BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Inspiring Excellence

DETECTION OF ACUTE LYMPHOCYTIC LEUKEMIA (ALL) AND ITS TYPE BY IMAGE PROCESSING AND MACHINE LEARNING.

AUTHORS
Himadri Chowdhury- ID 14201008
Shounak Banik- ID 14201022
Arafat Hossain- ID 14201023
Md. Imran Khaled- ID 14201034

SUPERVISOR
Amitabha Chakrabarty, PhD
Associate Professor
Department of Computer Science and Engineering

A thesis submitted to the
Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh.

December 2018

To our supporting faculty body, seniors and well wishers of the department and beyond. Love goes out to our friends and family for giving us the latent energy we always needed to get through this.

Declaration

We hereby declare that the thesis titled 'Detection of Acute Lymphocytic Leukemia (ALL) and its Type by Image Processing and Machine Learning, a thesis submitted to the Department of Computer Science and Engineering of BRAC University in partial fulfillment of the Bachelor of Science in Computer Science and Engineering is our own work. The work has not been presented elsewhere for assessment. The materials used from other sources have been acknowledged.

Authors:

Himadri Chowdhury
Student ID: 14201008

Shounak Banik
Student ID: 14201022

Arafat Hossain
Student ID: 14201022

Md. Imran Khaled
Student ID: 14201034

Supervisor:

Amitabha Chakrabarty, PhD
Associate Professor, Department of Computer Science and Engineering,
BRAC University.

The thesis titled Detection of Acute Lymphocytic Leukemia (ALL) and its Type by Image Processing and Machine Learning
Submitted by:
Himadri Chowdhury Student ID: 14201008
Shounak Banik Student ID: 14201022
Arafat Hossain Student ID: 14201023
Md. Imran Khaled Student ID: 14201034

of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of B.Sc. Engineering in Computer Science and Engineering.

1. Amitabha Chakrabarty, PhD
Associate Professor,
Department of CSE,
BRAC University.

2. Md. Abdul Mottalib, PhD
Professor and Chairperson
Department of CSE,
BRAC University.

December 2018

Acknowledgements

We want to dedicate our acknowledgement of gratitude to our thesis supervisor Amitabha Chakrabarty, PhD, Associate Professor, Department of Computer Science and Engineering of BRAC University for his guidance for the completion of our thesis. We are grateful to Professor Dr. M A Khan, Head of Department, Department of Hematology, Dhaka Medical College Hospital, Dhaka, Bangladesh who helped us the most to collect the dataset of our thesis work. Last but not the least, we are thankful to CSE department, BRAC University for providing us the necessary equipment for the completion of this project.

Abstract

Cancer starts when cells of body begin to grow rapidly. Cells in nearly any part of the body can become cancer and can spread to other areas of the body. The origin of Chronic Lymphocytic Leukemia (CLL) in the bone marrow and causes the random growth of a large number of unnatural cells. The leukemia cells start in the bone marrow. By the time, access into the blood cells and cause fatal disease. Mainly, there exist 4 types of leukemia which are Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Lymphocytic Leukemia (CLL) and Chronic Myeloid Leukemia (CML). In this paper, we proposed to build a methodology to detect the Leukemia (Cancer) by the help of image processing and machine learning. We are using the two stage otsu-optimization approach algorithm, Lab color space algorithm and wrapper method. For image preprocessing to be fit in the classifiers Image to Feature Vector method and Label Encoding methods have been applied on the dataset. Furthermore, we applied various machine learning algorithms, Logistic Regression, Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbor (KNN) and from neural network algorithm Convolutional Neural Network (CNN) has been applied. We made an effort to build a comprehensive comparison among machine learning algorithms. Though it has been done in past research papers but in this paper we collected few image data from Dhaka Medical College and preprocessed it with another public image data set named ADL to attain at least a promising test accuracy. Moreover, in this research paper we tried to break a superstition of recent age which is Convolutional Neural Network (CNN) is the only appropriate model to train an image dataset. We implemented AdaBoost Classifier which has given 87% of test accuracy with a glimpse of high cross validation accuracy of 90%. We also brought Voting Classifier in process, mixing AdaBoost, Gaussian Naive Bayes, K-Nearest Neighbor (KNN) classifiers together has given 89% of test accuracy as much as like Convolutional Neural Network (CNN) 90%. Thus, we can conclude the debate that image dataset can be trained for pattern recognition with simple machine learning algorithm with the minimum computational cost with higher accuracy.

Table of contents

List of figures

List of tables

Nomenclature

Greek Symbols

β	An association between set of input and output values are termed as beta function
δ	Shows difference between two numbers
ε	The permitivity of free space
γ	A simply closed curve on a complex plane
θ	theta is used to represent an angle
ι	Unit imaginary number $\sqrt{-1}$
α	representing quantities such as angles

Other Symbols

exp	The exp stands for exponential
ln	The natural logarithm
$\subset$	The set of which all the elements are contained in another set
$\sum$	The process of summing something up

Acronyms / Abbreviations

ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
ANN	Artificial Neural Networks
CART	Classification and Regression Trees

CLL Chronic Lymphocytic Leukemia

CML Chronic Myeloid Leukemia

CNN Convolutional Neural Network

GLM Generalized Linear Model

KNN K- Nearest Neighbor Algorithm

ROC Receiver Operating Characteristic

SVM Support Vector Machine Algorithm

Chapter 1

Introduction

A multi-focus doctor's facility based review illustrative investigation of more than 5000 affirmed hematological disease cases in the middle of January 2008 to December 2012. A sum of 5013 patients aged between 2 to 90 years had been determined to have threatening hematological scatters. A 69.2% were males (n=3468) and 30.8% females (n=1545), with a male to female proportion of 2.2:1. The general middle age at analysis was 42 years. Acute myeloid leukemia was most regular (28.3%) with a middle age of 35 years, trailed by chronic myeloid leukemia with 18.2% (middle age 40 years), non-Hodgkin lymphoma (16.9%; middle age 48 years), acute lymphoblastic leukemia (14.1%; middle age 27 years), mutual myeloma (10.5%; middle age 55 years), myelodysplastic disorders (4.5%; middle age 57 years) and Hodgkin's lymphoma (3.9%; middle age 36 years). The slightest basic was chronic lymphocytic leukemia (3.7%; middle age 60 years). Beneath the age of 20 years, acute lymphoblastic leukemia was transcendent (37.3%), trailed by intense myeloid leukemia (34%). Endless lymphocytic leukemia and numerous myeloma had for the most part happened among more seasoned patients, matured 50-over [9]. We observed from the analysis that patients aged below 20 years are affected mostly with ALL which is why it is the main focus of the study.

1.1 Motivation and Background

Acute leukemias are threatening neoplastic ailments that emerge from either lymphoid (acute lymphoblastic/ lymphocytic/ lymphoid leukemia, or ALL) or myeloid (acute myeloid/

myelogenous/ myelocytic leukemia, or AML) cell lines. Acute leukemias are portrayed by the expansion of youthful, non-utilitarian cells in the bone marrow that are consequently discharged into the circulation system. ALL is the most widely recognized youth harm, while AML for the most part influences adults. The two maladies are related with genetic disorders, for example, Down disorder and past exogenous bone marrow harm. The over the top expansion of juvenile impacts in the bone marrow impedes all other cell lines, bringing about paleness, coagulating scatters, and expanded weakness to contaminations [2]. We will be focusing on ALL as it mainly occurs in children from 2-5 years and acute means fast growing so if not identified and treated in the right time the blast cell will spread quickly.

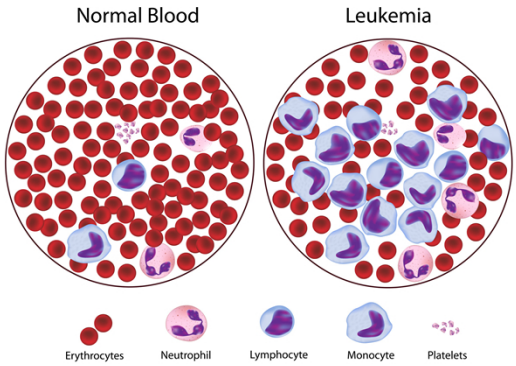


Fig. 1.1 Visual Difference Normal Blood Cell and Leukemia [33]

“Lymphoblastic” means it occurs from early immature forms of lymphocytes, a kind of white blood cell. Regularly, the leukemia cells attack the blood decently fast. They can likewise now and again spread to different parts of the body, including the lymph nodes, liver, spleen, central nervous system , and testicles [5, 21]. There are highly trained surgeons in Bangladesh for identifying the blast cell and no doubt they are good at what they do, but it takes a lot of time to identify it and anything can happen during that time to a patient. Moreover, while testing the surgeon has to be precise. There are quite a few steps for detecting ALL and there is a high possibility of making any slight mistake as we know the phrase “To err is human” [1]. As there are no digitized system in Bangladesh for identifying ALL so, to detect the symptoms of acute lymphocytic leukemia (ALL) or the presence of the immature lymphocytes in the blood , we have come up with a procedure which is more efficient, precise and less error prone. Moreover, the procedure we are following saves a lot of time and we are focusing on time because the faster we classify the ALL, the faster treatment can be started in turn there will be less loss of life.

Nowadays, the contribution of image processing and machine learning in medical science

is highly noticeable. Many medical research and inventions are created by the help of this sector of Computer Science. Medical images are considered as a vital tool to utilize for the diagnosis and analysis of many diseases such as breast, chest illness, blood disorder etc. The digital medical images provide opportunities for further analysis for more accuracy in diagnosis [46].
This thesis project is an attempt to apply image processing and machine learning in the area of medical science. The focus of this work is on developing a methodology to detect "Acute Lymphoblastic Leukemia (ALL)" based on the size and other features of the nucleus. We implemented a couple of classification methods to detect the abnormal cell of nucleus in the blood sample. However, we found out the maximum accuracy rate to extract the features and classify the nucleus from Convolutional Neural Network (CNN). So that, we have used Convolutional Neural Network (CNN) as a classification method. The digitizing experiment of detecting ALL has been implemented by researchers of other countries. Therefore, our goal is to help the children of our society through our integrated system.

1.2 Methods

The momentous portions of the project which vivid our proposed model are the machine learning methods or algorithms, we used in this thesis. These are Logistic Regression Algorithm, Decision Tree, Naive Bayes Theorem, Random Forest Algorithm, AdaBoost Classifier Algorithm, K-Nearest Neighbors (KNN) Algorithm and Convolutional Neural Network (CNN). All the algorithms we have used for the feature extraction and classification of nucleus of the blood sample to identify the abnormal blood cell. We obtained different results by applying different algorithms with our datasets collected from Dhaka Medical College Hospital and the online resource. Finally, chose the algorithm for continuing the process which performs swiftly, efficiently and give the most accurate result.

1.3 Objective

The main objective of this project is to develop a system that can detect the presence of the abnormal cell of blood in a human body which indicates the sign of cancer cell. In order to detect the lymphocytic leukemia we decided to use several methods in a proper sequence which will give a significant result about the patient . The Python language has been used to develop the project overall. There is no needed of any extra component without a desktop computer and internet connection. We are looking forward to see the success of this project

in the sector of modern medical science.

1.4 Overview of Contents

The rest of the dissertation is organized as follows:

Chapter-2: Literature Review:

The previous study related to our proposed model which we have done, described elaborately in this chapter.

Chapter-3: Data Collection and Project overview:

In this chapter, we mentioned about how we collected the dataset (microscopic images of blood sample of leukemia affected patients and normal patients) and the works related to this project. The project overview also has been discussed here. How we want to develop the system, what we want to implement and what results we expect to get has been discussed in short.

Chapter 4: Algorithms:

The detailed discussion of the algorithms, we used in our thesis are included in this chapter.

Chapter 5: Image Processing:

This chapter contains discussions about the overall process elaborately of preprocessing of image, segmenting of the microscopic image of blood by showing of several steps and the extraction of the feature of nucleus.

Chapter 6: Result and Analysis:

The Accuracy, Recall, Precission and Validation rate has been checked by different algorithms and got the maximum rate. The details of the observation and the results obtained have been discussed in this chapter. Future perspective of our proposed model is also included in this section.

Chapter 7: Conclusion:

In this last chapter the main results of the system has been summarized and some concluding remarks has been provided.

Chapter 2

Literature Review

To accustom with the methodologies of existing works, we studied a lot of papers related to our topic. As blood cancer is an uprising issue and medical procedure is too sensitive and time consuming to detect any blast cell, so there was no doubt that Scholars would come up with the idea of computer aided diagnosis to detect blast cell. The most common algorithm in this approach consists of several firm steps: image pre-processing, segmentation, feature selection or extraction, classification, and evaluation.

Rehman A et al. proposed a vigorous segmentation and profound deep learning procedures with the convolutional neural system used to prepare the model on the bone marrow pictures to accomplish precise arrangement results. Exploratory outcomes hence got and contrasted and the consequences of different classifiers Naïve Bayesian, K- Nearest Neighbor Algorithm (KNN), and Support Vector Machine (SVM). Trial results uncover that the proposed strategy accomplished 0.9778 exactness [32].

Sarmad Shafique et al. discussed about the whole process from pre-processing to classification and described all different approach of every step. For the sake of learning and understanding the different methods of classification, researchers have shown comparison between the methods even showed the accuracy of each and every method [36].

Sachin Kumar et al. describe the method of automatic detection of Acute Leukemia by basic enhancement, morphology, filtering and segmenting technique to extract region using K-Means Algorithm which has an accuracy of 0.9280 by testing with 60 sample data [18].

Another research of detection and classification of Acute Leukemia supervised by Giao N. Pham et al. used Convolutional Neural Network(CNN) as a mechanism for classification and extraction of features from raw images. In this research they postulated a network containing 4 layers. The first three layers used for detecting features and the last layer containing 2

neral (Fully connected and Softmax) used for identifying the features. The experiment was executed in Matlab and it had a very good accuracy of 0.9643 [42].

Preeti Jagadev et al. inscribed a paper on the "Detection of leukemia and its types using image processing and machine learning". Initially, the smear images are segmented using 3 algorithms which are k-means clustering, Marker controlled watershed and HSV color based segmentation algorithm. The difference between the morphological components of normal and Leukemic lymphocytes is worthy of attention, therefore, heterogeneous features are extracted from the segmented lymphocyte images. For the classification process machine learning classifier is used which is Support Vector Machine(SVM) and it aims to identify which type of leukemia it is among the 4 types [13].

Himali P. Vaghela et al. have done a research with an objective of detection of leukemia affected cells and count it. To detect the immature cells they implemented couple of methods in their paper. These are Watershed Transform, K means Clustering Algorithm, Histogram Equalization Linear Contrast Stretching, Shape based features. After executing all the techniques they discovered that shape based features are manoeuvred for better result and accuracy. The method is used to detect different shapes like circle, rectangle, ellipse, square etc. So to identify geometrical shapes of cells, the shape based features are very efficient and promising to detect different type of cells and their shapes. Shape based features also manifests that it has the highest accuracy of 0.9780 among the other three methods for counting leukemic cells [44].

Further work similar to our proposed model was introduced by T. T. P. Thanh et al. in 2018. They used Convolutional Neural Network (CNN) based method to show the difference between the normal and abnormal blood cell images. Conventionally the architecture of CNN comprise of three layers convolutional layer, pooling layer, and fully connected layer. Here, the researchers use a network containing 7 layers. The first 5 layers are executed for feature extraction and the last 2 layers (fully connected and softmax) used for the classification of extracted features. Significantly, the proposed model attainted an accuracy of 0.966 [43].

A cost efficient solution to detect the presence of abnormal growth of white blood cell (WBC) has introduced by Subrajeet Mohapatra et al. They used a Fuzzy Clustering based two stage color segmentation strategy for separating the leukocytes or white blood cell (WBC) from the other blood component. They implemented Hausdorff Dimension and Contour Signature to classify a lymphocytic cell nucleus and after that they used Support Vector Machine (SVM) to define the type of leukemia [23].

Nimesh Patel et al. founded a research of the automated detection of leukemia using microscopic blood cell. K-mean clustering is implemented for the detection of white blood cells. Histogram equalization is generated and then Zack algorithm is applied for grouping

the lymphocytes and myelocytes. For further steps the grouped lymphocytes need to be identified and rejected because the nucleus are appended in adjacent cells for which the exact area of the nucleus cannot be calculated. So, they used roundness measure to identify the grouped lymphocytes. Further, the image is cleaned and feature extraction is done so that it can be compared with the standard values. For the classification part machine learning algorithm "Support Vector Machine"(SVM) is used which manifest an accuracy 0.9357 [26].
Indira P. et al. in 2016 presented a research paper where they proposed a model to detect Acute Myeloid Leukemia (AML) which is also fast growing but the contrast is that it occurs in adults. K-means clustering is implemented for segmentation. For feature extraction both the spatial and spectral features are extracted. Spectral features are much more reliable and stable but in spectral data much of the information is replicated from image to image which makes classification insignificant. Thus to overcome the problem genetic algorithm is used to optimize the spectral features. Further, for classification linear Support Vector Machine was applied [12].
Luis Henrique Silva et al. deployed a strategy to detect leukemia automatically and in an efficient way. They used the Convolutional Neural Networks (CNNs) to extract the features from a blood smear and to classify the type of leukemia they used Support Vector Machine, Multilayer Perceptron and Random Forest. They obtained an accuracy rate of 1.00 [45].
Himani Sharma et al. published a journal on Decision Tree Algorithms where they have done a broad discussion about the different type of Decision Tree Algorithm, the characteristics of these algorithms, pros and cons and also the challenges generally we face in the implementation [37].
Keiron Teilo O'Shea et al. introduces the Convolutional Neural Networks Algorithm where they discussed about different papers about Convolutional Neural Networks and the techniques developed recently which are using in different models [25].
An article about the Logistic Regression and the pattern for the application of the logistic methods used by the researchers in testing is introduced by Chao-Ying Joanne Peng et al. They also showed the mathematical equations in details of the Logistic Regression in this article [28].
Eesha Goel et al. has formed an journal with a discussion about an ensemble Machine Learning Technique named Random Forest Algorithm introduced by Brieman in 2001 where the implementations in various sectors of this algorithm has presented also [6].
A significant study has made under a popular classification method named Bayes Classification method by Harshad M. Kubade. In his paper, he discussed about the 3 types of method of Bayes Method. Among the types, we have decided to use Naive Bayes Classifier in our project and is a probabilistic classifier [17].

Anna Jurek et al. has introduced a research on the well-known techniques which are using the most in Ensemble-Based Classification techniques. They also discussed about the modifications of these techniques, the drawbacks and the reviews about these techniques particularly [15].
Sadegh et al. proposed a method for classifying objects based on closest training examples in the feature space.It simply retains the entire training set during learning and assigns to each query a class represented by the majority label of its k-nearest neighbor in the training set [11].

Chapter 3

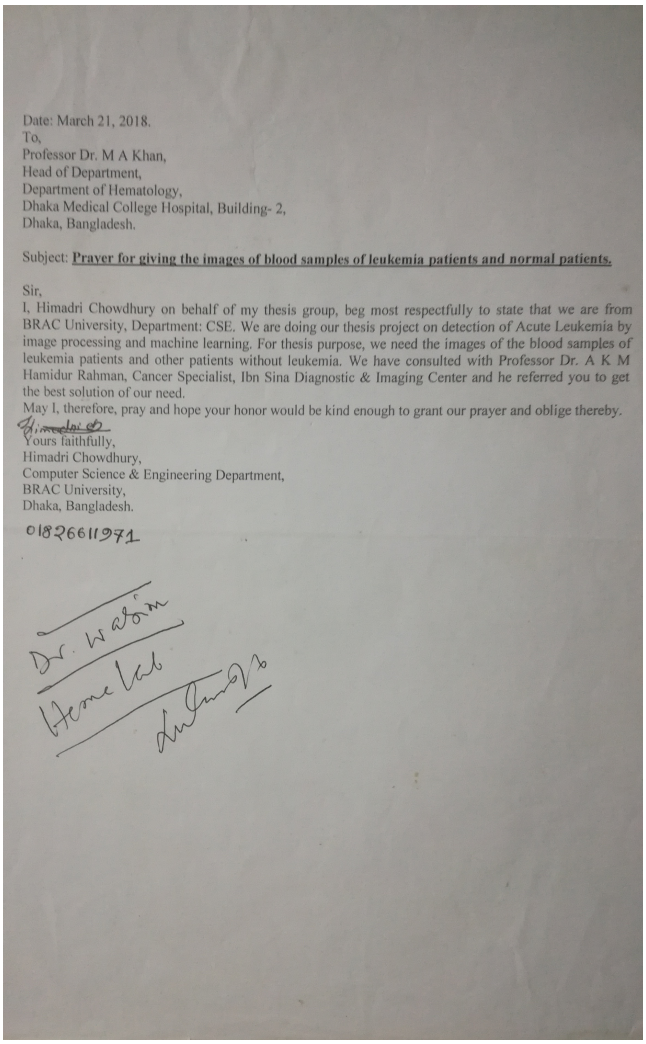
Data Collection and Project Overview

In this chapter, we discussed about, from which source and in what procedure we gathered our dataset. Overview of the project has been claimed in this chapter also.

3.1 Collection Process

Generally, there are two types of dataset. These are preprocessed dataset and real world dataset. With the aim to work with real world dataset, we made an effort to collect microscopic blood images with the help of few medical colleges. At first, we thought collecting data would be easy as our research topic is on a sensitive and arising issue. Unfortunately, it was quite the opposite as it turned out most of the medicals we visited was not willing give any data because of security purposes. At first we went to Professor Dr. A K M Hamidur Rahman, Cancer Specialist, Ibn Sina Diagnostic Imaging Center who referred us to a doctor who was suppose to help us but when the meeting was set he refused to help us with data because he thinks it was dangerous to hand over the such sensitive data. Then, Professor Dr. A K M Hamidur Rahman referred us to Professor Dr. M A Khan, Head of department of Hematology of Dhaka Medical College Hospital and scheduled a meeting with us. Then we presented our topic to the Professor with full details and was fascinated with our idea. Finally, we were approved for the collection of microscopic blood images.
To train and test our data we needed at least 1000 data samples for our methodology to give accurate results. But, unfortunately we discovered that Dhaka medical do not store the microscopic images in their database. So, our only way to get the data samples was to go there everyday and collect the data. We knew it would be hard to collect the amount of data we needed because it was not possible to go there everyday as we had classes to attend. Therefore, we took turns and went there and sometimes we had to miss our class to collect

data. But, unfortunately there are days we had to come back with no data, as the person in charge to provide us with the data was not always available because of the official workload of the hospital. Even after getting permission it was taking to much time to gather the amount of data we needed and as the deadline was approaching we could not afford to waste any time for gathering data from Dhaka Medical College Hospital. As a result, we hardly managed to collect only 20 microscopic image of peripheral blood sample.



Date: March 21, 2018.
To,
Professor Dr. M A Khan,
Head of Department,
Department of Hematology,
Dhaka Medical College Hospital, Building- 2,
Dhaka, Bangladesh.

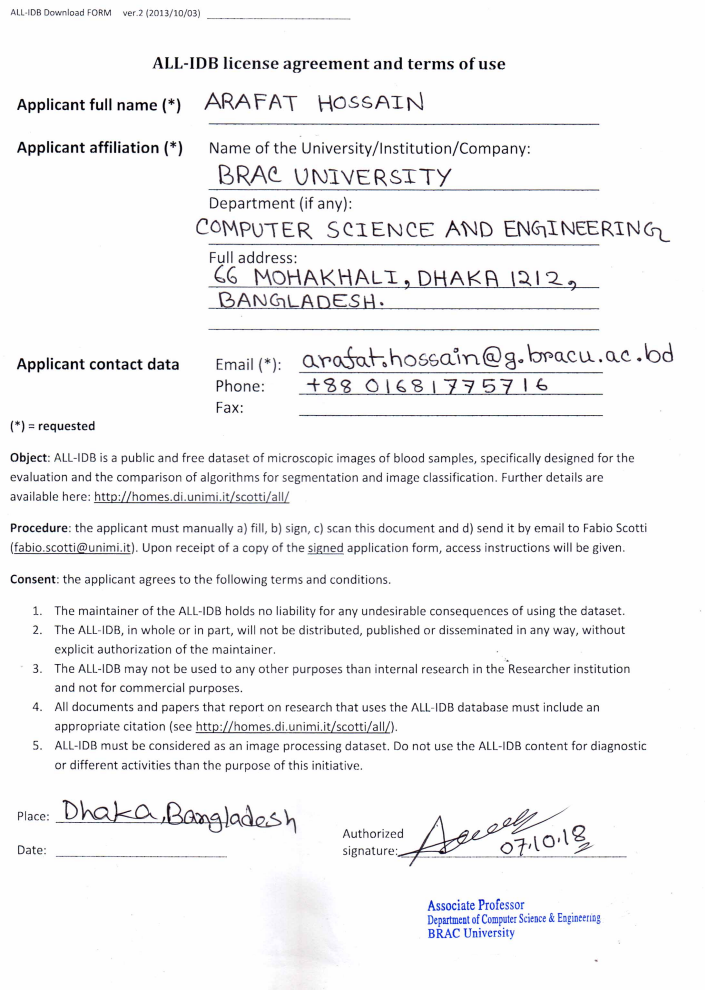
Subject: **Prayer for giving the images of blood samples of leukemia patients and normal patients.**

Sir,
I, Himadri Chowdhury on behalf of my thesis group, beg most respectfully to state that we are from BRAC University, Department: CSE. We are doing our thesis project on detection of Acute Leukemia by image processing and machine learning. For thesis purpose, we need the images of the blood samples of leukemia patients and other patients without leukemia. We have consulted with Professor Dr. A K M Hamidur Rahman, Cancer Specialist, Ibn Sina Diagnostic & Imaging Center and he referred you to get the best solution of our need.
May I, therefore, pray and hope your honor would be kind enough to grant our prayer and oblige thereby.

Yours faithfully,
Himadri Chowdhury,
Computer Science & Engineering Department,
BRAC University,
Dhaka, Bangladesh.
01826611971

Fig. 3.1 Application Form to Collect Blood Samples Affected by Leukemia.

The rest of the data were collected from a website [19] with the purpose of providing free public dataset of microscopic images of blood samples significantly designed for the evaluation and comparison of algorithms for segmentation and image classification. So, we had to fill up a form and submit it after which we got access to the dataset.



ALL-IDB Download FORM ver.2 (2013/10/03) _____

ALL-IDB license agreement and terms of use

Applicant full name (*) ARAFAT HOSSAIN

Applicant affiliation (*) Name of the University/Institution/Company:
BRAC UNIVERSITY
Department (if any):
COMPUTER SCIENCE AND ENGINEERING
Full address:
66 MOHAKHALI, DHAKA 1212,
BANGLADESH.

Applicant contact data Email (*): arafat.hossain@g.bracu.ac.bd
Phone: +88 01681775716
Fax:

(*) = requested

Object: ALL-IDB is a public and free dataset of microscopic images of blood samples, specifically designed for the evaluation and the comparison of algorithms for segmentation and image classification. Further details are available here: http://homes.di.unimi.it/scotti/all/

Procedure: the applicant must manually a) fill, b) sign, c) scan this document and d) send it by email to Fabio Scotti (fabio.scotti@unimi.it). Upon receipt of a copy of the signed application form, access instructions will be given.

Consent: the applicant agrees to the following terms and conditions.

1. The maintainer of the ALL-IDB holds no liability for any undesirable consequences of using the dataset.
2. The ALL-IDB, in whole or in part, will not be distributed, published or disseminated in any way, without explicit authorization of the maintainer.
3. The ALL-IDB may not be used to any other purposes than internal research in the Researcher institution and not for commercial purposes.
4. All documents and papers that report on research that uses the ALL-IDB database must include an appropriate citation (see http://homes.di.unimi.it/scotti/all/).
5. ALL-IDB must be considered as an image processing dataset. Do not use the ALL-IDB content for diagnostic or different activities than the purpose of this initiative.

Place: Dhaka, Bangladesh
Date: _____

Authorized signature: 07.10.18

Associate Professor
Department of Computer Science & Engineering
BRAC University

Fig. 3.2 Online Collection Form of Blood Samples Affected by Leukemia.

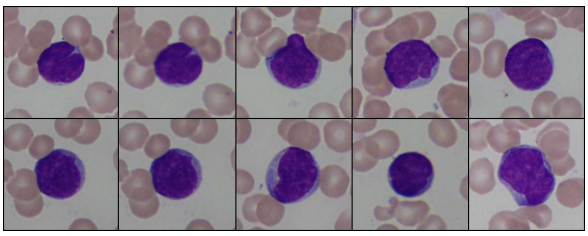


Fig. 3.3 Some of the Blood Samples (Collected Dataset) Affected by Leukemia.

.

3.2 Project Overview

In this section, the overview of the system has been discussed in details. This section contains brief explanation of the algorithms that we used to build up the system. This chapter also contains the discussions of the methods that has been used.
From the literature review, we found out that there are many ways to distinguish the abnormal blood cell and come to a conclusion by detecting the Acute Leukemia in a human body. In our system, we have decided to implement several methods in a sequential process to detect the presence of leukemia in human. Our main goal is to perform the processes in such a way that will give us the maximum accuracy in detection of leukemia which will differentiate our project from the others.
The details of what we plan to, how we plan to do it and what results we expect have been explained here in a summarize version. Discussions about the path in which we are going through to get the success in this project and the methods we will use have been provided.

3.2.1 Programming Language and Environment

To implement our proposed model for our research, we select the Python (Programming Language) to make our project functional. As an interpreted high-level programming language for general purpose programming, Python has fame all over the world. As a popular programming language and its environment is installed in most of the desktops, we easily can work with our project in our home as well in the university lab. The reason behind choosing the Python as programming language is its progressive features. Supporting of multiple programming paradigms including object-oriented, imperative, functional and procedural accomplish our tasks spontaneously.

3.2.2 Algorithms

By going through several research papers and journals related to our proposed model we have selected a few algorithms by which we have implemented to reach to our appropriate goal. All the algorithms are included of Machine Learning Models or Techniques. We have performed these algorithms to extract the features and classify the nucleus sample after the segmentation of blood samples and evaluated the accuracy of detecting the abnormal cell of nucleus with the maximum accuracy. Logistic Regression Algorithm, Decision Tree, Naive Bayes Theorem, Random Forest Algorithm, K-Nearest Neighbors (KNN) Algorithm,

Convolutional Neural Network (CNN) and AdaBoost Classifier Algorithm has been used in this thesis to find the proper and maximum rate of accuracy.

3.2.3 Image Processing

Image Preprocessing

Before the segmentation of image we need to prosess the these. In our proposed model we used wherein filter of image cleaning or preprocessing.

Image Segmentation

The first approach to detect the presence of acute lymphocytic leukemia is to segment the blood sample. Here, an effective technique has been used to segment the nucleus from the blood in a automatic process. Gray scale contrast enhancement and filtering is he main basic of this technique. 12 consecutive steps have to be applied to implement the technique where the nucleus will be detached from the blood sample. In a short brief, the steps are given below which are described properly in the Chapter- 4, the segmentation of the blood portion of the paper.

1. Conversion of the image (input), X to a grayscale image, Y (output).
2. Adjustment of the grayscale image, Y, intensity values with a linear contrast stretching to get image L.
3. Enhancement of the contrast of the grayscale image by the histogram equalization method to get image H.
4. Obtain the image R1= L+H to brighten all other image components except cell nucleus.
5. Obtain the image R2= L-H to highlight the entire image objects including cell nucleus.
6. Obtain the image R3= R1+R2 to remove all other components of blood.
7. Implementation of 3 by 3 minimums filter for the 3 (three) times on e image R3 for reducing the noise.
8. Using Otsu's method, calculate the global threshold value.
9. Conversion of binary image from R3 image by using the threshold value.

10. Use the morphological opening for removing the small pixel groups.
11. Give connection of the neighboring pixels to form objects.
12. Implementation of size test removal of all objects that are less than 0.50 of average RBC area [22, 10, 14].

By following the steps given above the nucleus will be separated from the blood sample which will help us to extract the feature of nucleus in further steps.

Image-Feature Exploration

After the segmentation of the microscopic image of blood sample we separated the nucleus from other particles of blood. The extraction process has been implemented on this nucleus to find the condition of it whether it is normal or abnormal which helps us to reach to the actual result of our proposed thesis project. Here, we used several algorithms separately to find the rate of accuracy, recall, precision score and validation score. The algorithms are

1. Logistic Regression Algorithm,
2. Decision Tree,
3. Naive Bayes Theorem,
4. Random Forest Algorithm,
5. K-Nearest Neighbors (KNN) Algorithm,
6. Convolutional Neural Network (CNN).

At the time of evaluation, we got different accuracy, recall, precision score and validation score for the different algorithms. Among all, we got the maximum success from the Convolutional Neural Network (CNN). The idea of the algorithms are described explicitly in the Feature Extraction portion of this paper.

3.2.4 Result and Analysis

After the feature extraction, we obtained results for different algorithms. By comparing all the results we came to know that the Convolutional Neural Network (CNN) is working most efficiently with respect to our datasets. The mechanism that we followed for each algorithm

to know the behaviour of these and the evaluation processes are discussed in details in the "Result and Analysis" chapter of this paper.

3.2.5 Total Workflow of the Proposed Model

The Flowchart of the Total Workflow of the Proposed Model is given below.

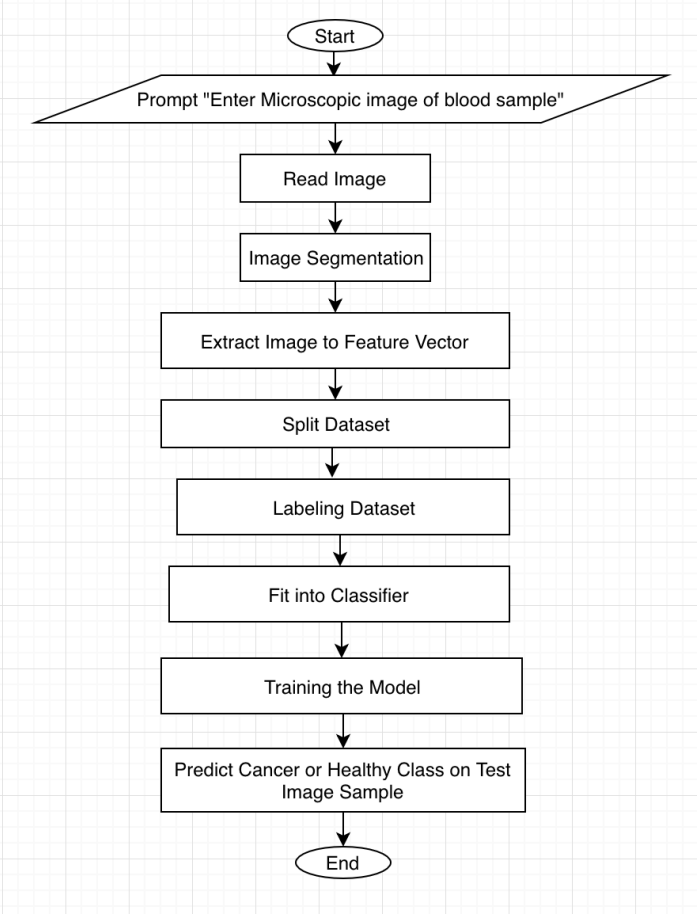


Fig. 3.4 Workflow of the whole project.

Chapter 4

Algorithms

In this chapter, we discussed in details about how did we extracted the features and classify the nucleus which we got from the segmentation of the sample of blood. Here, we used a couple of algorithms to extract the features and classify the nucleus. Implementation of the Logistic Regression Algorithm, Decision Tree, Gaussian Naive Bayes Theorem, AdaBoost Classifier Algorithm, Random Forest Algorithm, K-Nearest Neighbors (KNN) Algorithm and Convolutional Neural Network (CNN) maximized our success of our proposed model by showing the different accuracy rate, recall score, precision score and validation score of each algorithm. Efficiency of the algorithms we have used in our project in extracting the features and classification of neucleus and detection of the presence of Acute Lymphocytic Leukemia in blood samples are described elaborately in below.

4.1 Logistic Regression Algorithm

Logistic Regression is a classification algorithm. It is used to anticipate a binary outcome (1/0, Yes/No, etc) provided a set of independent variables. Simply, it speculates the probability of event of a phenomenon by attaching data to a logit function. Logistic Regression is a member of a larger class of algorithms known as Generalized Linear Model (GLM) .
The fundamental equation of generalized linear model:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2 \tag{4.1}$$

Here, g() is the link function, E(y) is the expectation of target variable and

$$\alpha + \beta x1 + \gamma x2 \tag{4.2}$$

is the linear predictor. The job of link function is to 'link' the expectation of y to linear predictor.
Equation of Simple Linear Regression:

$$g(y) = \beta o + \beta(Age) \tag{4.3}$$

As g() is the link function so it is constructed using two things, probability of success (p) and probability of failure (1-p) and they should meet the criteria:
a) p>= 0
b) p<= 1
Since, the probability should always be positive, we will put the linear equation in exponential form and g() is denoted with 'p'.

$$p = exp(\beta o + \beta(Age)) = e^{\beta o + \beta(Age)} \tag{4.4}$$

To make the probability less than 1, we divide p by a value greater than p:

$$p = exp(\beta o + \beta(Age))/exp(\beta o + \beta(Age)) + 1 = e^{\beta o + \beta(Age)}/e^{\beta o + \beta(Age)} + 1 \tag{4.5}$$

Using the equations (4.3), (4.4) and (4.5) we get the Logit Function:

$$p = e^y/1 + e^y \tag{4.6}$$

where p is the probability of success

$$q = 1 - p = 1 - e^y/1 + e^y \tag{4.7}$$

where q is the probability of failure.
Dividing, (4.6)/(4.7), we get,

$$p/(1-p) = e^y \tag{4.8}$$

After taking log on both side,

$$log[p/(1-p)] = y \tag{4.9}$$

log(p/1-p) is a link function. After substituting value of y we get the equation of Logistic Regression:

$$log[p/(1-p)] = \beta o + \beta(Age) \tag{4.10}$$

Here, (p/1-p) is the odd ratio, so when the log of odd ratio is positive, the probability of success is always more than 0.50.
We trained our dataset with "Logistic Regression" where the testing accuracy was 76.92%, recall score 76.92%, precision score 78.19% and cross-validation score 77.55% [4].

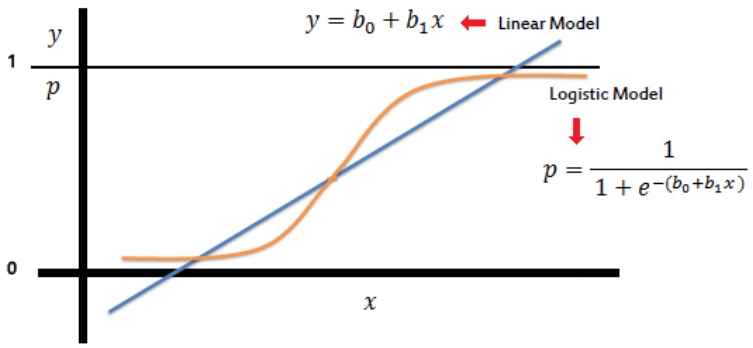


Fig. 4.1 Graph of Logistic Regression Algorithm [35]

4.2 Decision Tree Algorithm

A decision tree is a flowchart like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test and each leaf node represents

a class label. The paths from root to leaf represents classification rules. Decision Tree Classifier presents a string of crafted questions about the attributes of the test record. After receiving each answer, there is follow-up question until it reaches to the class label of the record. Initiating from the root node the test condition to the record is applied and based on the outcome of the test the appropriate branch is followed. A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. To build a decision tree, we need to calculate two types of entropy using frequency tables as follows [41, 3].

a) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{e} -P_i \log_2(P_i) \tag{4.11}$$

b) Entropy using the frequency table of two attributes:

$$E(T,X) = \sum_{c \subset X} P(c) * E(c) \tag{4.12}$$

The visual graph of the Decision Tree Algorithm is given below for understanding the method easily.

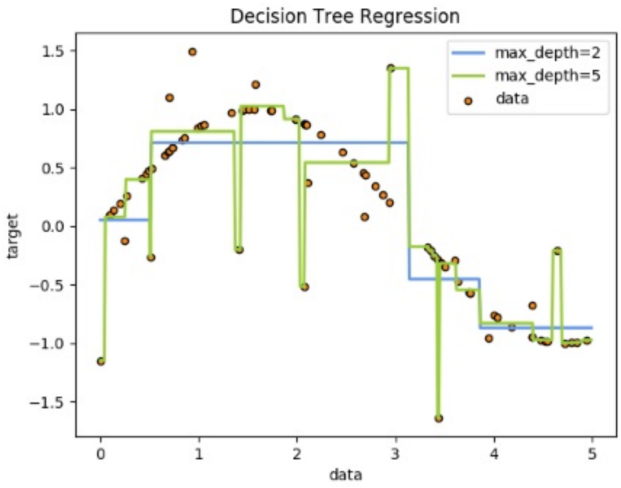


Fig. 4.2 Graph of Decision Tree Algorithm [29]

4.3 Gaussian Naive Bayes Algorithm

Naive Bayes is a classification expertise based on Bayes' Theorem with a suspicion of independence among predictors. In basic terms, a Naive Bayes classifier expect that the nearness of a specific component in a class is inconsequential to the nearness of some other element. For instance, a natural product might be viewed as an orange in the event that it is yellow, round, and around 4 inches in diameter. Regardless of whether these features rely upon one another or upon the presence of alternate features, these properties autonomously add to the likelihood that this organic product is an orange and that is the reason it is known as 'Naive'. Bayes hypothesis gives a method for ascertaining back likelihood P(c|x) from P(c), P(x) and P(x|c) and the equation:

$$P(c|x) = P(x|c)P(c)/P(x) \tag{4.13}$$

* P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
* P(c) is the prior probability of class.
* P(x|c) is the likelihood which is the probability of predictor given class.
* P(x) is the prior probability of predictor.

We trained our dataset with "Naive Bayes Gaussian" where the testing accuracy was 71.79%, recall score 71.79%, precision score 71.96% and cross validation score 67.30% [31].

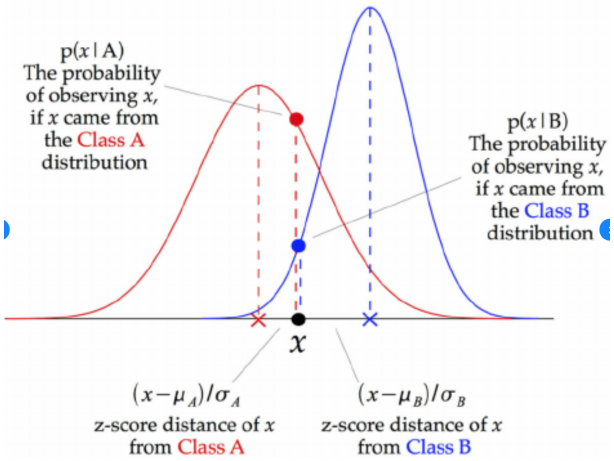


Fig. 4.3 Graph of Gaussian Naive Bayes Algorithm [30]

4.4 Random Forest

A random forest multi-way classifier consists of a number of trees, with each tree grown utilizing randomization. The leaf nodes of each tree are labeled by estimates of the posterior distribution over the image classes. Each internal node contains a test that best splits the space of data to be classified. An image is classified by sending it down every tree and aggregating the reached leaf distributions. The trees here are binary and are constructed in a top-down manner. The binary test at each node can be chosen in one of two ways: (i) randomly, i.e. data independent; or (ii) by a greedy algorithm which picks the test that best separates the given training examples. “Best” here is measured by the information gain

$$\Delta E = -\sum_{i} \frac{|Q_i| E(Q_i)}{|Q_i|} \tag{4.14}$$

caused by partitioning the set Q of examples into two subsets Qi . according the given test. Here E(q) is the entropy

$$-\sum_{i}^{N} =_1 P_j \log_2(P_j) \tag{4.15}$$

with Pj the proportion of examples in q belonging to class j, and | . | the size of the set. The process of selecting a test is repeated for each nonterminal node, using only the training examples falling in that node. The recursion is stopped when the node receives too few examples, or when it reaches a given depth. The test image is passed down each random tree until it reaches a leaf node. All the posterior probabilities are then averaged and the argmax is taken as the classification of the input image. We prepared our dataset with "Random Forest" where the testing accuracy was 84.61%, recall score 84.1%, precision score 88.03% and cross validation score 85.00%. [39].
Visual graph of this algorithm is given into the next page.

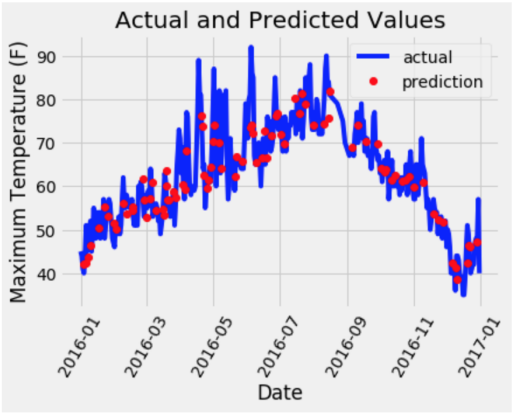


Fig. 4.4 Graph of Random Forest Algorithm [16]

4.5 K-Nearest Neighbors (KNN) Algorithm

The K-Nearest Neighbors (KNN) algorithm is a robust and flexible classifier that is frequently utilized as a benchmark for more intricate classifiers, for example, Artificial Neural Networks (ANN) and Support Vector Machines (SVM). K-Nearest Neighbors (KNN) used in variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. K-Nearest Neighbors (KNN) is a non-parametric and lazy learning algorithm. Non-parametric means there is no supposition for fundamental information conveyance. Lazy algorithm it does not need any training data points for model generation. All preparation information utilized in the testing stage. This makes preparing quicker and testing stage slower and costlier. Expensive testing stage implies time and memory. In the most pessimistic scenario, K-Nearest Neighbors (KNN) needs more opportunity to examine all information focuses and filtering all information focuses will require more memory for putting away preparing information.
In the classification setting, the K-closest neighbor calculation basically comes down to forming a majority vote between the K most comparable instances to a given "inconspicuous" perception. Similarity is characterized by a separation metric between two data points. A well known decision is the Euclidean separation given by

$$d(x,x') = \sqrt[0.5]{(x1-x1')^2+(x2-x2')^2+...+(xn-xn')^2} \tag{4.16}$$

To start with, you discover the k nearest to the designated point and after that group focuses by larger part vote of its k neighbors. Each object votes in favor of their class and the class with the most votes is taken as the expectation. We arranged our dataset with "K-Nearest

Neighbors" where the testing accuracy was 58.79%, recall score 58.79%, precision score 65.58% and cross validation score 68.46% [24, 47].

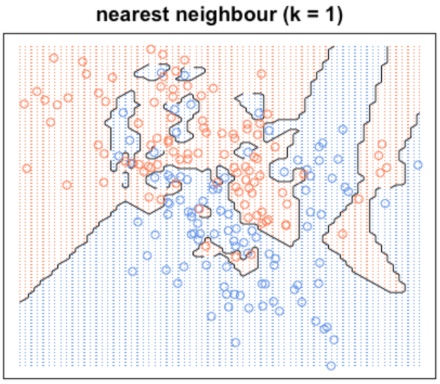


Fig. 4.5 Graph of K-Nearest Neighbors (KNN) Algorithm [47]

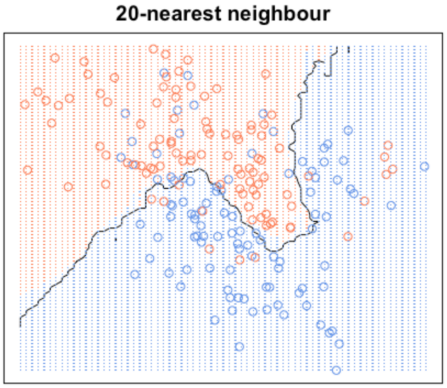


Fig. 4.6 Graph of K-Nearest Neighbors (KNN) Algorithm [47]

4.6 Convolutional Neural Network (CNN)

At its most essential, Convolutional Neural Network (CNN) can be thought of as a sort of neural system that utilizes numerous indistinguishable duplicates of a similar neuron. This enables the system to have loads of neurons and express computationally vast models while keeping the quantity of genuine parameters—the values depicting how neurons behave—that should be adapted genuinely little.

An input image is passed to the first convolutional layer. The output is acquired as an activation map. The channels connected in the convolutional layer remove significant feature

from the input image to pass further. Each channel will give an alternate feature to help the right class prediction. On the off chance that we have to hold the size of the image, we utilize same padding (zero padding), other savvy substantial padding is utilized since it decreases the quantity of features. Pooling layers are then added to additionally decrease the quantity of parameters. A few convolution and pooling layers are included before the prediction is made. Convolutional layer help in separating features. As we go further in the system more explicit features are removed when contrasted with a shallow system where the features extricated are more conventional. The output is then produced through the output layer and is contrasted with the output layer for error generation. A loss function is characterized in the fully connected output layer to figure the mean square loss. The gradient of error is then calculated.The error is then back propagated to refresh the filter (weights) and inclination esteems. One preparing cycle is finished in a solitary forward and in reverse pass. We have passed our dataset through the system and achieved an accuracy of 90.09% [20, 7].
We can summarized the theory of convolutional layer by some mathematical equations. Accept a volume of size:

$$W_1 * H_1 * D_1 \tag{4.17}$$

Requires of four filters K, their spatial extent F, the stride S, the amount of zero padding, P. Produces a volume of size

$$W_2 * H_2 * D_2 \tag{4.18}$$

where:

$$W_2 = (W_1 - F + 2P)/S + 1 \tag{4.19}$$

$$H_2 = (H_1 - F + 2P)/S + 1 \tag{4.20}$$

$$D_2 = K \tag{4.21}$$

With parameter sharing, it introduces

$$(F.F.D_1).K \tag{4.22}$$

weights and K biases.
In the output volume, the d- th depth slice (of size W2*H2) is the result.

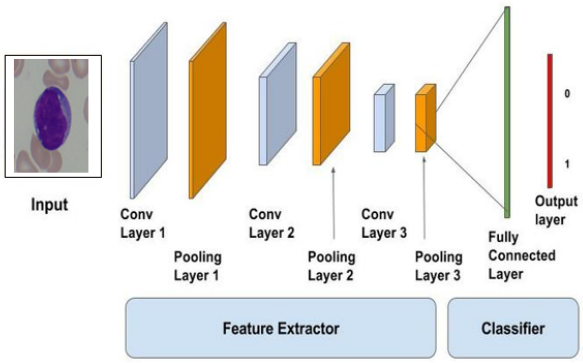


Fig. 4.7 Graph of Convolutional Neural Network (CNN) Algorithm [8]

4.7 AdaBoost Classifier Algorithm

AdaBoost, another way to say "Adaptive Boosting", is the primary pragmatic boosting calculation proposed by Freund and Schapire in 1996. It centers around classification issues and intends to change over an arrangement of frail classifiers into a solid one. The final equation for classification can be represented as

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)) \tag{4.23}$$

where f_m represents the m_th feeble classifier and theta_m is the comparing weight. It is actually the weighted blend of M frail classifiers.
Given a set of data containing n points, where

$$x_i \in \mathbb{R}^d, y_i \in (-1, 1) \tag{4.24}$$

Here -1 represents the negative class whereas 1 denotes the positive class.
Instate the weight for every data point as:

$$w(x_i, y_i) = (1/n), i = 1, ..., n \tag{4.25}$$

For iteration m = 1, ..., M:
a) Fit weak classifiers to the data set and select the one with the most minimal weighted

characterization blunder:

$$\varepsilon_m = E_{wm}[1_{y \neq f(x)}] \tag{4.26}$$

b)The weight for the m_th weak classifier is calculated:

$$\theta_m = (1/2)ln((1-\varepsilon_m)/\varepsilon_m) \tag{4.27}$$

For any classifier with exactness higher than half, the weight is sure. The more exact the classifier, the bigger the weight. While for the classifier with under half exactness, the weight is negative. It implies that we join its prediction by flipping the sign. For instance, we can transform a classifier with 40% precision into 60% exactness by flipping the indication of the expectation. In this way even the classifier performs more terrible than arbitrary speculating, regardless it adds to the last prediction. We just don't need any classifier with correct half exactness, which doesn't include any data and in this way contributes nothing to the last prediction [34].

c) The weight for each data point is updated as:

$$w_{m+1}(x_i, y_i) = (w_m(x_i, y_i)exp[-\theta_m y_i f_m(x_i)])/Z_m \tag{4.28}$$

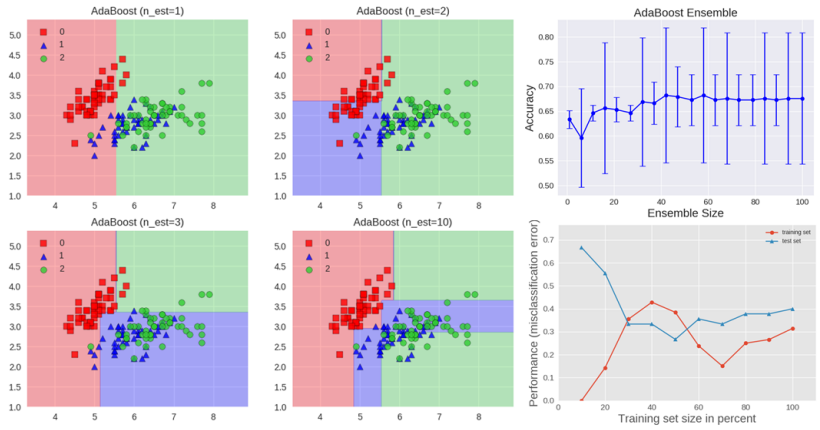


Fig. 4.8 Graph of AdaBoost Classifier Algorithm [38]

Chapter 5

Image Processing

Once all the images are acquired the images are needed to go through certain steps before to fit into classification model. In our proposed model images are went through following steps.

5.1 Image Preprocessing

Data collected from nature may have problems. In case of image data it can have blurred regions which can hamper accuracy rate of any classification. In solution to that we have filtered the images with wiener filter. In the image cleaning, the filter removed all the leucocyte which are at the edge of the image and components are not leucocyte. Solidity need to be measured for better image cleaning. The images we got in RGB form we need to convert it into gray scale image for further processing. Solidarity needs area and convex of each leukocyte.

$$Solidity = area/ConvexArea \tag{5.1}$$

5.2 Image Segmentation

Segmentation means festering of image in different area for various enactments. The segmentation part is very decisive because the accuracy of the successive feature extraction and classification hinge on the unerring segmentation of white blood cells. Further, it is arduous and exigent problem because of the composite nature of the cells and apprehension in the microscopic image.

In the first instance, the blood sample is taken from the dataset section. Then, we started with converting the images into grayscale images so that the nucleus part of the cell comes into view as the darkest part of the image.

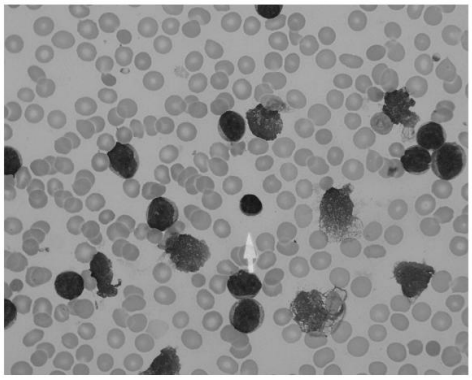


Fig. 5.1 Conversion into grayscale image from the normal image of blood.

The localization of the white blood cell nucleus is constructed on linear contrast stretching, histogram equalization and image arithmetic. The first copy of the grayscale image was intensified with linear contrast stretching which was denoted as A1 and the other copy was enhanced with histogram equalization which was denoted as A2.

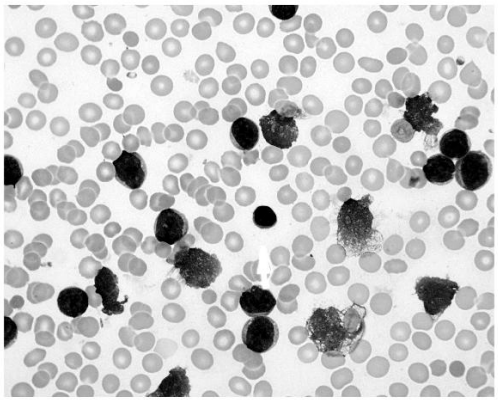


Fig. 5.2 Adjustment of the grayscale image of blood.

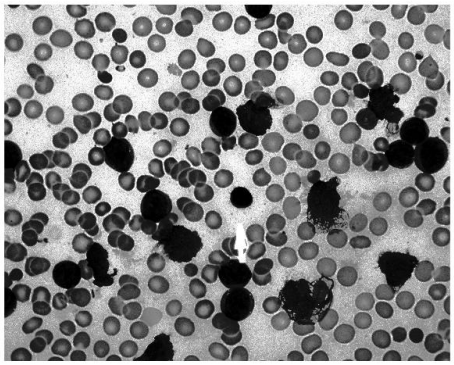


Fig. 5.3 Enhancement of the contrast of the grayscale image by the histogram equalization.

After applying the methods on the grayscale images the resultant images were added. Execution of the addition of the image, all the resultant pixels exceeding the intensity value of 255 were truncated to 255 which lightened most of the attributes in the image excluding the nucleus.

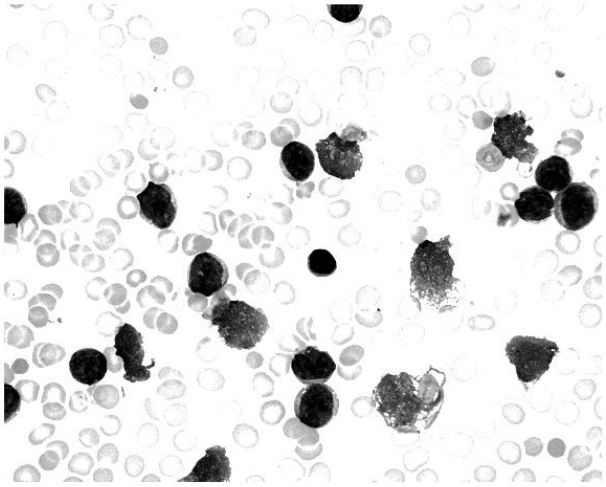


Fig. 5.4 Image of all other components after brighten except cell nucleus.

The histogram equalized image (A2) was subtracted from the resultant image of the addition (I1). This action spotlighted all the objects and all its border in the image incorporating the cell nucleus.

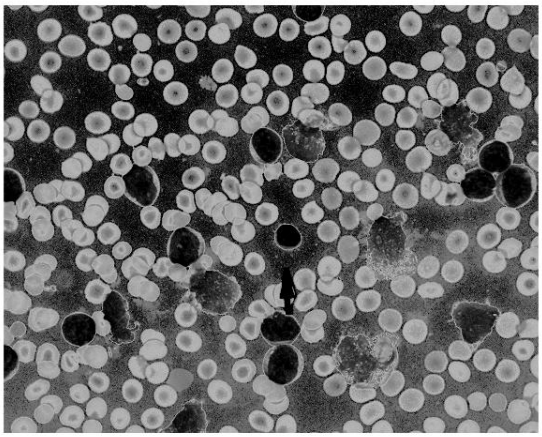


Fig. 5.5 Highlighted image of the entire objects including cell nucleus.

The resultant of both the addition (I1) and subtraction (I2) were added which gives a new resultant I3. The addition helped to separate almost all the other blood components while it clinged to the nucleus with minimum influence of disformity on the nucleus part of the white blood cell.

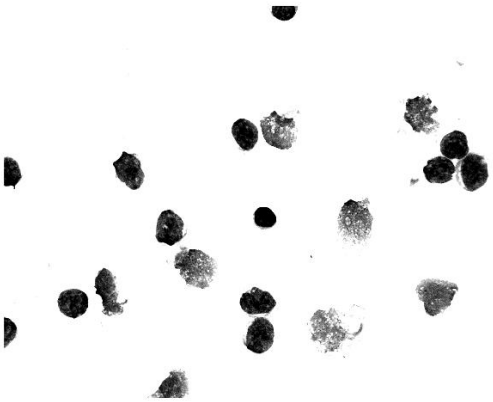


Fig. 5.6 Image after removing all other components of blood.

After executing all the arithmetic operations a global threshold using Otsu's method is required to detect the nucleus but executing it at this level is risky at it might lead to miss-segmentation of some part of the nucleus due to the effect of deformity after the last arithmetic operation. Therefore, to circumvent the issue a 3 by 3 minimum filter was used to raise the intensity value which made the nucleus part darker, after which global threshold Otsu's

method was applied.

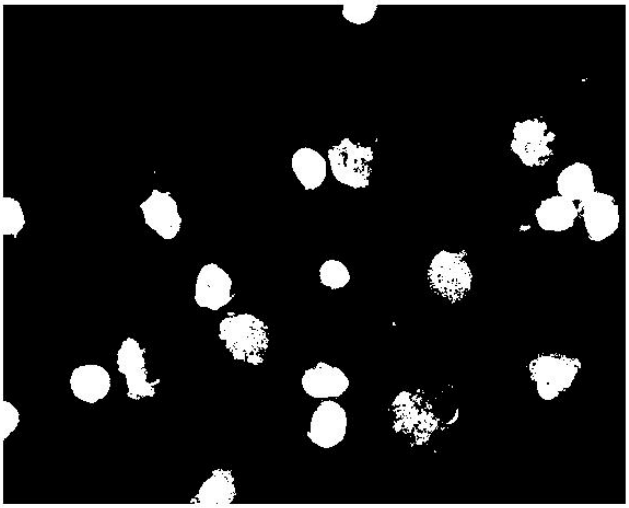


Fig. 5.7 Image after thresholding the blood sample.

Using the threshold value from the last step I3 was transformed to binary image. Then, morphological opening was used to separate small pixel groups.

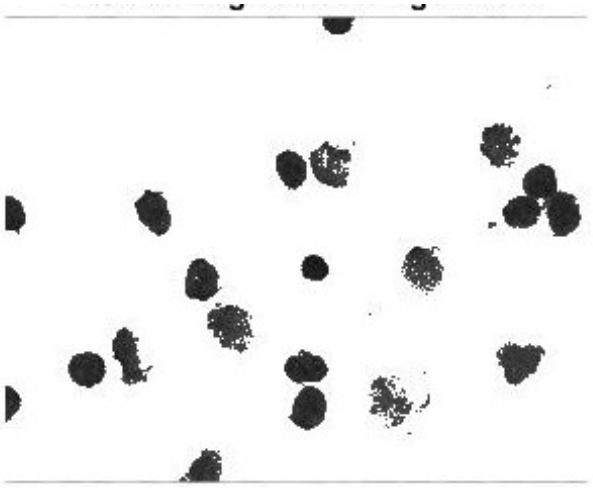


Fig. 5.8 Image of finally segmented (separation of nucleus) blood sample.

Inspite of having the detailed description in this section we are giving the total workflow is given in the next page so that one can understand the total method at a glance.

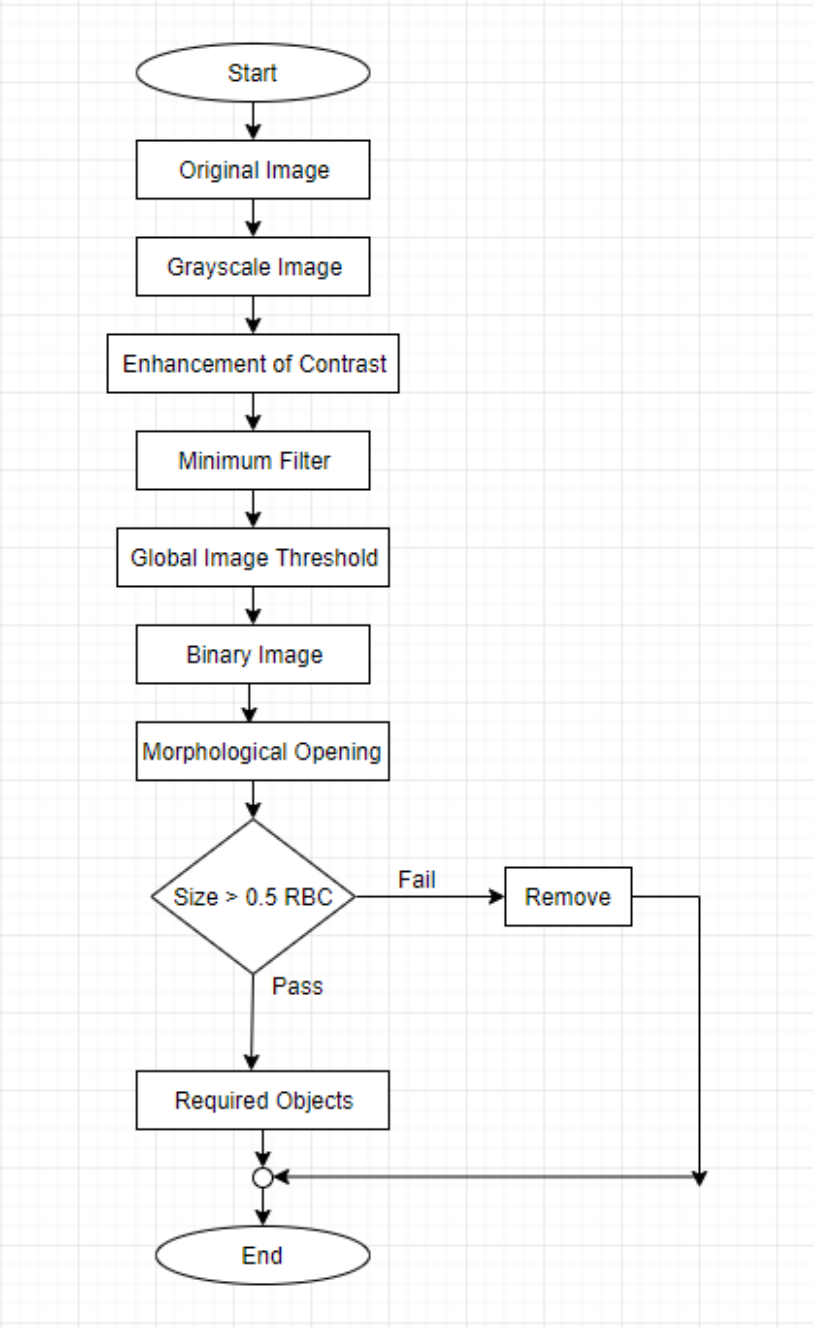


Fig. 5.9 Workflow of segmentation.

5.3 Image-Feature Exploration

Image feature is needed to be fit in the algorithms for classification. Image feature is a pattern of an image which explains what we see in the image. We Extracted image into vector space for computer vision. This eases the possibility to perform mathematical calculation on them for instance to find similar images. Among two of the methods to extract feature named, 'Image Descriptor', 'Neural Nets', we have used Image Descriptor method. For this, the best possible way is use of 'Raw Pixel Descriptor'. The following steps has been followed :

1. Detecting Key point descriptors of the image.
2. Giving Images a same size of 257x257
3. Execution Flatten Operation to fit all of the images into one big vector.
4. Making Descriptor of same size, 64. So our dimension of the vector is 4096
5. Extracting feature and storing feature vectors in rawImages array.
6. Labeling based on image path.

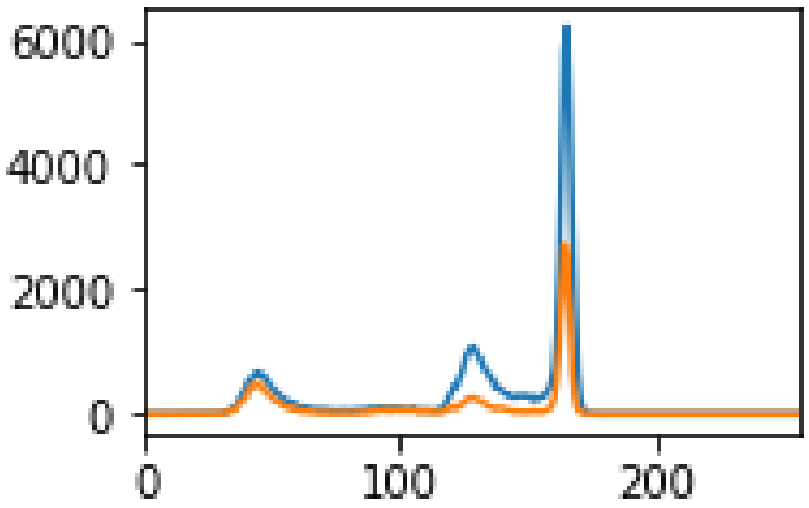


Fig. 5.10 Feature vector plot.

5.3.1 Feature Extraction and Labeling

```
# loop over the input images
for (i, imagePath) in enumerate(imagePaths):
# load the image and extract the class label
# our images were named as labels.image_number.format
image = cv2.imread(imagePath)
# get the labels from the name of the images by extract the string before "."
label = imagePath.split(os.path.sep)[-1].split(".")[0]

# extract raw pixel intensity "features"
#followed by a color histogram to characterize the color distribution of the pixels
# in the image
pixels = image_to_feature_vector(image)
print('————————pixels')
print(pixels)
print('————————end')
hist = extract]_color_histogram(image)

#print(pixels)
#print(hist)

# add the messages we got to the raw images, features, and labels matricies
rawImages.append(pixels)
#features.append(hist)
labels.append(label)
# show an update every 200 images until the last image
if i > 0 and ((i + 1)% 200 == 0 or i ==len(imagePaths)-1):
print("[INFO] processed /".format(i+1, len(imagePaths)))
```

5.3.2 Image Preprocess for Neural Network

Above procedure for fitting image data into machine mcahine learning classifier. For implementing neural network we preprocessed the image dataset in a different method. There are various approaches for this for example, Image Scaling, Uniform Aspect Ratio, Mean, Standard Deviation of input data, Normalizing Image Inputs, Dimensionality Reduction, Data

Augmentation. Although CNN itself handle the large scale color space data but for better accuracy we used Data Augmentation before classifying with CNN. Data Augmentation does following tasks:

1. Applies several Transformations of the original input which increases the number of training dataset [40].

2. Duplicates the instances of training set by various transformations for example translation, rotation, symentries [40].

Chapter 6

Result and Analysis

In this chapter, we will try to compare and find reasons behind the accuracy rate of each algorithm on Acute Leukemia Cancer dataset. We are going to illustrate relation based on the table stated below.

6.1 Accuracy of All Models

Results are acquired by implementing different classification and concern algorithms. The classification models are trained with 80% of training data and 20% of test data dividing into 2 classes, cancer and healthy data samples. The classification array we have, consists of Logistic Regression Algorithm, Decision Tree Algorithm, Random Forest Algorithm, Gaussian Naive Bayes Algorithm, K-Nearest Neighbour Algorithm (KNN), AdaBoost Algorithm, Voting and Convolutional Neural Network (CNN) Algorithm.
The results we got by evaluating our datasets with several algorithms are given into the next page.

Name of Algorithm	Accuracy Score	Precision Score	Recall Score	F1 Score	Cross Validation Score	ROC Score
Logistic Regression	76.92%	78.19%	76.92%	71.76%	77.55%	70.10%
AdaBoost Algorithm	87.17%	88.41%	87.17%	87.19%	90.07%	88.18%
Decision Tree	69.92%	71.17%	69.23%	70.18%	76.53%	68.69%
Gaussian Naive Bayes	71.79%	71.96%	71.79%	71.87%	67.30%	86.95%
Random Forest	84.61%	88.03%	84.61%	86.62%	85.00%	62.36%
K-Nearest Neighbors	58.79%	65.58%	58.79%	66.26%	68.46%	72.22%
Voting with ANK	89.74%	90.35%	89.74%	95.09%	87.30%	90.35%
Voting with ADR	79.48%	83.47%	79.48%	81.43%	86.92%	81.65%
Voting with NRD	76.92%	78.19%	76.92%	77.55%	82.69%	77.85%
Convolutional Neural Network	90.09%					

Table 6.1 Result After Feature Extraction and Classification From Different Algorithms.

6.2 Model Evaluation

Model evaluation is essential part of any machine learning project. A model may give satisfactory result while testing but can perform poor against any evaluation method like as accuracy score, Receiver Operating Characteristic Graph (ROC), Confusion Matrix etc. In our proposed model we evaluated our models with Cross Validation Accuracy, Receiver Operating Characteristic Graph (ROC) and Classification Report.

6.2.1 Ensemble Model

The main purpose of ensemble methods is to combine the accuracy of several base estimators built with a given learning algorithm in order to improve robustness over a single estimator. Most ensemble methods use a single base learning algorithm to produce homogeneous base learners. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a 'vote'. It works by first creating two or more standalone models from your training dataset. A Voting Classifier can then be used to wrap models and average the predictions of the sub-models when asked to make predictions for new data. The predictions of the sub-models can be weighted, but specifying the weights for classifiers manually or even heuristically is difficult. More advanced methods can learn how to best weight the predictions from sub models. This is called stacking (stacked aggregation). Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses.

1. The Accuracy of AdaBoost Classifier, Decision Tree Algorithm and Random Forest Algorithm: 89.88825671973453

2. The Accuracy of Decision Tree, Gaussian Naive Bayes Algorithm, Random Forest Algorithm: 79.48717948717948

3. The Accuracy of AdaBoost Classifier, K-Nearest Neighbor (KNN) Algorithm and Gaussian Naive Bayes Algorithm: 94.458257455672455

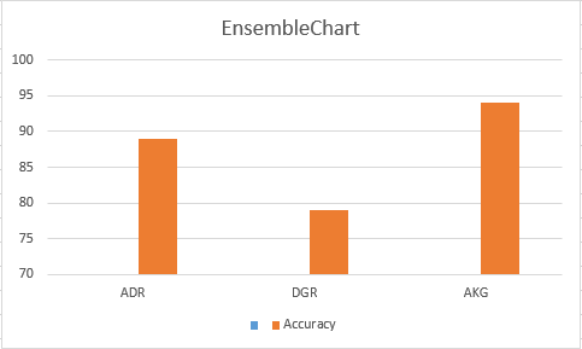


Fig. 6.1 Accuracy of AdaBoost Classifier, Decision Tree Algorithm and Random Forest Algorithm; Decision Tree, Gaussian Naive Bayes Algorithm, Random Forest Algorithm; AdaBoost Classifier, K-Nearest Neighbor (KNN) Algorithm and Gaussian Naive Bayes Algorithm.

6.2.2 Cross Validation

Knowing the parameters of a prediction function, tuning and testing it on same data samples is a methodological mistake [27]. A model that would repeat the same labels of the samples that have seen before tend to attain training accuracy of 1. But it may fail when it comes across an unknown test data. This case scenario is known as overfitting. To solve this issue we can use cross validation method. Generally, obtaining cross validation score by different approaches in the best practice but it is not always possible. The basic approach is called K-fold CV, the dataset is splitted into K subsets. The following procedure is followed for every K folds.

1. The model is trained with the K-10 of the folds of the training data.

2. The testing is applied on the remained data sets.

Usually 5 fold CV is enough but as we have less data samples we have used 10 fold CV which is again took much time to compute the cross validation score.
Among all the classifiers, till now AdaBoost , Random Forest, Voting (AdaBoost, Gaussian Naive Bayes, K-Nearest Neighbor (KNN)) came out as most possible perfect model with the test accuracy of 87.17%, 84.61%, 89.74% and respectively cross validation score of 90.07%, 85.00%, 87.30%.

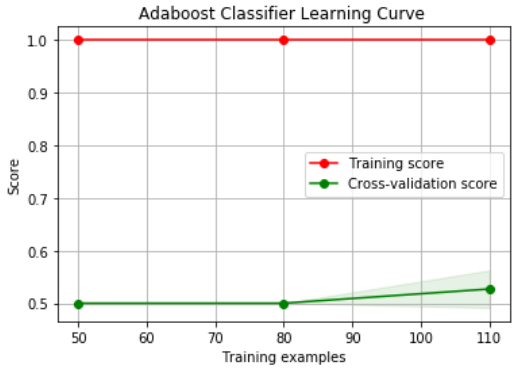


Fig. 6.2 Adaboost Classifier Learning Curve.

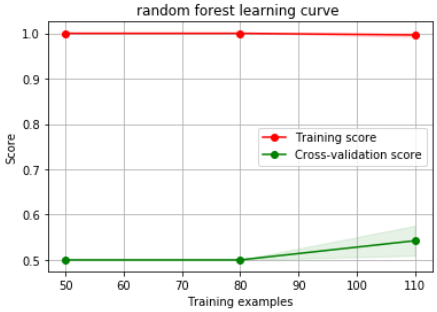


Fig. 6.3 Random Forest Learning Curve.

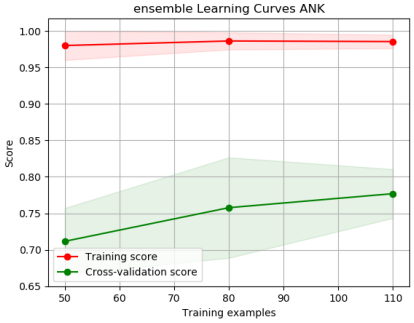


Fig. 6.4 Ensemble Learning Curve: AdaBoost, Gaussian Naive Bayes, K-Nearest Neighbor (KNN) Algorithm.

Our Train Accuracy is almost 1.0 in every classifier. It can happen because of the short length of dataset and less variance in the dataset. So by learning curves we actually can not say which classifier is working best on image data set. So we implemented Cross Validation; k=10. Then, we compared the Test Accuracy against the Cross Validation Score.

Classifier	Test	Train	Cross Validation
Random	84.61	89.61	85
AdaBoost	87.17	100	90.07
Voting	89.74	100	87.3

Fig. 6.5 Cross validation table.

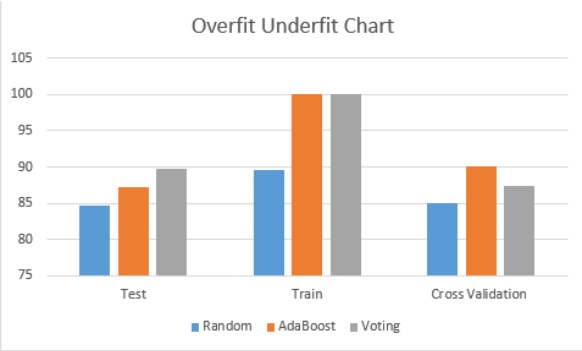


Fig. 6.6 Cross validation chart.

From the above table we can observe, among all the classifiers Random Forest is best fit with the cross validation accuracy of 85% against the Test Accuracy of 84.61%.

6.2.3 Area Under The Curve

The score of Area Under the Curve can validate a classifier efficiency. Two rates are measured here, (i) True Positive Rate (Sensitivity), (ii) False Positive Rate (Specificity). Four variables can affect the result:

1. **True Negative:** The classifier predicted Negative class true. Negative class is healthy data set.

2. **False Positive:** The classifier predicted Positive class when the class is actually Negative.

3. **True Positive:** The classifier predicted Positive class true. Positive class is cancer data set.

4. **False Negative:** The classifier predicted positive class when the class is actually negative.

True Positive Rate (TPR)

True Positive/ (False Negative + True Positive). Higher the rate higher efficient the classifier is.

False Positive Rate (FPR)

False Positive/ (False Positive + True Negative). Higher the rate higher inefficient the classifier is.
Among the classifiers we attained promising ROC score for AdaBoost and Voting (Adaboost, Gaussian Naive Bayes) classifier respectively 88.18%, 90.35%.

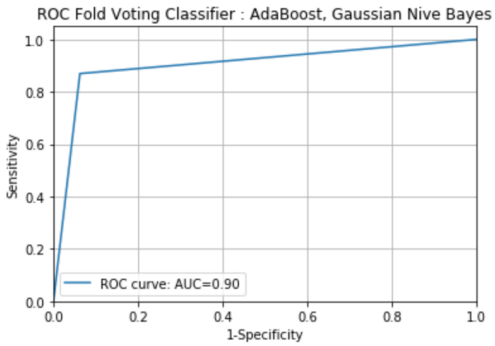


Fig. 6.7 ROC Fold Voting Classifier: AdaBoost, Gaussian Naive Bayes.

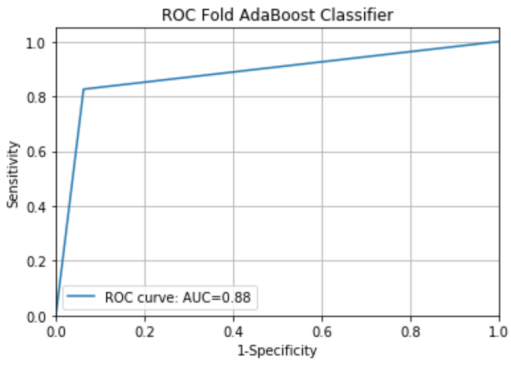


Fig. 6.8 ROC Fold AdaBoost Classifier.

Analyzing the above graphs we can observe, AdaBoost and Voting (AdaBoost, Gaussian

Naive Bayes) classifier have the highest sensitivity score of 88.18% and 90.35% against minimum specificity score.

6.2.4 Classification Report

Classification Report is one of the popular method to compare machine learning algorithms. The report consists of Precision Score, Recall Score, F-1 Score.

1. **Precision Score:** Ability of a classifier to not to label as positive samples when it is negative.

$$PrecisionScore = tp/(tp + fp) \tag{6.1}$$

2. **Recall Score:** Ability of a classifier to predict all positive sample with the true value.

$$RecallScore = tp/(tp + fn) \tag{6.2}$$

3. **F-1 Score:** The Weighted harmonic mean of Precision and Recall score which signifies that both the Precision and Recall scores are important.

Classifier	Precssion	Recall	F-1 Score
Random	88.03	84.61	86.62
AdaBoost	88.41	87.17	87.17
Voting	90.35	89.74	9509.00%

Fig. 6.9 Classification report table.

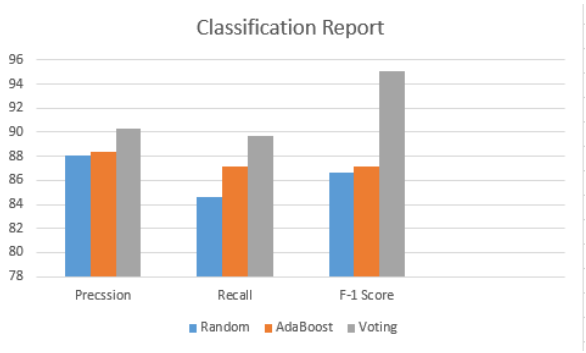


Fig. 6.10 Classification report chart.

From the above table we can nominate Voting Classifier (AdaBoost, Gaussian Naive Bayes, KNN) as the best fit one as it has 90.35% of Precision Score and 89.74% Recall Accuracy.

6.3 Candidate Classifier

In the Model Evaluation chapter three classifiers can be nominated as candidate classifiers, AdaBoost Ensemble Classifier, Voting Ensemble Classifier mixing AdaBoost, Gaussian Naive Bayes, KNN classifiers together and Random Forest. In the recent age when it comes to pattern recognition task Convolutional Neural Network (CNN) is thought to be best fitted. So in this paper we implemented CNN on our image dataset.

6.3.1 Convolutional Neural Network

The process is a tree of Multi Layer Parser classifier. It consists of 4 layers, (i) Convolutional Layer, (ii) ReLU Activation Layer, (iii) Pooling Layer, (iv) Fully Connected Layer. Before giving input in this layers CNN model needs to fix features or filters for the first layer which is Convolutional Layer. As the dataset is about circle shape recognition we have taken 4 filters of 3x3 consisting of random pixel values. The further process is followed by,

1. In the convolutional layer the 4, 3x3 matrix are iterated on the whole image and created a dot product of filter matrix and every 3x3 matrix of the image and stored in another map or array.

2. ReLU is nothing but a activation function. It replaces the negative values in the matrix with zero. After ReLU activation we have 4 matrix as we have 4 filters.

3. In the Pooling layer we selected 2x2 window size. Iterated on the whole image, took the highest value in a particular window and stored in a different matrix. So we can have a shrinked image.

This 3 steps are repeated 2 times so we have 2 CNN layers and after that the image matrix size becomes shorter than before. Now the Fully Connected Layer is implemented which stores all the pixel values of 4 outcome images into one stack. Thus, when a test image is given CNN executes all the above steps and compare the stack value with the predefined values and give prediction regarding pattern recognition.After applying epoc size of and batch size of 10 the CNN model worked on our image dataset with 100% of train accuracy and and total 90.09% test accuracy.

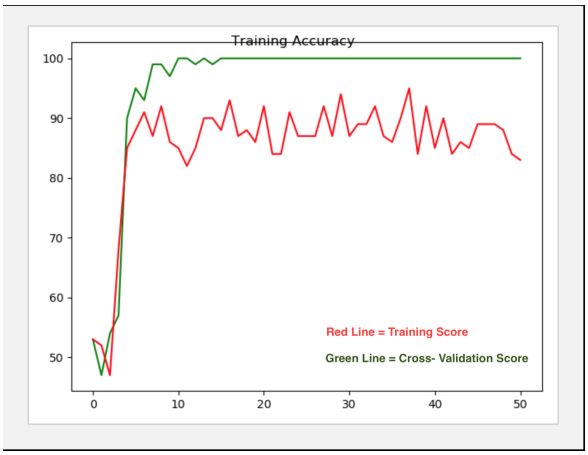


Fig. 6.11 Convolutional Neural Network (CNN) Learning Curve.

But CNN has certain disadvantages like as it has high computational cost. And it needs a lot of works to be done before initializing the problem. The overfitting issue is also present in CNN as we have less dataset. In Spite of having overfitting issue which can be solved in further improvisation of dataset CNN is performed as like as AdaBoost which also have 90% of cross validation accuracy and it less complex and needs less computational cost to perform. In this paper we made an effort to show that a supervised binary probabilistic classifier can perform with high accuracy on image dataset which has not been mentioned any of the paper that we have studied so far.

Chapter 7

Conclusion

As human race is going further with the change in the technologies, which is making our life easier. The inventions from image processing and machine learning in Medical Sector is one of the fastest growing field of technology. The main objective of our work was to come up with a system that can detect the presence of Acute Lymphocytic Leukemia in the blood cell of human body. After studying a lot of literature review, we found several methods to detect the rapid growth of white blood cell (WBC) which is a sign of affecting of Acute Leukemia. Firstly, we segmented the microscopic images of blood cell by following several steps to separate the nucleus from the blood cell. After that, we applied total 07 (seven) machine learning algorithms on our dataset. We got different accuracy rate from different algorithms and got the maximum accuracy rate (90.09%) from the Convolutional Neural Network (CNN). About 80% of the 260 data (microscopic image of blood cell) has been used to train the system and rest of the data has been used to test the system. Regardless, the accuracy of the trained system will be increased by instructing the process with more data.

7.1 Future Perspective

Though we successfully implemented our proposed model to distinguish the abnormal blood cell of a human body which is the sign of affecting of Acute Lymphocytic Leukemia, we have some minor limitations in this project. Our future perspective is to update the model by withdrawing all the limitations.

We will extend our proposed model by including the use of Support Vector Machine (SVM) algorithm. This algorithm will show us the type of the cancer that attacked in the patient

body. We already built the model however we could not show the uses of it because of the lacking of data (microscopic image of different type of leukemia affected blood sample) and the detail informations about data collection are given into the “Data Collection” part of this paper. We are collecting the data and shortly can update our model.
User friendliness is another important point to be concerned for introducing a new product in the market. At present we are circulating our project as a prototype which has to be operate through several steps to get the result; is not user friendly at all. To operate our proposed model, we need a desktop setup or a laptop to execute the whole system and sum up with a result. We operated all the methods and algorithms in Python Development Environment and the testing of datasets are also done into the desktop. However, in future when we will commercially launch our project we want to introduce a small device which can be settled down in a human hand. The reason behind this is it will not be convenient to an user to use this system by operating a desktop or a laptop again and again. Moreover, these devices take much more space in a office or diagnostic center. A concerning matter of commencement of a product commercially in market is packaging. Without a good look of a system or device, users do not seem to be interested on that product.

References

[1] Acute lymphocytic leukemia (ALL) in adults. [online] https://www.cancer.org/cancer/acute-lymphocytic-leukemia/references.html.

[2] Childhood acute lymphoblastic leukemia treatment (PDQ) health professional version. [online], https://www.cancer.gov/types/leukemia/hp/child-all-treatment-pdq.

[3] Decision tree classifier. [online] http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html.

[4] Simple guide to logistic regression in R. [online] https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/.

[5] Ferri, F. F. (2017). *Ferri's Clinical Advisor 2018 E-Book: 5 Books in 1*. Elsevier Health Sciences.

[6] Goel, E. and Abhilasha, E. (2017). Random forest: A review. *International Journal of Advanced Research in Computer and Communication Engineering*, 7:251– 257.

[7] Gupta, D. (2018). Architecture of convolutional neural networks (CNNs) demystified. *Analytics Vidhya*.

[8] Gupta, V. (2017). Image classification using convolutional neural networks in keras. *Learn OpenCV*.

[9] Hossain, M. S., Iqbal, M. S., Khan, M. A., Rabbani, M. G., Khatun, H., Munira, S., Miah, M. M. Z., Kabir, A. L., Islam, N., Dipta, T. F., Rahman, F., Mottalib, A., Afrose, S.,

Ara, T., Biswas, A. R., Rahman, M., Abedin, A. M., Rahman, M., Yunus, A., Niessen, L. W., and Sultana, T. A. (2014). Diagnosed hematological malignancies in bangladesh - a retrospective analysis of over 5000 cases from 10 specialized hospitals. *BioMed Central*.

[10] Hussain, C. A., Sushma, T., Varsa, P. S., and Manidepu, M. R. (2018). Detection of white blood cells nuclei using automatic segmentation. *International Multidisciplinary Conference on "Knowledge Sharing, Technological Advancements and Sustainable Development"(IMC2k18)*, pages 448–454.

[11] Imandoust, S. B. and Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *Int. Journal of Engineering Research and Applications*, 3:605– 610.

[12] Indira, P., Ganeshbabu, T. R., and Vidhya, K. (2016). Detection of leukemia in blood microscope images. *International Journal of Control Theory and Applications*, 9(5):63–67.

[13] Jagadev, P. and Virani, H. G. (2017). Detection of leukemia and its types using image processing and machine learning. *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, pages 522–526.

[14] Joshi, M. D., Karode, A. H., and Suralkar, S. R. (2013). White blood cells segmentation and classification to detect acute leukemia. 2:147–151.

[15] Jurek, A., Bi, Y., Wu, S., and Nugent, C. (2013). A survey of commonly used ensemble-based classification techniques. *The Knowledge Engineering Review*, 29:551–581.

[16] Koehrsen, W. (2017). Random forest in python. [online] https://towardsdatascience.com/random-forest-in-python-24d0893d51c0.

[17] Kubade, H. M. (2018). The overview of bayes classification methods. *International Journal of Trend in Scientific Research and Development (IJTSRD)*, 2:2801– 2802. [online] https://www.ijtsrd.com.

[18] Kumar, S., Mishra, S., Asthana, P., and Pragya (2018). Automated detection of acute leukemia using k-mean clustering algorithm. *CoRR*, abs/1803.08544.

[19] Labati, R. D., Piuri, V., and Scotti, F. (2011). All-idb: The acute lymphoblastic leukemia image database for image processing. In *Image processing (ICIP), 2011 18th IEEE international conference on*, pages 2045–2048. IEEE.

[20] Le, J. (2018). Convolutional neural networks: The biologically-inspired model.

[21] Longo, D., P. Hunger, S., and Mullighan, C. (2015). Acute lymphoblastic leukemia in children. *New England Journal of Medicine*, 373:1541–1552.

[22] Mohammed, M., Far, B., and Guaily, A. (2012). An efficient technique for white blood cells nuclei automatic segmentation. *IEEE International Conference on Systems, Man, and Cybernetics*, pages 220–225.

[23] Mohapatra, S., Patra, D., and Satpathi, S. (2010). Image analysis of blood microscopic images for acute leukemia detection. In *Industrial Electronics, Control & Robotics (IECR), 2010 International Conference on*, pages 215–219. IEEE.

[24] Navlani, A. (2018). Knn classification using scikit-learn. [online] https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn.

[25] O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. *CoRR*, abs/1511.08458.

[26] Patel, N. and Mishra, A. (2015). Automated leukaemia detection using microscopic images. *Second International Symposium on Computer Vision and the Internet (VisionNet' 15)*, 58:635– 642.

[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[28] Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research - J EDUC RES*, 96:3–14.

[29] Prateek (2017). Machine learning-decision trees and random forests. [online] http://dsbyprateekg.blogspot.com/2017/09/machine-learning-decision-trees-and.html.

[30] Raizada, R. D. and Lee, Y.-S. (2013). Smoothness without smoothing: why gaussian naive bayes is not naive for multi-subject searchlight studies. *PloS one*, 8(7):e69566.

[31] Ray, S. (2017). 6 easy steps to learn naive bayes algorithm (with codes in python and r). [online] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[32] Rehman, A., Abbas, N., Saba, T., Rahman, S., Mehmood, Z., and Kolivand, H. (2018). Classification of acute lymphoblastic leukemia using deep learning.

[33] Rendon, A., Gap, M., and Insew, J. (2017). Understanding and recognizing leukemia symptoms.

[34] SauceCat (2017). Boosting algorithm: Adaboost – towards data science. [online] https://towardsdatascience.com/boosting-algorithm-adaboost-b6737a9ee60c.

[35] Sayad, S. (2017). Logistic regression. [online] https://www.saedsayad.com/logistic_regression.htm.

[36] Shafique, S. and Tehsin, S. (2018). Computer-aided diagnosis of acute lymphoblastic leukaemia. *Computational and mathematical methods in medicine*, 2018.

[37] Sharma, H. and Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, 5(4):2094–2097.

[38] Smolyakov, V. (2018). Ensemble learning to improve machine learning results. [online] https://blog.statsbot.co/ensemble-learning-d1dcd548e936.

[39] Srivastava, T. (2014). Introduction to random forest – simplified. [online] https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/.

[40] Tabik, S., Peralta, D., Herrera-Poyatos, A., and Herrera, F. (2017). A snapshot of image pre-processing for convolutional neural networks: Case study of mnist. *International Journal of Computational Intelligence Systems*, 10:555.

[41] Tan, P.-N., Steinbach, M., and Kumar, V. (2006). Introduction to data mining. pages 196– 199.

[42] Thanh, T., N. Pham, G., Park, J.-H., Moon, K.-S., Lee, S.-H., and Kwon, K.-R. (2017). Acute leukemia classification using convolution neural network in clinical decision support system. pages 49–53.

[43] Thanh, T. T., Vununu, C., Atoev, S., Lee, S. H., and Kwon, K. R. (2018). Leukemia blood cell image classification using convolutional neural network. *International Journal of Computer Theory and Engineering*, 10(2):54– 58.

[44] Vaghela, H., Modi, H., Pandya, M., and Potdar, M. B. (2016). A novel approach to detect chronic leukemia using shape based feature extraction and identification with digital image processing. *International Journal of Applied Information Systems (IJAIS)*, 11:63–67.

[45] Vogado, L. H., Veras, R. D., Andrade, A. R., Aires, K. R., Silva, R. R., and Araujo, F. H. (2017). Detection of leukemia in blood microscope images. *International Journal of Control Theory and Applications*, pages 367– 373.

[46] Wahhab, H. T. A. (2015). Classification of acute leukemia using image processing and machine learning techniques.

[47] Zakka, K. (2016). A complete guide to k-nearest-neighbors with applications in python and r. *Retrieved April*, 24:2018.

Appendix A

How to run the code for segmenting the image of blood sample

Operating System: Windows

MATLAB R2018a package-Full version 64 bit

1. Go to the URL given below
   https://www.mathworks.com/products/matlab.html
2. Open an account in the URL given in the Point No. 1.
3. Save the Activation Key
4. Go to the License center
5. Choose Associate by Activation Key
6. Use the Activation Key
7. Download the software
8. Click the installer button to download the installer
9. Run matlab.exe
10. Create a new file
11. Write your code and compile

Appendix B

How to run the code for feature extraction and classify the image of nucleus

Operating System: Windows

Anaconda 5.3.1 for python 3.6- 64 bit

1. Download python 3.6 version-64 bit from https://www.anaconda.com/download/
2. Install it on your computer according to instruction
3. Open your Anaconda Prompt from the start menu
4. Navigate to the anaconda directory.
5. Install all the necessary packages from Anaconda prompt
6. Run Anaconda Navigator
7. Launch Spyder or Jupyter notebook
8. Create a new file
9. Write your code and compile

Appendix C