly from both edges and the center of the blood smear. All slides in the dataset were stained with Giemsa stain and were captured under the same lighting conditions by an Olympus CX43 microscope under a magnification of 50 times with an oil immersion objective lens and an effective magnification of 500.

The manual examination of blood smear images was conducted by local domain experts. During this visual examination, the hematology specialists used several morphological criteria to distinguish between lymphoblasts and normal cells. The most significant criteria included the nucleus position and shape, chromatin structure, presence of nucleoli, nucleocytoplasmic ratio, size of the cell, and color or structure of the cytoplasm. Following the WHO classification system, ALL is divided into B-lymphoblastic leukemia/lymphoma, T-lymphoblastic leukemia/lymphoma, and acute leukemia of ambiguous lineage. Because, from a morphological point of view, there are no reproducible criteria to distinguish between B and T lineage lymphoblastic leukemia, ALL subtype classification is not considered in this study.

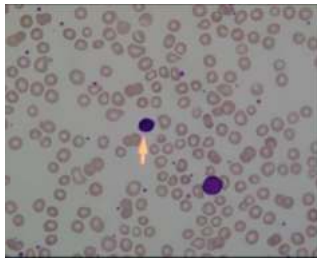
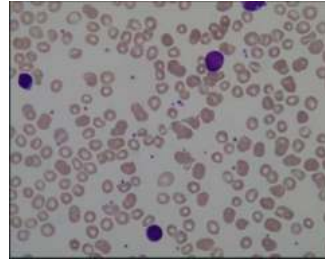
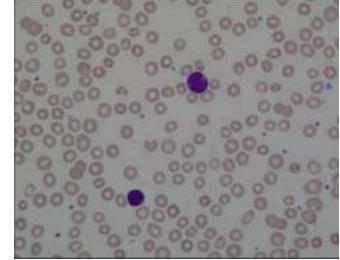


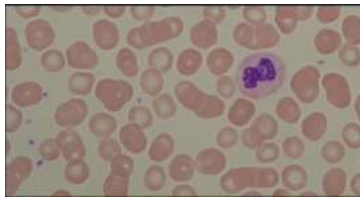
Fig (a)



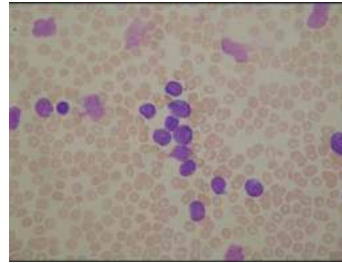
fig(b)



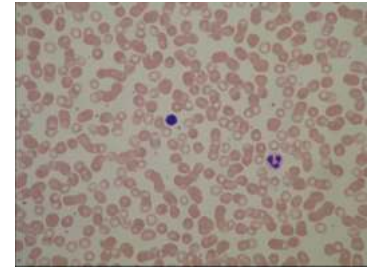
fig(c)



Fig(d)



fig(e)



fig(f)

Fig 5.1 Blood Smear Image dataset

The above image set has been designed for testing the performances of classification systems.

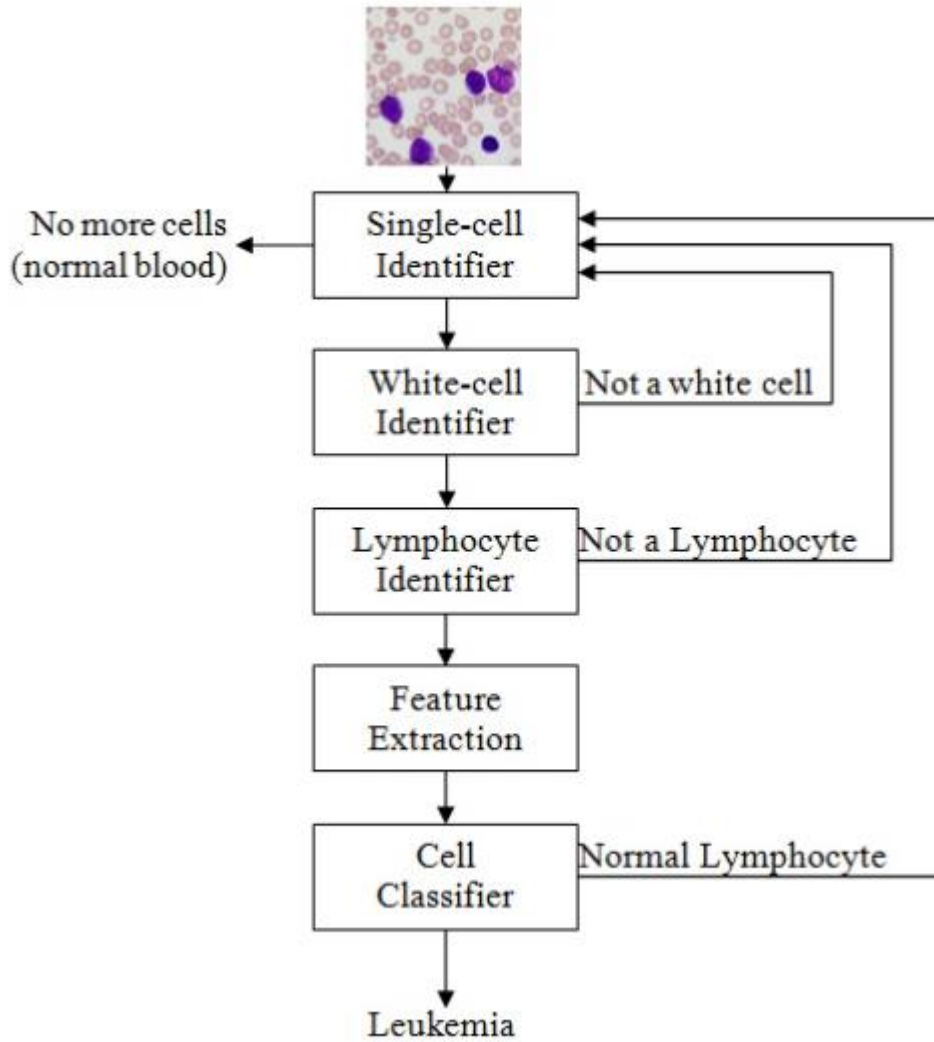
Examples of the image contained: healthy cells from non-ALL patients (a-c), probable lymphoblasts from ALL patients (d-f).

5.3 Morphological features of ALL blast cells:

The classification of the lymphocyte in microscope images is quite complex since even an expert operator can have dubs in classifying some lymphocyte cells. Actually, the morphological distinctive aspects of ALL blast and normal lymphocytes are very smooth.

5.4 Image processing:

The identification and classification of white blood blast cells had been tackled by a classic sequence of steps as shown in the next figure. a hierarchical classification approach can be followed, where the segmentation of white cells is achieved and then each single cell is classified after a feature extraction phase.



A microscopic blood image of size 183×275 (Fig. 1(a)) is considered for evaluation.

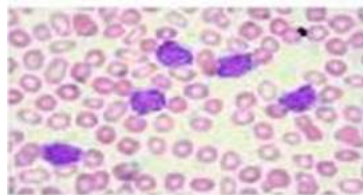


Fig 5.2 Original Image

The input image is processed sequentially. The aim of pre-processing is an improvement of the image data that suppresses unwilling distortions or enhances. The steps to be taken are :

- Read image.
- Resize image

- Remove noise (Denoise)
- Segmentation.
- Morphology (smoothing edges)

In the above figure a, b are the images which contains noise and c, d are the noiseless images which are obtained using median filtering technique. Median filtering is a nonlinear process useful in reducing impulsive, or salt-and-pepper noise.

Segmentation is performed in two stages for extracting WBC nucleus from the blood microscopic images using colour based clustering. The segmented output of cell nucleus image obtained after applying Fuzzy C-means clustering algorithm is shown in Fig.1(b).



Fig 5.3 Segmented output

The cluster image containing only blue nucleus is used to obtain the sub images containing a single nucleus each as shown in fig.2(a).



Fig 5.4 Separated Nucleus sub images using bounding box technique

Feature Extraction: Feature extraction in image processing is a technique of redefining a large set of redundant data into a set of features (or feature vector) of reduced dimension. This transformation of the input data into the set of features is called feature extraction [10]. In this three types of features are extracted i.e. fractal dimension, shape features including contour signature and texture. In addition color features are also extracted from the nucleus image.

1.Fractal Dimensions: Fractals have been used in medicine and science earlier for various quantitative measurement Hausdorff dimension is an essential feature for fractal geometry and will be an essential quantitative measure for cell boundary roughness measurement. Fig. 3(a) shows the nucleus boundary whose roughness is measured by Hausdorff Dimension using box counting method. A graphical plot showing box counting algorithm results is presented in Fig. 4. The straight line in the plot represents a line of best fit. And HD is obtained from the polynomial coefficients of the line of best fit. The HD is found to

be 1.0255 for the nucleus image shown in figure.3(a)



Fig 5.5 Nucleus Image and Edge image

Euclidean distance between the centroid and boundary pixels is measured as shown in Fig. 4. The variance of all the distances is found to be 31.6296.

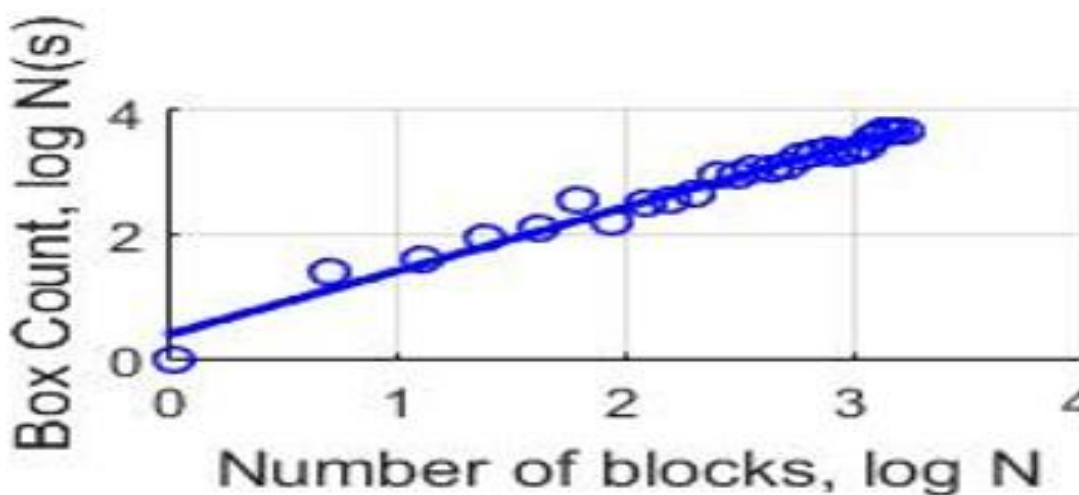


Fig 5.6 Box counting algorithm results

Similarly the shape features are also measured using the relations. The features are Area, Perimeter, Solidity, Compactness, Eccentricity, Form Factor. Few shape feature comparison between a mature lymphocyte and a lymphoblast is tabulated in Table 1.

Table-5.1: Results of Shape Features

Measure	Leukemia	Non-Leukemia
Area	457	101
Perimeter	85.3900	34.2720
Eccentricity	0.8841	0.9352
Compactness	0.8215	0.5302
Solidity	15.9550	11.6294
Formfactor	0.6991	0.6894

Color and texture features are also extracted for the image nucleus sample and recorded. Few texture measurements are tabulated in Table 2.

Table-5.2: Results of Texture Features

Measure	Leukemia	Non-Leukemia
Homogeneity	0.6726	0.6654
Energy	0.1978	0.2125
Entropy	0.1698	0.1926

Among all the features the most relevant features are selected and used to train the SVM.

Support vector machines (SVM) are employed for classification. It is a powerful tool for data classification based on hyper plane classifier . This classification is achieved by a separating surface (linear or non linear) in the input space of the dataset. They are basically two class classifiers that optimize the margin between the classes. The classifier training algorithm is a procedure to find the support vectors. Relevant extracted features as described in Section II-F are used as input to the SVM.

CHAPTER VI

ANALYSIS

Our goal is to obtain a new set of features for better classification. We have tried to identify such features which are basically followed by hematologists. The results obtained in terms of features are also verified by an expert. For training we considered out of 108 blood cell images, we have extracted 45 different types of leukemia cells and non leukemia cells. 90 images are used for feature extractions. For a better classification between leukemia and non leukemia we have extracted 9 different features from this images.

The feature extracted from 25 leukemia and non leukemia cells are used for training and remaining are used for testing. We considered the output of the leukemia data as 1 and non leukemia data as 0. The advantage of the proposed scheme over existing schemes is that we are considering smear images with many lymphocytes. Existing schemes mostly consider those images which only have one lymphocyte under the field of view.

Images from microscope are usually difficult and always require human intervention which is not desirable in an automated system. The proposed scheme using the new features is certainly aiming one step ahead towards automated system. The features extracted in our proposed scheme from the available images were used for training SVM classifier and an accuracy of 94.7368% was observed.

CONCLUSION AND FUTURE WORK

A two stage WBC nucleus segmentation of stained blood smear images followed by relevant feature extraction for leukemia detection is the main theme of the project. The project mostly concentrates on measuring nucleus boundary irregularities using two methods i.e. hausdorff dimension and contour signature. Along with this shape, color and texture features are also considered for better detection accuracy. Leukemia detection with the proposed features were classified with SVM classifier.

Further research into this area, like classification of lymphoblast into various subtypes, can be taken into consideration because of the obtained results. Also, different techniques can be innovated or improved for touching cells, leukemia type classification and image segmentation without staining.

Table-7.1 Results for Features:

Measure	Leukemia	Non-Leukemia
Area	481	102
Perimeter	93.0480	36.4220
Eccentricity	0.8579	0.8146
Compactness	17.9999	13.0055
Solidity	0.9007	0.9273
Formfactor	0.6981	0.9662
Homogeneity	0.6822	0.6732
Energy	0.1928	0.2017
Entropy	0.1784	0.1992

REFERENCES

- [1] B. J. Bain . A Beginner's Guide to Blood Cells. Blackwell Publishing 2nd edition, 2004.
- [2] Childrens Hospital of Wisconsin Website. <http://www.chw.org>
- [3] C. Haworth, A. Hepplestone, P. Morris Jones, R. Campbell, D. Evans, M. Palmer. Routine Bone Marrow Examination in the Management of Acute Lymphoblastic Leukaemia of childhood. . Journal of Clinical Pathology, 34: 483 – 485, 1981.
- [4] N. Sinha and A. G. Ramakrishnan. Automation of Differential Blood Count. In Proceedings Conference on Convergent Technologies for AsiaPacific Region, 2:547 – 551, 2003.
- [5] G. Ongun, U. Halici, K. Leblebicioglu, V. Atalay, M. Beksac, and S. Beksac, An Automated Differential Blood Count System.. In Int. Conf. of the IEEE Engineering in Medicine and Biology Society , volume 3, pages 2583 - 2586, 2001.
- [6] Subrajeet Mohapatra Development of Impulse Noise Detection Schemes for Selective Filtering Master Thesis, National Institute of Technology Rourkela, 2008
- [7] Daniel Graves and Witold Pedrycz Fuzzy C-Means, Gustafson Kessel FCM, and Kernel Based FCM: A Comparative Study Analysis and Design of of Intelligent Systems using Soft Computing Techniques , Springer, 2007.
- [8] K. S. Ravichandran and B. Ananthi. Color Skin Segmentation using K–Means Cluster. International Journal of Computational and Applied Mathematics, 4(2):153 – 157, 2009.
- [9] A. K. Jain, Fundamentals of Digital Image Processing. Pearson Education, 1st Indian edition, 2003.
- [10] The Wikipedia the Free Encyclopedia Website. <http://en.wikipedia.org>
- [11] B. B. Mandelbrot. How long is the coast of Britain? Statistical self similarity and fractional dimension. Science, 156:636 – 638, 1967.
- [12] B. T. Milne. Measuring the fractal geometry of landscapes. Applied Mathematics and Computation, 27:67 – 79, 1988.
- [13] A. P. Pentland. Fractal based description of natural scenes . IEEE Transactions on Pattern Analysis and Machine Intelligence, 6:661 – 674, 1984.
- [14] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

