AUTOMATED CLASSIFICATION PROCEDURE FOR ACUTE MYELOID LEUKEMIA CELLS BASED ON BONE MARROW SAMPLES

by

LIM HUEY NEE
(1041310492)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

School of Mechatronic Engineering
UNIVERSITI MALAYSIA PERLIS

2019

ACKNOWLEDGMENT

I am grateful and greatly indebted to my supervisor, Prof. Dr. Mohd Yusoff Mashor. I have the great privilege and honor to express my whole hearted gratitude to him for his guidance, supervision, inspiring encouragement, constructive criticism and help in carrying out this thesis work. In addition, my gratitude goes to my co-supervisor, Prof. Dr. Rosline Hassan, Deputy Dean of Research, Hospital University Sciences Malaysia (HUSM) for her expert, sincere and valuable guidance on improving my research work. My gratitude is extended to the staffs of Hematology Lab of HUSM, especially Miss Selamah Ghazali and Puan Narishah for the assistance in knowledge sharing and data collection.

My sincere thanks to all the comrades and fellow research teammates of Electronicf & Biomedical Intelligent Systems (EBItS) research group in the Seriab cluster for the teamwork and friendship during hard times and sunny days.

I would like to express my deepest gratitude to my parents and sisters for sharing my pain and pleasure, providing me constant love and support. Their encouragement and sacrifice was in the end what made this thesis possible.

Thank you very much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial intelligence
ALL	Acute lymphoblastic leukemia
AML	Acute myeloblastic leukemia
ANN	Artificial neural network
ATRA	All-trans retinoic acid
BDT	Bagged decision tree
CART	Classification and regression tree
CBC	Complete blood count
CLL	Chronic lymphoblastic leukemia
CML	Chronic myeloblastic leukemia
DBC	Differential blood count
DNA	Deoxyribo Nucleic Acid
DT	Decision Tree
FAB	French American British
FFS	Filter-based feature selection
GLCM	Gray-level co-occurrence matrix
GVF	Gradient vector flow
K-NN	*K*-nearest neighbor
LM	Levenberg-Marquardt
LVQ	Learning vector quantization
MA	Mean amplitude (Gabor)
MCE	Misclassification error
MLP	Multilayer Perceptron
NN	Neural network
OOB	Out-of-bag
PCA	Principal component analysis
PSO	Particle swarm optimization
RBC	Red blood cell
RBF	Radial basis function
RGB	Red green blue
RNG	Random number generator
SD	Standard deviation
SE	Squared Energy (Gabor)

SKIZ	Skeleton by influence zones
SRG	Seeded region growing
SVM	Support vector machine
WBC	White blood cell
WHO	World Health Organization

LIST OF SYMBOLS

C_k	K^{th} cluster centre
E	Euclidean distance
$in(x,y)$	Value of the input pixel
n_k	Number of pixels belonging to centre C_k
$B(M)$	Watershed catchment basin influence zone
M	Watershed minima markers
F	Digital grayscale image
$\nabla^2 V(\underline{x})$	Hessian matrix
$J(\underline{x})$	Jacobian matrix
$I_G(\theta)$	Gini impurity
τ	Decision tree node
T	Binary decision tree
L_P	Positive Lagrange multiplier
L_D	Dual Lagrange multiplier
ξ_i	Slack variable
C	Penalty parameter
$\phi(.)$	Adjustable parameter of certain kernel function
$\bar{x}_{int}$	Mean of pixels intensity
σ_{int}	Standard deviation of pixels intensity
$P_\delta(i,j)$	Co-occurrence matrix of GLCM
$G_{s,o}$	Gabor filter at scale *s* and orientation *o*
$V_{s,o}$	Output convolution of Gabor filter
$A_{s,o}(x,y)$	Amplitude of Gabor filters
$E_{s,o}(x,y)$	Energy of Gabor filters
C	Box constraint (soft margin) of SVM RBF kernel
σ	Scaling factor sigma of SVM RBF kernel

PROSEDUR PENGKELASAN AUTOMATIK UNTUK SEL LEUKEMIA MYELOID AKUT BERDASARKAN SAMPEL SUMSUM TULANG

ABSTRAK

Leukemia adalah penyakit barah darah yang merupakan penyakit barah paling biasa di kalangan kanak-kanak berumur antara umur 0 hingga 13 tahun. Penyakit leukemia akut merebak dengan pantas dan tidak terkawal jika diagnosis dan rawatan tergendala. Diagnosis leukemia umum bermula dengan proses pemeriksaan darah diikuti dengan ujian sumsum tulang. Ujian ini termasuk ujian genetik dan molekul yang mahal dan memerlukan masa yang lama. Di samping itu, kebanyakan kajian selama ini hanya bertumpu kepada spesimen darah dan pengkelasan atas kes normal dan berpenyakit sahaja. Kajian ini melibatkan perkembangan rekabentuk prosedur pengkelasan automatik untuk *leukemia myeloid akut* (AML) berdasarkan sampel sumsum tulang. Cabaran dalam kajian ini adalah untuk mencari prosedur automatik yang paling berkesan dari peruasan imej ke proses pengkelasan yang boleh mengklasifikasikan kelas AML. Teknik peruasan imej automatik termasuk gabungan algoritma pengelompokan purata-*k* berbilang lapis, peruasan batas air dan algoritma pertumbuhan titik benih telah dicadangkan. Konsep baharu iaitu pengelompokan berbilang lapis telah dicadangkan dan dibentangkan untuk menjalankan proses peruasan imej secara automatik. Gabungan teknik peruasan imej automatik dalam kajian ini telah berjaya menyingkirkan latar belakang, sel darah merah dan sel-sel yang tidak berkenaan serta mengekalkan sel leukemia. Teknik ini menghasilkan spesifisiti 98.17%, sensitiviti 99.40% dan ketepatan 98.61%, yang mana merupakan peratusan paling tinggi berbanding dengan kaedah peruasan imej konvensional. Kajian ini juga melibatkan pengekstrakan ciri di mana 92 ciri telah diperolehi. Empat kumpulan ciri telah dipertimbangkan iaitu bentuk, warna, tekstur dan Gabor. Selain itu, dua kaedah pemilihan ciri, iaitu Pemilihan Ciri Bertapis (PCB) dan Pokok Keputusan Berkumpul (PKB) dipilih untuk memilih ciri yang menonjol dan mengurangkan jumlah ciri. Dua kaedah klasifikasi telah diuji iaitu kaedah klasifikasi berbilang dan kaedah klasifikasi berhierarki. Pengkelas MLP, SVM dan DT dipilih untuk menilai ciri-ciri yang telah diekstrak. Prestasi klasifikasi adalah berdasarkan peratus ketepatan pada data latihan, pengesahan dan ujian. Pengelas MLP telah terbukti sebagai pengkelas yang terbaik kerana ia mencapai peratus ketepatan yang tertinggi dalam kaedah klasifikasi berhierarki, iaitu 98.56% untuk latihan, 99.55% untuk pengesahan dan 97.78% untuk ujian. MLP berhierarki telah digunakan dalam kaedah pemilihan ciri. 30 ciri telah dipilih menggunakan kaedah PCB manakala 48 ciri telah dipilih menggunakan kaedah PKB. Ciri warna terbukti penting dan menonjol kerana mereka tergolong dalam sepuluh ciri paling penting dalam kedua-dua kaedah PCB dan PKB. Dalam perbandingan prestasi yang terakhir, MLP menggunakan ciri terpilih PKB telah memperolehi peratus ketepatan yang tertinggi, iaitu 98.21% untuk latihan, 98.22% untuk pengesahan dan 99.00% untuk ujian. Pada keseluruhannya, keputusan prestasi menunjukkan MLP berhierarki menggunakan ciri terpilih PKB adalah gabungan terbaik dalam rekabentuk sistem klasifikasi automatik pengkelasan AML.

Automated Classification Procedure for Acute Myeloid Leukemia Cells based on Bone Marrow Samples

ABSTRACT

Leukemia is a cancer of blood that is the most common cancer among children between the ages of 0 to 13 years old. The development of acute leukemia progress rapidly and uncontrollably if the diagnosis and treatment is delayed. The conventional leukemia diagnosis begins with screening process on peripheral blood followed by bone marrow test. The test includes genetic and molecular tests which are expensive and time consuming. Moreover, the previous works were mostly focused on peripheral blood samples and studies on screening and diagnosis, i.e. the recognition between normal and abnormal cases but not the leukemia subtypes classification tasks. This study involves in the development of automated classification procedure for acute myeloid leukemia (AML) based on bone marrow samples. The challenge in this study is to find the best automated procedure from image segmentation to classification that can classify the subtypes of acute myeloid leukemia. Automated image segmentation technique based on the combination of multilayer *k*-mean clustering algorithm, watershed segmentation and seeded region growing algorithm has been proposed. A new concept of multilayer clustering procedure is proposed and presented to implement the idea of automated image segmentation process. The combination of multilayer clustering, watershed and seeded region growing method successfully eliminates background, red blood cells and unwanted cells while retaining the cells of interest. The proposed image segmentation method achieves the highest specificity of 98.17%, highest sensitivity of 99.40% and highest accuracy of 98.61%, where it is the highest percentage as compared to the conventional segmentation method. The study also includes feature extraction where 92 features are extracted in classifying AML subtypes. Four feature categories are considered namely simple shape descriptors, colour, texture, and Gabor features. Besides that, two feature selection methods, Filter-based Feature Selection (FFS) and Bagged Decision Tree (BDT) are chosen to select prominent features and reduce feature dimensionality. Two classification methods are tested, namely multiclass classification and hierarchical classification method. The extracted features and classification methods are fed into three classifiers which are Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Decision Tree (DT). The classification performance is based on the accuracy of training, validation and testing data. MLP has shown to yield good classification and generalization performance as it achieves accuracies of 98.56%, 99.55% and 97.78% for training, validation and testing data respectively. Hierarchical MLP classifier is used in feature selection process. 30 features are selected using FFS method while 48 features are selected using BDT method. Colour features have proven to be prominent and important as six of the colour features are among the top ten selected features. MLP with BDT feature set achieved the highest accuracy of 98.21%, 98.22% and 99.00% for training, validation and testing data respectively. Overall, the performance comparison results indicate that MLP implemented with BDT feature set is the best option for AML subtypes classification.

CHAPTER 1

INTRODUCTION

1.1 Background

Leukemia, also known as cancer of the blood, is one type of cancers that is aggressive and rate of mortality will be increased if it is not controlled in its early stage (Theml *et al.*, 2004). Leukemia is a malignant disease of the blood that is characterized by an abnormal accumulation of immature white blood cells and its precursors which disable its normal function of fighting infections. In normal condition, the production and derivation of blood cells occur in bone marrow, follow by the release of mature leukocytes into the lymphatic system to defend the body against infectious disease and foreign materials. These mature leukocytes are essential in the immune system. In the body of a leukemia patient, the cell lineage retains the ability to proliferate but differentiation to mature cells is compromised. As a result, excessive accumulation of immature cells occurs and the cells can be spread to other vital organs through the blood circulation system in the body (Haferlach *et al.*, 2005).

In Malaysia, cancer has been classified as one of the major causes of health problems, with the occurrence of 100,000 cases each year (Azizah *et al.*, 2016; Omar *et al.*, 2006). Generally, cancer can be categorized into several types including organ induced cancer such as liver and kidney, and blood induced cancer. Blood induced cancers include leukemia, lymphoma and myeloma. According to the Malaysian National

Cancer Registry Report 2007 – 2011 that was published by the Ministry of Health Malaysia, leukemia is among the top ten most common cancers among Malaysians include breast cancer, colorectal cancer, lung cancer and lymphoma. Leukemia is the top most commonly occurred cancer among the age group of 0 – 24 years old. It was reported that the occurrence of leukemia was as high as 47.1% among the young male cancer patients while 45.5% among the young female cancer patients in children below age of 14 years.

There are two main types of leukemia which are acute leukemia and chronic leukemia. Acute leukemia occurs for about 85% of childhood leukemia cases while chronic leukemia is commonly found among adults (Azizah *et al.*, 2016). Acute leukemia progresses rapidly and can cause fatality if it is not treated immediately. There are two types of acute leukemia, namely acute myeloid leukemia (AML) and acute lymphoid leukemia (ALL). AML can be classified into eight subtypes according to the French-American-British (FAB) system, namely M0, M1, M2, M3, M4, M5, M6 and M7. This research focuses on the AML and its subtypes. ALL is excluded from this study due to the lack of FAB-classified samples available.

1.2 Problem Statement

Accurate and fast diagnosis of leukemia is important for optimal treatment. Early diagnosis of the disease is of paramount to ensure treatment can be pursued promptly and expecting good prognosis. Analysis by (Padilha, *et al.*, 2015) stated that the median overall survival is 7.4 months for M4, M5, M6 and M7; 23.5 months for M0, M1 and M2, and 97.7 months for M3. The survival rate of subtype M3 was the highest as

compared to other subtypes, hence the confirmation and classification of subtype M3 is to be emphasized.

The conventional leukemia diagnosis begins with a screening process on peripheral blood smear where the total number of normal and abnormal white blood cells (WBCs) is determined. This process is also known as complete blood count (CBC). If the blood test suggests the possibility of leukemia, it is followed by the bone marrow test where differential blood count is executed through microscopic examination. When the case is suspected of leukemia according to the FAB system, genetic and molecular test will be conducted for diagnostic confirmation (Falini *et al.*, 2010). Generally, the microscopic examination on peripheral blood smear cannot be used to conclude and confirm the leukemia diagnosis. Furthermore, the genetic and molecular test which are required to define the prognosis are costly and time consuming. Rapid diagnosis of acute leukemia is required for early treatment of patients. Many countries do not have the facility and equipment to conduct genetic tests to classify leukemia according to the prognostic biomaker as it is expensive (Basso *et al.*, 2007). Besides, the method also requires a well-trained hematologist and the diagnosis results are highly dependent on the expertise and experience of the hematologist. The equipment required to run molecular and genetic tests are expensive and time consuming. The importance to identify the good prognosis of AML subtypes and the technical limitation of bone marrow morphology features have stressed the need for a computer-aided system that able to automate the AML subtypes classification process.

Advances in computer hardware, software and image processing algorithms have led to implementation of computer-aided system to leukemia diagnostic. Several attempts

to develop computer-aided diagnosis of acute leukemia were carried out by researches. By providing quantitative measurements and objective-based decisions, computer-aided leukemia diagnostic system facilitates rapid decision making and contributes to minimizing diagnosis error. However, most studies have focused the attention on peripheral blood samples (Reta *et al*., 2015). Moreover, the studies were conducted on diagnosis phase, e.g. between normal case and abnormal case, or between ALL and AML. Very little attention has been given to date on bone marrow samples and AML subtypes classification. There were several studies on AML classification using peripheral blood images, mainly focused on AML subtypes M2, M3 and M5 (Harjoko *et al.*, 2018; Kahaki *et al.*, 2017; Mohamed *et al.*, 2018; Rawat *et al.*, 2017; Sarrafzadeh *et al.*, 2015; Tran *et al.*, 2016). Examination of bone marrow sample is more difficult compared to that of peripheral blood, due to the complexity and density of bone marrow structures and cells components. Presently, computer-aided diagnosis of AML using bone marrow samples can only be found in Francis *et al.* (2011) and Reta *et al.* (2015). Unfortunately, both studies concentrated on detection and identification of AML using morphological features by classifying between normal and non-AML classes.

The assessment of peripheral blood sample can only be used in screening purpose. Leukemia diagnosis confirmation can only be done using bone marrow sample. The main aim of this study is to extend the leukemia diagnosis scope into AML subtypes classification. The subtypes of AML is important to help determine the patient's prognostic factors. Prognostic factors help doctors determine the best treatment, a more or less intensive treatment and the risk of relapse after treatment. Therefore, the subtypes classification for AML is of paramount. The primary commitment of this work is the utilization of classification procedure for the recognition of AML subtypes, i.e M1 to M6

using a prominent set of features accompanied with intelligent classifiers which has not been done before. This fact can be considered as the work novelty.

1.3 Research Objectives

The main objective of this study is to develop an automated classification procedure for acute myeloid leukemia cells based on bone marrow samples. This main objective covers the following sub-objectives:

i) To develop an automated image segmentation procedure for AML bone marrow samples based on unsupervised method.

ii) To propose and evaluate the significance of morphological and statistical features for AML subtypes classification.

iii) To evaluate and identify the best performing intelligent classifiers and classification concept for AML subtypes classification.

1.4 Research Scopes

This research focuses on developing an automated classification procedure for AML cells based on bone marrow samples. The procedure includes image processing technique and classifier to perform the classification. All the analyses are performed offline. The microscopic images are limited to the images of bone marrow obtained from a light microscope with 100× objective, and stained by Wright-Giemsa stain. This study focuses only on the AML and its subtypes, subjected to the availability of the bone marrow samples. The samples collected for this study were of subtypes M1, M2, M3, M4, M5, and M6. The subtypes M0 and M7 were excluded from this study due to lack

of samples. For initial study, performance evaluation of image segmentation and AML subtypes classification are based on 20 de-identified bone marrow slides which were prepared and validated by two experienced haematologists from Hematology Department, Hospital Universiti Sains Malaysia (HUSM), Kelantan, Malaysia. The subtypes of the samples have been confirmed according to its flow cytometry, genetic and molecular tests by HUSM. To avoid color degradation problem of the samples, the bone marrow smears are collected within three years from the date of smear preparation.

1.5 Thesis Organization

This thesis is organized into five chapters including this introduction chapter. Chapter 1 starts with brief introduction and overview of the research. Explanation in this chapter includes introduction, problem statement, research objective as well as the research scope.

Chapter 2 focuses on literatures related to the research. This chapter continues presenting descriptions of leukemia in detail. The descriptions include types and classification of leukemia, risk factors and symptoms as well as the current clinical diagnosis methods for leukemia. Then, a brief overview of digital image processing is described. Some examples of medical application using image processing are described. The last part of this chapter provides an overview of intelligent classifiers and their applications in several areas of medical field.

Chapter 3 describes the details of the proposed methodology. The method generally comprises of five main parts namely image acquisition, image segmentation,

feature extraction and selection, and classification. In image acquisition, the method and equipments used for capturing bone marrow AML slide images are explained. In image segmentation part, a new procedure for segmenting AML images, which combines the multilayer *k*-mean clustering algorithm, watershed segmentation and seeded region growing is introduced. In feature extraction and selection stage, four types of potential features namely simple shape descriptors, colour, texture and Gabor features are described. Two feature selection methods are applied to find the optimal subset of features and remove the irrelevant features. In classification stage, the classification methods, classifier input data preparation and the parameters of the classifiers used in this research are presented.

Chapter 4 presents and discusses the experimental results of the evaluation. The chapter begins with the results of proposed automated image segmentation methods and its segmentation performance evaluation. Next, the classification performances of the classifiers are compared and discussed based on various features subsets. The chapter is concluded with the best feature subsets as well as the best method used in classifying AML subtypes.

Finally, Chapter 5 draws the overall conclusions and highlights the main contributions of the thesis. Also, some possible improvements and direction of future research works are suggested at the end of the chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Cancer of the blood or blood cancer is an umbrella term for cancers that affect the blood, bone marrow and lymphatic system. Leukemia, together with lymphoma and myeloma, are the three main groups of blood cancer that is common in society (Kahaki *et al.*, 2017; Rawat *et al.*, 2017; Rodellar *et al.*, 2018). Generally, leukemia is divided into two types which are acute leukemia and chronic leukemia. The current issues and pitfall in leukemia diagnosis is that, 1) for an accurate diagnosis and classification, it requires few other blood investigation and not limited to morphology alone; 2) at times, by morphology examination, second opinion is required; 3) this will interfere with the duration needed to visually inspect the image due to the repetitive work. All of the processes eventually delay the treatment. Besides that, the process of inspecting images under the microscope is straining to the eyes (Safabakhsh and Zamani, 2006; Scotti, 2006; Theera-Umpon and Dhompongsa, 2007).

Studies by (Rajendran *et al.*, 2008) stated that among the reasons that burden the workload is the information management, retrieval and storage of leukemia patient data which are done manually despite the increment of cases handled every year. Therefore, currently many researchers focus on digital image processing technique on the medical images as well as automation of the visual image analysis process (Agaian *et al.*, 2014; Díaz and Manzanera, 2015; Putzu *et al.*, 2014; Sarrafzadeh *et al.*, 2015). Morphological

analysis and utilizing the designated equipment in the hematology laboratory are useful in classifying the leukemia types. Moreover, the classification of the leukemia types depends highly on the experience and expertise, the quality of the blood and marrow smears prepared and the usage of light microscope with good optical features.

This chapter is divided into three major sections. The first section provides a brief description of leukemia. Types of leukemia, causes and risk factors, sign and symptoms, and finally the explanation of current trend for leukemia diagnosis and classification are briefly discussed. The second section describes the field of digital image processing and existing techniques used to process the digital images. Application of digital image processing in medical field using existing technique is also highlighted. The fundamental of *k*-means clustering and watershed segmentation is presented here. Finally, the last section highlights on an overview of classification as well as their application in various areas of medical field.

2.2 Leukemia

The term leukemia is derived from the Greek words for white *(leukos)* and blood *(haima)*. Leukemia generally describes a kind of blood disease where immature white blood cells are abundant. The production of these immature white blood cells begins in bone marrow. The excessive immature white blood cells, or blasts in another expression, are produced as a result of abnormal and damaged Deoxyribo Nucleic Acid (DNA) in the white blood cells (WBC) (Theml *et al.*, 2004).

When a person has leukemia, the bone marrow starts proliferating blast cells while the maturation of these cells is halted (Ravandi & Giles, 2009). The overcrowding of blasts makes the bone marrow unable to develop healthy WBCs and interferes with the functionality of WBCs. Over time, the crowding blasts cells will spill out into the circulatory system and this can lead to serious problems such as anemia, bleeding and infection. It may also affect other organs of the body. Despite the lack of understanding on the cause and prevention, it is likely that several risk factors are involved and several symptoms that might suggest leukemia. Blood and bone marrow tests are important in detecting and diagnosing leukemia. The process of blood progression in bone marrow of a normal person and person with leukemia is explained in the following sub-section.

2.2.1 Blood and Bone Marrow

Blood is a unique tissue in the circulatory system and it is fundamental to human life. Blood consists of straw-colored plasma and several suspended cells such as red blood cell (RBC, erythrocyte), white blood cell (WBC, leukocyte) and platelet (thrombocyte). The total blood volume of an adult male who weighs 70 kg is approximately five to six litres and accounts for eight percent of the body weight (Miwa, 1998).

The production of blood cellular components including formation, development and differentiation is called haematopoiesis. This process originates from bone marrow which is part of the bone in the human body. Bones are made up of two regions: the hard, calcium-based outside shell or compact tissue that gives structure of the human body; and the cancellous tissue or medullary cavity which contains fat cells, fluid, fibrous tissue,

blood vessels and blood forming cells (Barbara *et al.*, 2001; Clarke, 2008). The anatomy of the bone is shown in Figure 2.1.

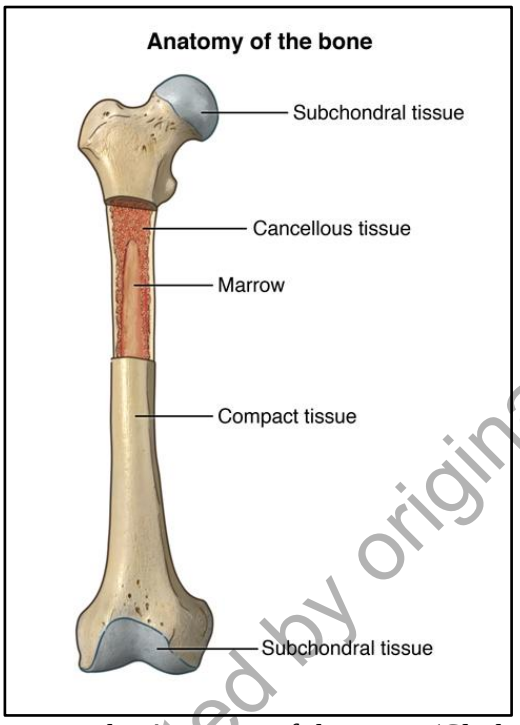


Figure 2.1 The Anatomy of the Bone (Clarke, 2008)

The variety of blood cells in the body is derived from an undifferentiated bone marrow cell. This cell, which is named haematopoiesis stem cell, is able to produce any type of precursor cells (Miwa, 1998; Theml *et al.*, 2004). These precursor cells then give rise to all of the different blood lineages and undergo proliferation and maturation into mature blood cells, as shown in Figure 2.2. The red box shows the cells where AML develops from.

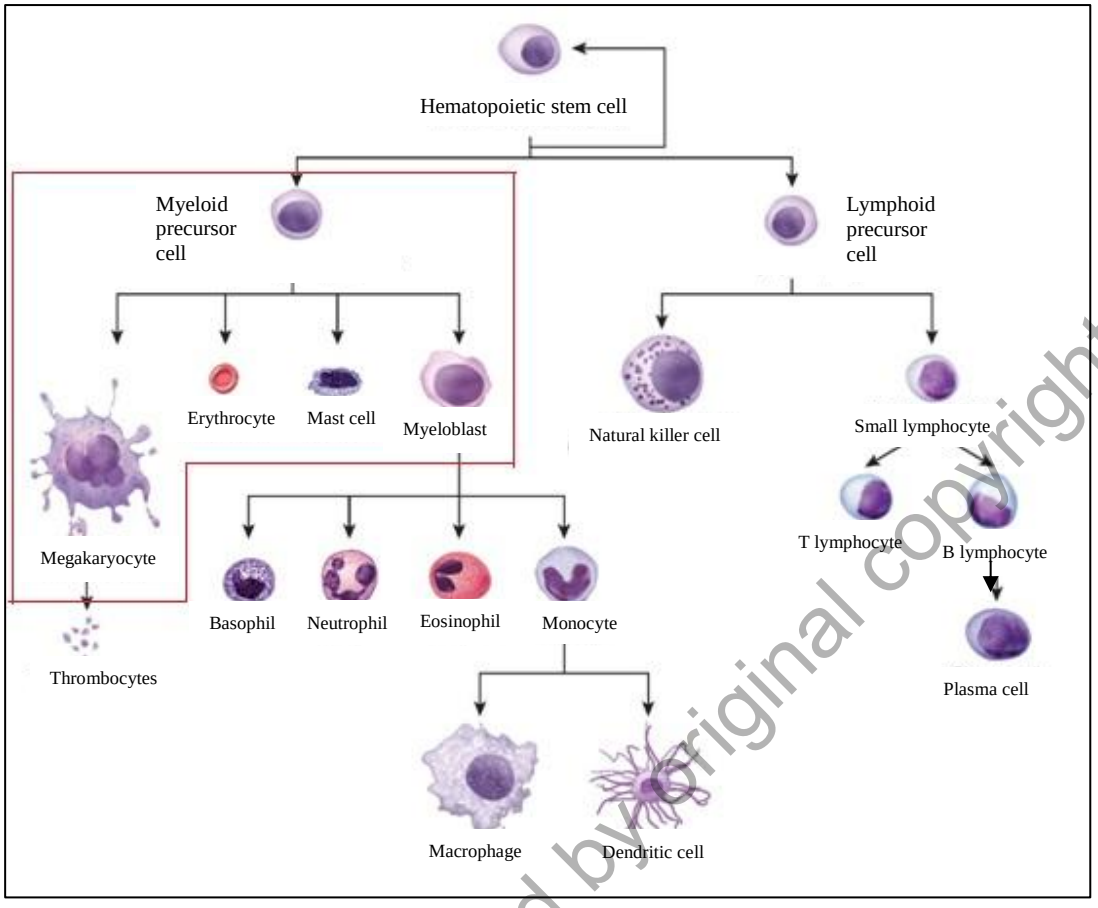


Figure 2.2 Blood Cell Development from Differentiation of Hematopoietic Stem Cell (Miwa, 1998)

The four main elements of blood cells that are formed in the bone marrow are (Bell & Sallah, 2005):

i) Red blood cells– transport oxygen from the lungs to organs and peripheral site.

ii) White blood cells – body's main defence system in destroying invading organisms such as bacteria and viruses, fight infections and assist in the removal of dead or damaged tissue cells.

iii) Platelets – responsible for blood clotting.

iv) Plasma cells – carries nutrients, antibodies and the proteins involved in blood clotting.

Biologically, blood is essential in maintaining homeostasis, or a steady state, a continual balancing act of the body systems to provide an internal environment that is comparable with life (Bell & Sallah, 2005). It involves actions like hydration, temperature regulation and ion concentration. The main blood functions include:

i) Transportation of oxygen and carbon dioxide, chemical substances (hormones, nutrients, salts), and cells that defend the body.

ii) Regulation of the body's fluid and electrolyte balance, acid-base balance and body temperature.

iii) Protection of the body from viral and bacterial infection.

iv) Protection of the body from loss of blood by the action of clotting.

2.2.2 Types and Classification of Leukemia

The blood cell progression starts from stem cell and differentiate into two groups: the myeloid lineage and the lymphoid lineage. WBCs from either lineage can be affected by leukemia. Leukemia that affects the myeloid lineage is called Myeloblastic Leukemia (AML & CML) while leukemia that affects the lymphoid lineage is called Lymphoblastic Leukemia (ALL & CLL) (Haferlach *et al.*, 2005). Acute Leukemia (AML & ALL) progresses rapidly and may cause discomfort and sickness almost immediately. Chronic Leukemia (CML & CLL) on the other hand develops slowly over time and it may take years for the symptoms to surface. Generally, there are four main types of leukemia (Theml *et al.*, 2004):

i) Acute Myeloblastic Leukemia (AML) – Affects both adults and children.

ii) Acute Lymphoblastic Leukemia (ALL) – Most common leukemia in children.

iii) Chronic Myeloblastic Leukemia (CML) – Occurs mainly in adults.

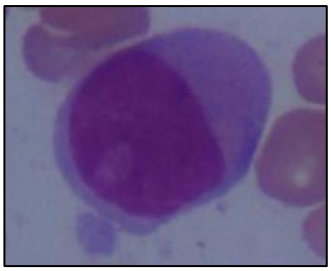
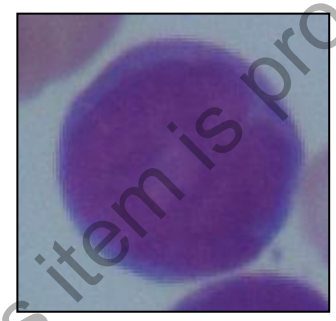
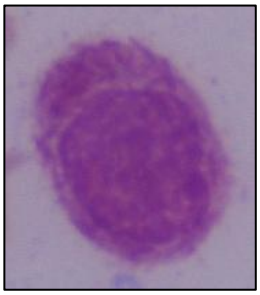
iv) Chronic Lymphoblastic Leukemia (CLL) – Mostly affects adults older than 55, almost never affects children.

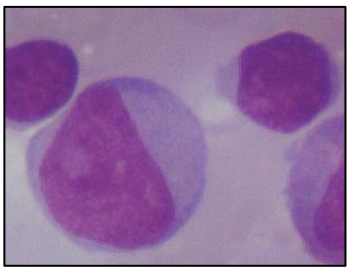
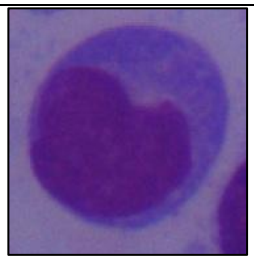
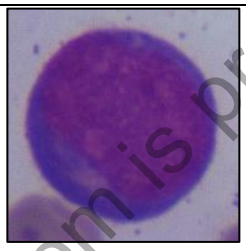
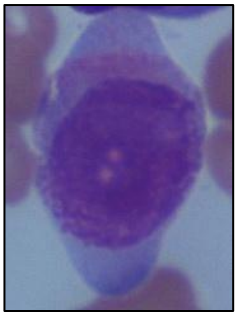
For most types of cancer, the staging describes the severity of a patient's cancer based on the magnitude of the primary tumour and on the extent cancer has spread in the body. On the other hand, AML does not form tumours. It is generally widespread throughout the bone marrow and has spread to other organs such as the liver and spleen. Therefore, AML is not staged like most other cancers, instead is classified into subtypes as it affects both a patient's prognostic factor and the best treatment.

Two types of classification schemes for AML are commonly used, the French-American-British (FAB) system and the World Health Organization (WHO) system (Bennett *et al.*, 1976; Falini *et al.*, 2010; Verdiman *et al.*, 2002). FAB classification system is based on morphology to define specific immunotypes while WHO classification reviews chromosome translocation and evidence of dysplasia. In this context, FAB classification relies on the morphological outlook of leukemia cells under microscope, whereas WHO classification divides AML into molecular stage (Schoch & Haferlach, 2002). This study follows the FAB classification system, which classifies acute leukemia by morphology and cytochemistry supplemented by immunophenotyping. In 1976, the FAB classification system was developed by the FAB Cooperative Group based on marrow and blood morphology and cytochemical staining (Bennett *et al.*, 1976). The system defines groups of AML based on the percentage of maturing cells beyond the myeloblast stage, dividing AML into eight subtypes, M0 through M7. It remains the most accepted classification system of AML (Bell and Sallah, 2005). Subtypes M0 through M5 start in immature form of white blood cells. M6 starts

in immature form of red blood cells while M7 starts in immature form of platelets. The description of each AML sub-types is presented in Table 2.1.

Table 2.1 The Description of AML Subtypes

M0: Undifferentiated myeloblastic leukemia	
	The blasts are large and agranular (resemble ALL L2), rounded nucleus, fine chromatin and prominent nucleoli.
M1: Myeloblastic leukemia with minimal maturation	
	The blasts are variation in size with agranular cytoplasm, prominent nucleoli and regular nuclear outline. More than 90% of total WBCs population are M1 type myeloblast cells (granular and agranular types). It is found in all aged group with highest incidence seen in adult and in infants less than a year old.
M2: Myeloblastic leukemia with maturation	
	The blasts are moderate to large size with moderate amount of granular cytoplasm and irregular nuclear outline. 30% to 89% of total WBCs population are M2 type myeloblast cells (granular and agranular types). The presenting symptoms are similar to those of the M1 type yet the distinguishable characteristic between these two types is the presence of maturation at or beyond the promyelocyte stage.
M3: Promyelocytic leukemia (APML)	
	Abnormal promyelocytes with heavy granulations, cells contain bundles of Auer rods. The granules in the cell's cytoplasm are toxic and can be fatal if the cytoplasm burst and the granules flow into the circulatory system causing bleeding abnormalities (Mittal and Meehan, 2001). The identification of this sub-type is clinically important as it usually has the best prognosis

	should treatment starts instantly. The median age and survival average of APML is about 18 months and occurred in younger adult.
M4: Myelomonocytic leukemia	
	There are two populations of blasts in this sub-type: Monoblast (left) – larger in size, moderate bluish agranular and double membrane cytoplasm, irregular nucleus outline. Myeloblast (right) – smaller in size, agranular cytoplasm, high nucleus/cytoplasm ratio. It is distinguishable from M1, M2 and M3 by an increased proportion of monocytic cells in the bone marrow or blood or both.
M5: Monoblastic monocytic leukemia	
	80% of marrow precursors in this sub-type are monoblasts, promonocytes or monocytes. The morphological characteristics of the blast cells are double membrane cytoplasm and perinucleur 'haloes' with kidney-shaped nucleus
M6: Erythroid leukemia	
	The blasts are moderate to large in size with abundant bluish cytoplasm. More than 50% of marrow cells are erythroid (RBC) precursors or called as erythroblasts. It is a rare form of leukemia that primarily affects the peripheral cells and non-sexist in children.
M7: Megakaryocytic leukemia	
	The blasts are of megakaryocytic lineage (platelet), medium to large in size, irregular cytoplasmic borders and cytoplasm blebbing (represents early platelet formation), hence seems to have two different layers of cytoplasm. The diagnosis of M7 should be suspected when the blasts show such features. M7 is rare and it usually occurs as a leukemia transformation of chronic granulocytic leukemia and myelodysplastic syndrome (MDS).

2.2.3 Risk Factors and Symptoms of Leukemia

To this day, the real cause of leukemia is still largely unknown; however, the researchers and scientists in the medical division have identified the possible risk factors. The combination of genetic and environmental factors is considered to be associated with an increased risk of leukemia since leukemia results from the mutation in the DNA. The possible risk factors include (Deschler and Lubbert, 2006):

i) Environmental factors such as high radiation exposure, such as the atomic bombing incident in Hiroshima and Nagasaki, Japan. However, the usage of *x*-ray machine with low radiation dosage is not harmful to human.

ii) Exposure to certain chemicals in high volume and long-time duration. For example, benzene, a type of colorless and highly inflammable liquid that is found in gasoline and glue. Other chemical examples include pesticide used in agriculture to kill insects, herbicide used to defoliate or slow down the growth of weed and formaldehyde used in manufacturing industry and household products (Checkoway*et al.*, 2015).

iii) People who have had chemotherapy and radiation therapy for other cancers.

iv) Genetic disorders such as Down syndrome or myelodysplastic syndromes (MDS).

v) Family history of leukemia.

The aforementioned risk factors highlight the possibility of certain people to have an increased risk in having leukemia. Although it is not clinically proven that the risk factors are the main cause of leukemia, people who are exposed to any of the risk factors need to be extra cautious of the possibility of deteriorating health condition for early detection. The clinical manifestations are similar among the different types of leukemia. It causes lack of platelets in the blood, preventing blood clotting, thus causing easy

bleeding, bruising and recurrent nosebleed. Dysfunctional immune system makes people with leukemia vulnerable to infections. Swollen lymph nodes and enlarged spleen and liver causes the patients to experience nausea or a feeling of fullness, resulting in unintentional weight loss. Some other signs and symptoms are fatigue, weakness, fever and night sweats. It could cause neurological symptoms such as headache if leukemic cells invade the central nervous system.

2.2.4 Current Laboratory Diagnosis of Leukemia

A correct and rapid diagnosis of leukemia is of utmost importance for optimal treatment. The diagnostic gold standard and classification of leukemia involves various methods such as morphological studies, cytochemistry, cytogenetics and molecular analysis, immunophenotyping and molecular biology (Basso *et al.*, 2007). Morphology forms the initial diagnosis of leukemia, follow by cytogenetic analysis and flow cytometric studies from bone marrow samples (Bell and Sallah, 2005).

2.2.4.1 Peripheral Blood Test

Peripheral blood tests are used to evaluate the type and quantity of blood cells that are present, the blood chemistry and other factors. Blood samples are generally obtained from a vein in the arm and the examination is performed manually by visualizing and counting the cells through a light microscope after the blood sample is smeared on a slide and stained. The hematologist or technician will identify and count the cells in a systematic prescribed method (Padilha *et al*., 2015). Although the method is generally

accurate, the process is very technologist-dependent and time consuming. The peripheral blood tests include:

i) Complete Blood Count (CBC) – It is the widely requested and single most important lab test on blood. CBCs are done as a routine screening that provides general information about the patient's status. The cellular measurements include WBC count, which counts the total number of WBCs present regardless of the types, and 3-part differential count which reports neutrophils, lymphocytes and monocytes. These are the most useful cells to count and they make up 80% of WBC.

ii) Differential Blood Count (DBC) –It is a count of the number of WBCs present in a known volume of blood. The WBC count is further identified by the major WBC sub-populations, thus named as the differential count. DBC is normally measured by hematology analysers. If the analyser indicates abnormalities, the technologist may perform microscopic smear review or an actual manual differential count for accuracy of the reported results (Basso *et al.*, 2007).

2.2.4.2 Bone Marrow Tests

Although peripheral blood tests are invaluable in routine screening and initial diagnosis, further tests are required should a person is suspected leukemia. For diagnostic confirmation and sub-types identification, a needle biopsy and aspiration of bone marrow from pelvic bone needs to be done to test for leukemic cells, DNA markers and chromosome changes in the bone marrow (Barbara *et al.*, 2001; Felming *et al.*, 2003). It is done by studying the cell morphology and bone marrow architecture. Figure 2.3 graphically depicts the process of obtaining a bone marrow sample. The bone marrow

biopsy and aspiration are often extracted at the same time but each method targets different sample. Bone marrow biopsy extracts a sample of the actual marrow from inside bone while bone marrow aspiration removes a small amount of the marrow tissue in liquid form. Besides diagnosis, bone marrow samples may also be used to evaluate if the cancer is not in remission or responded to treatment.

Although the most recent WHO classification relies more on chromosomal abnormalities than on morphology, a correct morphological diagnosis remain the basis and is essential for correct initial treatment (Ravandi and Giles, 2009; Voute, 2012). It is essential to quickly distinguish AML from ALL in case of a high WBC count and the need to install anti-leukemia treatment immediately. Furthermore, the early morphological recognition of AML sub-type M3 or even the suspicion of M3 should prompt immediate treatment of all-trans retinoic acid (ATRA). One of the life threatening complications in patients with M3 is the early bleeding and it is fatal in 5-10 per cent of children with M3. ATRA is important in the initial treatment of M3 because it quickly reduces the risk of bleeding complications. In addition, continual use of ATRA makes a very important contribution to the cure of most M3 patients with contemporary treatment.

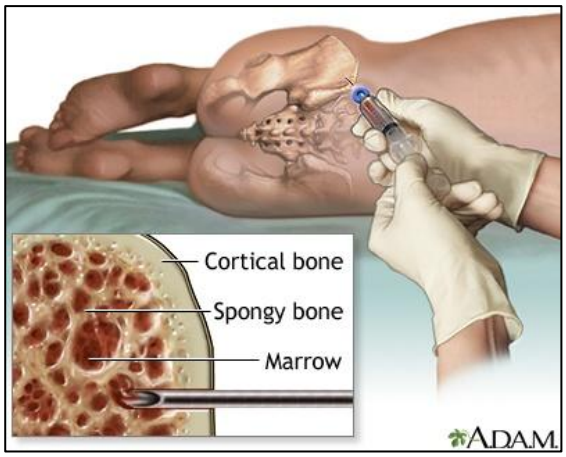


Figure 2.3 Bone Marrow Sample Extraction Procedure (A.D.A.M., 2014)

2.2.4.3 Molecular and Genetic Tests

Besides examination of the peripheral blood and bone marrow samples, additional tests such as flow cytometric immunophenotyping and genetic analysis are used to further classify the types of leukemia. The utility of flow cytometry in immunophenotyping is to count the total number of cells, identify the components and structural features of the blood cells. It is also used to measure each cells and is able to process thousands of cells in the matter of few seconds. One of the crucial importance of flow cytometry is the distinction between lymphoid and myeloid leukemia (Weir and Borowitz, 2001). Flow cytometry comprises of three main systems: fluidics, optics and electronics system. The fluidics system transports the cell suspension in a stream to a laser beam one cell at a time. Lasers and optical filters in the optics system illuminate the cells and direct the resulting light signals to the detectors. The detected light signals are converted into electronic signals to be processed by the computer. The experts analyse the difference and sorting features in the scatted lights to identify the cells existed in the blood sample (Basso *et al.*, 2007; Haferlach *et al.*, 2005). Immunophetotyping is often done with multi-color flow cytometry in high-income countries, but can be done on slides if such facilities are lacking.

In AML, genetic testing focuses on leukemia-specific clonal and prognostic markers. Standard cytogenetic analysis is a method of detecting and identifying clonal aberrations in AML patients. The procedure involves specimen preparation and analysis on the total number and the shape of chromosome structure. Cytogenetics has been used to develop the progression or regression of abnormal cell lineage, therefore it has become

a key part of acute leukemia evaluation, especially in assessing prognosis (Mrózek *et al.*, 2004; Schoch and Haferlach, 2002).

Polymerase chain reaction (PCR) is a very sensitive DNA replication technique that detects changes in the structure or function of genes and chromosome, particularly the mutation, inversion, fusion and deletion condition on the DNA. This technique only looks for certain genetic or chromosome changes, such as the PML-RARα fusion gene in acute promyelocytic leukemia M3, AML1-ETO in acute myeloid leukemia and BCR-ABL in chronic myeloid leukemia. Due to its specific function and the reliability of detecting leukemic cell in highly populated cells sample, PCR is very useful in monitoring the progress of the disease after treatment.

2.3 Digital Image Processing

Image is a visual representation of an object or thing. A digital image is a two-dimensional array of pixels. Pixel is the term most widely used to denote the elements of a digital image. The value of each pixel is proportional to the intensity of the corresponding point in the scene. An image may be described as $N \times N$ m-bit pixels, where N is the number of points and m controls the number of brightness values. The m bits give a range of 2^m values, ranging from zero to $2^m - 1$. For example, in a grayscale image, each pixel of an eight bit grayscale image gives brightness level ranging between zero for black and 255 for white, with shades of grey in between. Colour image follows a similar storage strategy to specify pixels' intensities. However, color image are represented by three intensity components instead of using just one image plane. These components generally correspond to red, green and blue (RGB), although there are other colour schemes. The

most common storage format for a color image uses eight bits for each of the three RGB components, thus it has 24 bits of intensity resolution (Nixon and Aguado, 2008; Shih, 2009).

Digital image processing refers to a method of transformation and modification of digital image, in order to get an enhanced image or to extract useful information from it. It is a type of signal manipulation of pixels in which input is image such as individual video frame or photograph, and output may be image or characteristics associated with the image(Gonzalez and Woods, 2009; Hlavac, 2011). The commonly used methods in image processing include enhancement, segmentation, noise filtering, image analysis and so on, depending on the users' requirements. The application of image processing has been widely explored in various fields, such as computer vision, face and fingerprint recognition, medical imaging and exploration geophysics.

Gonzalez and Wood (2009) stated that image segmentation falls into the mid-level computerized processes in the continuum of image processing, where it partitions an image into regions or object. The mid-level processes involve also extraction of object descriptors and individual object classification (Gonzalez and Woods, 2009). Image segmentation is at the first stage of the classification system, and it is considered a critical and essential procedure for successful feature extraction and classification of acute leukemia in later stage. In computer vision, segmentation is defined as a process of partitioning a digital image into regions or objects. The goal is to simplify the representation of an image into something meaningful and easy to analyse. In general, the more accurate the segmentation, the more likely recognition is to succeed. On the other hand, weak or erratic segmentation algorithms always guarantee eventual failure.

Medical imaging analysis has been one of the main targets in image processing continuum. Many proposals have been presented for the benefit of analysing medical images efficiently and accurately, mostly in the context of screening and diagnosis. In this chapter, the fundamental of clustering and watershed segmentation is covered as it is the main frame of the image processing methods in this thesis.

2.3.1 Clustering

Clustering or cluster analysis is the process of segregate groups with similar traits and assign them into cluster while different groups are as dissimilar as possible from one another. Clustering is an unsupervised classification that has no predefined classes. Clustering algorithms can be categorized based on their cluster model. There are four most prominent clustering algorithms namely connectivity model, centroid model, distribution model and density model. In this study, *k*-means clustering that falls into the centroid model is implemented.

K-means clustering originates from signal processing, and it is a popular machine learning and data mining algorithm. It is a partitioning method, where its main aim is to partition observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. Due to its easy implementation and the ability to apply on large datasets, *k*-means clustering is usefully applied in various fields such as computer vision, astronomy, marketing and agriculture (Ng *et al.*, 2008).

The standard algorithm was first introduced by Lloyd (1957) as a least squared quantization technique for pulse-code modulation (Lloyd, 1982). A simple iterative