

**CLASSIFICATION OF ACUTE LEUKEMIA USING IMAGE
PROCESSING AND MACHINE LEARNING TECHNIQUES**

HAYAN TAREQ ABDUL WAHHAB

**THESIS SUBMITTED IN FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY**

**FACULTY OF COMPUTER SCIENCE &
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA,
KUALA LUMPUR**

2015

DECLARATION

I declare that this thesis is my own work and has not been submitted in any form for another degree at any university or other institute of tertiary education. Information derived from the published and unpublished work and others have been acknowledged in the text and a list of reference is given.

Date:

Hayan Tareq AbdulWahhab

WHA070026

ABSTRACT

Medical diagnosis is the procedure of identifying a disease by critical analysis of its symptoms and is often aided by a series of laboratory tests of varying complexity. Accurate medical diagnosis is essential in order to provide the most effective treatment option.

The work presented in this thesis is focused on processing of peripheral blood smear images of patients suffering from leukemia based on blast cells morphology.

Leukemia, a blood cancer, is one of the commonest malignancies affecting both adults and children. It is a disease in which digital image processing and machine learning techniques can play a prominent role in its diagnostic process.

Leukemia is classified as either acute or chronic based on the rapidity of the disease progression. Acute leukemia can be further classified to acute lymphoblastic leukemia (*ALL*) and acute myeloid leukemia (*AML*) based on the cell lineage. The treatment protocol is allocated based on the leukemia type. Fortunately, leukemia like many other cancer types are curable and patient survival and treatment can be improved, subject to accurate diagnosis. In particular, this research focuses on Acute Leukemia, which can be of two distinct types (*ALL*, *AML*), with the main objective to develop a methodology to detect and classify Acute Leukemia blast cells into one of the above types based on image processing and machine learning techniques using peripheral blood smear images.

The methodology presented in this research consisted of several stages namely, image acquisition, image segmentation, feature extraction/selection and, classification.

The data was collected from two different sources, University of Malaya Medical Center (UMMC), Malaysia and M. Tettamanti Research Center for childhood leukemia and hematological diseases, Italy.

The image segmentation addressed several key issues in blast cells segmentation including, the blast cell localization, sub-imaging, color variation and segregation of touching cells. This stage was accomplished using several image processing techniques including, color transformation, mathematical morphology, thresholding, and watershed segmentation. The seeded region growing was used to further segment the blast cell into nucleus and cytoplasm, respectively. This combination resulted in a new algorithm we named ***CBCSA***. Based on the Relative Ultimate Measurement Accuracy for Area, the proposed algorithm was able to achieve an accuracy of 96% and 94% in the extraction of the blast cell region and the nuclear region, respectively.

Various types of features were employed to address the blast cell's morphology, including shape, texture and color. In total, 601 features were extracted from each blast cell, and its nucleus: 31 of these were shape-based features, while 534 were texture-based features and 36 were color-based features.

Artificial Neural Network and Support Vector Machine were used to classify blast cells into either ***ALL*** or ***AML*** according to the extracted features. As a result, an accuracy rate of 96.93% was achieved in the classification of blasts cells.

The resulting system will subsequently act as a second reader after the manual screening of peripheral blood smears. It is believed that this system would increase the diagnostic accuracy and consistency of the hematologist and laboratory practitioner in the daily diagnostic routine.

ABSTRAK

Diagnosis perubatan adalah prosedur mengenal pasti gejala penyakit dengan analisis kritikal dan sering dibantu oleh satu siri ujian makmal dengan tahap kerumitan yang berbeza-beza. Diagnosis perubatan yang tepat adalah penting untuk menyediakan pilihan rawatan yang paling berkesan. Kerja-kerja yang dibentangkan di dalam tesis ini memberi tumpuan kepada pemprosesan imej periferi smear darah pesakit yang menghidap leukemia berdasarkan morfologi sel letupan. Leukemia, iaitu kanser darah, adalah salah satu penyakit berbahaya yang memberi kesan pada orang dewasa dan kanak-kanak. Pemprosesan imej digital dan teknik pembelajaran mesin boleh memainkan peranan penting dalam proses mengenalpasti punca penyakit penyakit ini. Leukemia diklasifikasikan sebagai sama ada akut atau kronik berdasarkan kepantasan penyakit berkenaan berkembang. Leukemia akut boleh diklasifikasikan secara lanjut kepada Leukemia Akut Lymphoblastic (ALL) dan Leukemia Mieloid Akut (AML) berdasarkan pada keturunan sel. Protokol rawatan ini diperuntukkan berdasarkan jenis leukemia. Mujurlah Leukemia juga seperti kanser-kanser yang lain yang mana boleh diubati dan boleh diperbaiki kelangsungan hidup pesakit dan serta rawatannya tertakluk kepada ketepatan diagnosis. Kajian ini secara khususnya memberi tumpuan kepada Leukemia akut, yang boleh menjadi dua jenis yang berbeza (ALL, AML), dengan objektif utama untuk membentuk suatu kaedah untuk mengesan dan membahagikan sel letupan Leukemia Akut ke dalam salah satu jenis di atas berdasarkan pemprosesan imej dan teknik pembelajaran mesin menggunakan imej periferi smear darah. Kaedah yang dikemukakan dalam kajian ini terdiri daripada beberapa peringkat iaitu, imej pengambilalihan, segmentasi imej, ciri pengekstrakan / pemilihan dan klasifikasi. Data diperolehi daripada dua sumber yang berbeza, Pusat Perubatan Universiti Malaya (PPUM), Malaysia dan Pusat Penyelidikan M. Tettamanti untuk Leukemia Kanak-Kanak dan

Hematologi Penyakit, Itali. Laporan segmen imej mengemukakan beberapa isu-isu utama dalam segmentasi sel-sel letupan termasuk, penyetempatan letupan sel, sub-pengimejan, variasi warna dan pembahagian sentuhan sel-sel. Peringkat ini telah dicapai dengan menggunakan teknik pemprosesan imej termasuklah; warna transformasi, morfologi matematik, ambang, dan segmentasi titik. Yang pilihan rantau berkembang digunakan untuk segmen selanjutnya sel letupan ke dalam nukleus dan sitoplasma. Gabungan ini menghasilkan algoritma baru yang dinamakan CBCSA. Berdasarkan Pengukuran relatif utama ketepatan bagi kawasan, algoritma yang dicadangkan telah berjaya mencapai ketepatan 96% dan 94% dalam pengestrakan rantau sel letupan dan rantau nuklear masing-masing. Pelbagai jenis ciri-ciri telah digunakan untuk menangani morfologi sel letupan, termasuk bentuk, tekstur dan warna. Secara keseluruhannya, 601 ciri ini dipetik daripada setiap sel letupan, dan yang nukleus; 31 adalah berdasarkan ciri-bentuk, manakala 534 adalah berdasarkan ciri-tekstur dan 36 adalah berdasarkan ciri-warna. Rangkaian Neural Buatan dan sokongan Mesin Vektor telah digunakan bagi mengelaskan sel letupan ke dalam sama ada ALL atau AML menurut ciri-ciri yang diekstrak. Hasilnya, kadar ketepatan 96,93% telah dicapai dalam klasifikasi sel letupan. Sistem yang terhasil kemudiannya akan bertindak sebagai pembaca kedua selepas pemeriksaan calitan darah periferi secara manual. Adalah diharapkan bahawa sistem ini akan meningkatkan ketepatan diagnostik para pengamal perubatan dan penyelidik di makmal dalam rutin diagnostik harian secara konsisten.

ACKNOWLEDGMENTS

First and foremost I am very grateful to Allah the Almighty for the blessings and guidance He has bestowed upon me throughout the entire period of my doctorate undertaking.

My deepest appreciation goes to my supervisor Associate Professor Datin Dr. Sameem Abdul Kareem for her support and kindness in advising me to keep improving my knowledge and to keep believing in my abilities. A similar level of gratitude is due to Prof. Hany Arrifin from University Malaya Medical Centre, Malaysia, for supplying me with the medical images needed, as well as her guidance and expertise. It is unlikely that I would have reached completion without their encouragement and support.

I express my appreciation to everyone involved directly and indirectly to the success of this research. Last but not least, my family for their understanding, support, patience, and encouragement. Thank you for all the support, comments and guidance.

TABLE OF CONTENTS

Declaration.....	II
Abstract.....	III
Abstrak.....	V
Acknowledgements.....	VII
Table of Contents.....	VIII
List of Figures.....	XIII
List of Tables.....	XVII
List of Abbreviations and Acronym.....	XIX

CHAPTER 1.....	1
INTRODUCTION.....	1
1.1 Preliminary Background.....	1
1.2 Problem Background and Problem Statements.....	6
1.3 Objectives of the Research.....	11
1.4 Research Questions.....	12
1.5 Relationship between Research Objectives and Research Questions.....	13
1.6 Research Contribution	13
1.7 Research Methodology and Proposed Approach.....	16
1.8 Thesis Overview.....	19
 CHAPTER 2.....	 22
LEUKEMIA.....	22
2.1 Introduction.....	22
2.2 Blood and its Components.....	22
2.2.1 White Blood Cells (Leukocytes).....	25
2.3 Types of Leukemia.....	27
2.4 Initial Symptoms of Leukemia.....	29
2.5 Laboratory Diagnosis of Acute Leukemia.....	29
2.5.1 Complete Blood Count.....	33
2.5.2 Peripheral Blood Smear Morphological Examination.....	34
2.5.3 Bone Marrow Aspirate Morphological Examination.....	37

2.5.4	Immunophenotyping.....	39
2.5.5	Cytogenetic.....	39
2.6	Classification of Acute Leukemia.....	40
2.6.1	The French-American-British (FAB) Classification System.....	40
2.6.2	The World Health Organization (WHO) Classification System.....	42
2.7	Leukemia Treatment Options.....	44
2.8	Leukemia Prognosis.....	44
2.9	Summary.....	45
 CHAPTER 3.....		46
BACKGROUND AND LITERATURE REVIEW.....		46
3.1	Introduction.....	46
3.2	Fundamental of Image Processing.....	46
3.2.1	Representation of Microscopy Blood Digital Images.....	46
3.2.2	Color Spaces in Microscope Blood Images.....	48
3.2.2.1	RGB Color Space.....	49
3.2.2.2	HSV Color Space.....	50
3.2.2.3	Lab Color Space.....	51
3.2.3	Image Segmentation.....	52
3.2.3.1	Selected Image Segmentation Techniques.....	54
3.2.3.1.1	Otsu Threshold.....	54
3.2.3.1.2	Seeded Region Growing.....	56
3.2.3.1.3	Mathematical Morphology.....	58
3.2.3.1.4	Watershed Segmentation.....	60
3.3	Feature Extraction and Analysis.....	61
3.3.1	Shape-Based Features.....	63
3.3.2	Texture-Based Features.....	64
3.3.2.1	Histogram-Based Approach.....	66
3.3.2.2	Gray Level Co-occurrence Matrix (GLCM).....	68
3.3.3	Color-Based Features.....	75
3.4	Feature Selection.....	75
3.5	Pattern Classification.....	77
3.5.1	Artificial Neural Network.....	77

3.5.1.1	Multi Layer Perceptron Feed-Forward Network.....	79
3.5.2	Support Vector Machine.....	82
3.6	Review of Computer-Based Acute Leukemia Diagnosis and Classification.....	86
3.6.1	Peripheral Blood Image Acquisition.....	87
3.6.2	Blast Cells Segmentation.....	90
3.6.3	Feature Extraction, Selection and Classification.....	99
3.7	Summary.....	106
 CHAPTER 4.....		107
RESEARCH METHODOLOGY.....		107
4.1	Introduction.....	107
4.2	Data Acquisition.....	110
4.2.1	Dataset-A.....	111
4.2.2	Dataset-B.....	115
4.2.3	Gold Standard.....	116
4.3	Image Segmentation.....	117
4.4	Feature Extraction and Selection.....	119
4.5	Classification.....	122
4.6	Parameters Selection.....	123
4.6.1	MLP-NN Parameters Optimization.....	123
4.6.2	SVM Parameters Optimization.....	125
4.7	Imbalance Data.....	125
4.8	Evaluation Measures.....	127
4.8.1	Blast Cells Segmentation Evaluation.....	127
4.8.2	Classification Performance Measures.....	129
4.9	Summary.....	131
 CHAPTER 5.....		132
PERIPHERAL BLOOD SMEAR IMAGE SEGMENTATION.....		132
5.1	Introduction.....	132
5.2	Blast Cells Localization.....	133
5.2.1	Color Transformation.....	134
5.2.2	Mask and Marker Preparation.....	136

5.2.3	Blast Cells Reconstruction.....	139
5.2.4	Sub-Imaging.....	141
5.3	A Completed Blast Cell Segmentation Algorithm (CBCSA).....	144
5.3.1	Erythrocytes Removal.....	145
5.3.2	Segregating Touching Cells.....	150
5.3.3	Marker Image Preparation.....	153
5.3.4	Nucleus/Cytoplasm Separation.....	154
5.4	Summary.....	158
 CHAPTER 6.....		159
FEATURE EXTRACTION, SELECTION AND BLAST CELL CLASSIFICATION...		159
6.1	Introduction.....	159
6.2	Feature Extraction.....	159
6.2.1	Shape Features.....	160
6.2.2	Texture Features.....	163
6.2.2.1	Histogram-Based Features.....	163
6.2.2.2	GLCM Features.....	165
6.2.3	Color Features.....	170
6.3	Feature Selection.....	170
6.4	Blast Cells Classification.....	171
6.4.1	Data Normalization.....	171
6.4.2	Training and Testing Data Separation.....	171
6.4.3	MLP-NN Optimization, Training and Testing.....	172
6.4.4	SVM Optimization, Training and Testing.....	174
6.4.5	Dataset Balancing.....	177
6.5	Summary.....	177
 CHAPTER 7.....		179
RESULTS AND DISCUSSION.....		179
7.1	Introduction.....	179
7.2	Test and Evaluation Results of the Proposed Blast cells Segmentation Algorithm.....	179
7.2.1	Test and Evaluation of the Blast Cell Localization Algorithm.....	179
7.2.1.1	Discussion of the Results Related to Blast Cell Localization.....	183

7.2.2	Test and Evaluation Results of the CBCSA	184
7.2.2.1	Discussion of Results Related to CBCSA	192
7.3	Comparison with Other Blast Cell Segmentation Methods.....	199
7.4	Results and Discussion Related to Feature Extraction and Selection.....	210
7.5	Experimentation Result of the MLP-NN Architecture Selection.....	214
7.6	Experimentation Result of the SVM Hyper-Parameters Selection.....	218
7.7	Results and Discussion of Acute Leukemia Classification.....	219
7.8	Results and Discussion of Acute Leukemia Classification after Oversampling.....	225
7.9	Comparison between the Proposed Acute Leukemia Classification Approach and Other Approaches in the Literature.....	229
7.10	Summary.....	231
CHAPTER 8.....		232
CONCLUSION AND FUTURE WORK.....		232
8.1	Conclusion.....	232
8.2	Main Contribution.....	234
8.3	Achievement of Research Objectives.....	239
8.4	Impact and Significance to the Medical Field.....	240
8.5	Future Expansion and Recommendation.....	241
8.6	Summary.....	243

LIST OF FIGURES

Figure No.		Page
1.1	The Most Common Childhood Cancer in the United States	4
1.2	The Most Common Types of Cancer in Malaysia in Male	5
1.3	Most Common Types of Cancer in Malaysia in Female	5
1.4	Systematic Diagram of the Proposed Research	17
2.1	Blood Flow System in Human Body	23
2.2	Blood Cell Lineage and Maturation Chart	25
2.3	Common Symptoms of Leukemia	29
2.4	Steps to Confirm Acute Leukemia Diagnosis (Part A)	30
2.4	Steps to Confirm Acute Leukemia Diagnosis (Part B)	31
2.5	Sysmex KX21N Hematology Analyzer	33
2.6	Illustration of PB Smear Preparation and Examination	35
2.7	Bone Marrow Sample	38
2.8	Blood Taken from Bone Marrow	38
3.1	Representation of Microscopic PB digital image	47
3.2	Digital image types (a) Binary image (b) Grayscale image (c) Color image	48
3.3	RGB Cube	50
3.4	HSV Color Space	51
3.5	Lab Color Space	51
3.6	Typical Histogram of a Bi-level Image	55
3.7	The second-order neighborhood $N(x, y)$ of current testing pixel at (x, y)	57
3.8	Simulations of the watershed transform. (a) Input image. (b) Punched holes at minima and initial flooding. (c) A dam is built when waters from different minima are about to merge. (d) Final flooding, with three watershed lines and four catchment basins.	61
3.9	Image Features Description	62
3.10	Classification of shape representation and description techniques.	63
3.11	Samples of textures	65
3.12	Histogram of image with 16 gray-level intensity. (a) Histogram bins (b) Number of pixels in each gray level intensity	67

3.13	Spatial relationships of pixels defined by offsets, where d is the distance from the pixel of interest	69
3.14	Illustration of the GLCM computation process	70
3.15	Neural Network	78
3.16	Basic Structure of Artificial Neuron	79
3.17	MLP-NN with Two Hidden Layers	80
3.18	Optimal Separating Hyperplane	82
3.19	Microscope with a Digital Camera	87
4.1 (a)	The Proposed Acute Leukemia Diagnostic Methodology Phases (Image Acquisition)	107
4.1 (b)	The Proposed Acute Leukemia Diagnostic Methodology Phases (Segmentation, Feature Extraction and Selection)	108
4.1 (c)	The Proposed Acute Leukemia Diagnostic Methodology Phases (Blast Cells Classification)	109
4.2	Equipment Used for <i>Dataset-A</i> Image Acquisition	112
4.3	Olympus UC30 Digital Camera	112
4.4	Olympus CX31 Optical Microscope	112
4.5	Sample Images from <i>Dataset-A</i>	114
4.6	Sample Images from <i>Dataset-B</i>	116
4.7	Sample of Gold Standard Images (a) original image (b) manual highlights of the blast (c) manual highlights of the nucleus	117
5.1	Blast Cell Extraction Flowchart (Part A)	133
5.1	Blast Cell Extraction Flowchart (Part B)	134
5.2	PB image Color Transformation (a) Original RGB Image, (b) Original HSV Image, (c) Saturation Band, (d) Hue Band	135
5.3	Histogram of the Hue channel image in Figure 5.2 (d)	136
5.4	Mask and Marker Preparation (a) Binary version of “ S ” image in Figure 5.2(c), (b) Binary version of “ H ” image in Figure 5.2(d)	137
5.5	The bwH Image after Morphological Opening (Mask)	138
5.6	The bwS Image after Morphological Erosion (Marker)	139
5.7	Illustration of the blast cells reconstruction from Marker and Mask	140
5.8	Reconstructed Blast Cells	140
5.9	The Localized Blast Cells with the Original RGB Pixels	141
5.10	Label Matrix where each blast cell are labeled with different number	142

5.11	Illustration of sub-imaging procedure	143
5.12	Stages of the Completed Blast Cells Segmentation Algorithm (CBCSA)	145
5.13	Flowchart of Erythrocytes Removal Process	146
5.14	Color Contrast Enhancement (a) Original image, (b) Enhanced image	147
5.15	Extracted b channel image before and after applying median filter. (a) b channel image, (b) Smoothed b channel image with median filter	148
5.16	Histogram of the smoothed b channel image	149
5.17	Binary image of highlighted Erythrocytes	149
5.18	Final steps to prepare the Mask image. (a) Original image after subtracting Erythrocytes, (b) Hue channel of (a), (c) Final Mask image	150
5.19	Distance map of the Mask image	151
5.20	3-D representation of the distance map	152
5.21	Segregated touching blast cells	152
5.22	The Localized Blast Cells	153
5.23	Single Blast Cell Sub-Image	154
5.24	Nucleus/Cytoplasm Separation Steps	155
5.25	Production of homogenous nucleus. (a) Saturation channel, (b) Saturation channel after histogram equalization, (c) Resulted image after arithmetic addition of (a) and (b).	156
5.26	Generating the seeded region. (a) Binary version of image in Figure 5.25(c), (b) Seeded region	157
5.27	The grown nucleus region	157
6.1	The proposed methodology stages with emphasis on feature extraction, selection and classification	159
6.2	Graphical representation of simple shape features. (a) Original blast cell, (b) Area, (c) Rectangular bounding box, (d) Convex Hull (e) circularity, (f) perimeter, (g) minimum bounding ellipse	161
6.3	Histogram-based features (ALL). (a) Original ALL sample, (b) Grayscale image of the ALL sample, (c) Histogram of the grayscale ALL Sample	163 164
6.4	Histogram-based features (AML). (a) Original AML sample, (b) Grayscale image of the AML sample, (c) Histogram of the grayscale AML Sample	164
6.5	The effect of gray level quantization on nucleus chromatic pattern	168
6.6	MLP-NN model optimization, Training and Testing process	173
6.7	SVM model optimization, Training and Testing process	176

7.1	Overview of Test Results and Evaluation of the Proposed CAD-AL	179
7.2	Localization of blast cells (example 1)	180
7.3	Localization of blast cells (example 2)	180
7.4	Localization of blast cells (example 3)	181
7.5	Experimental results of the proposed segmentation approach for ALL PB sample (a) original image (b) localized blast cells (c) Ground-truth of blast region (d) Ground-truth of nucleus region (e) Blast region obtained using the proposed segmentation approach (f) Nucleus region obtained using the proposed segmentation approach	186
7.6	Experimental results of the proposed segmentation approach for AML PB sample (a) original image (b) localized blast cells (c) Ground-truth of blast region (d) Ground-truth of nucleus region (e) Blast region obtained using the proposed segmentation approach (f) Nucleus region obtained using the proposed segmentation approach	187
7.7	Experimental results of the proposed segmentation approach for ALL PB sample from Dataset-B (a) original image (b) localized blast cells (c) Ground-truth of blast region (d) Ground-truth of nucleus region (e) Blast region obtained using the proposed segmentation approach (f) Nucleus region obtained using the proposed segmentation approach	188
7.8	Blast region segmentation difficulties in AML . (a) Erythrocytes color is analogues to M3 cytoplasm color, (b) M3 blast with vitreous cytoplasm, (c) M7 with protrusion cytoplasm.	196
7.9	Segmentation result of Image005 (a) Original image, (b) Segmented image with blast cells border overlaid by red line	203
7.10	Segmentation result of Image019 (a) Original image, (b) Segmented image with blast cells border overlaid by red line	203
7.11	Various scenarios of touching (overlapping) cells. (a) Chain of cells, (b) Cluster, (c) Ring, (d) Cluster with filled holes, (e) Ring with filled holes	205
7.12	Sample result of marker-controlled watershed segmentation. (a) nucleus used as a marker, (b) watershed segmentation boundaries superimposed on the original image	206
7.13	Comparison of nucleus segmentation results	207
7.14	SFS Performance	211
7.15	Validation Accuracy versus hidden nodes for three different learning rates	215
7.16	Validation Accuracy versus number of training cycles (epochs) for three different learning rates and four hidden nodes	216
7.17	Graphical representation of the selected MLP-NN Architecture	217
7.18	Validation and testing accuracy using SVM with various combination of C and γ	219

LIST OF TABLES

Table No.		Page
1.1	Cancer Incidence per 100,000 population (CR) and Age-Standardize incidence (ASR), by gender, Peninsular Malaysia 2003-2005	4
1.2	Leukemia Cancer Incidence per 100,000 population (CR) and Age-standardized incidence (ASR), by ethnicity and gender, Peninsular Malaysia 2003-2005	6
1.3	The Relationships between Objectives and Research Questions	13
2.1	The Four Major Components of Blood	24
2.2	White Blood Cells (Basophil, Eosinophil, Neutrophil, Monocyte, Lymphocytes)	26
2.3	The Four Main Types of Leukemia	28
2.4	Description of each step in the acute leukemia diagnosis process	32
2.5	Morphological features of <i>ALL</i> subtypes based on FAB classification system	41
2.6	Morphological features of <i>AML</i> subtypes based on FAB classification system	41
2.7	WHO classification system of <i>ALL</i>	43
2.8	WHO Classification System of <i>AML</i>	43
3.1	Texture features extracted from gray level histogram	68
3.2	GLCM Texture Features	72
3.3	Summary of the acquisition process characteristics reported in the literature	89
3.4	Review of Previous Segmentation Algorithms	97
3.5	Review of previous feature extraction, selection and classification methods with their reported results	104
4.1	Acquisition Characteristics of <i>Dataset-A</i>	111
4.2	Number of Images and Blast Cells (<i>Dataset-A</i>)	113
4.3	Acquisition Characteristics of <i>Dataset-B</i>	115
4.4	Confusion Matrix	129
6.1	Histogram-based features extracted from blast cell nucleus	165
6.2	GLCM texture features from (Haralick, 1973)	166
6.3	GLCM texture features from Soh & Tsatsoulis (1999)	166
6.4	GLCM texture features from Clausi (2002)	167
6.5	GLCM texture features from MATLAB Image Processing Toolbox	167

6.6	GLCM texture features calculated for each nucleus sub-image	169
6.7	Ratio of samples used for training and testing	172
7.1	Evaluation of the proposed BCL Algorithm	182
7.2	The difference between GT and CBCSA segmentation result for blast cells in Figure 7.5.	190
7.3	The difference between GT and CBCSA segmentation result for blast cells in Figure 7.6.	190
7.4	The difference between GT and CBCSA segmentation result for blast cells in Figure 7.7.	190
7.5	Mean \pm standard deviation for the difference (%) between GT and CBCSA segmentation results for all sub-images extracted from Dataset-A and Dataset-B	191
7.6	Performance comparison between the proposed CBCSA and the benchmark	202
7.7	Evaluation of Touching Blast Cells Segmentation Results	205
7.8	Morphological features extracted from blast cell region and nucleus region of ALL and AML	211
7.9	MLP-NN final Parameters setting	217
7.10	Classification performance using the MLP-NN as the learning machine	220
7.11	Classification performance using the SVM as the learning machine	221
7.12	Classification performance using the MLP-NN at three different oversampling rates	226
7.13	Classification performance using the SVM at three different oversampling rates	226
7.14	Performance comparison between the proposed method and other state-of-the-art methods	229

ABBREVIATIONS AND ACRONYMS

PB	Peripheral Blood
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
CLL	Chronic Lymphocytic Leukemia
BM	Bone Marrow
RBC	Red Blood Cell
WBC	White Blood Cell
FAB	French-American-British
CAD-AL	Computer-Aided Diagnosis System for Acute Leukemia
CAD	Computer Aided Diagnosis
ANN	Artificial Neural Network
SVM	Support Vector Machine
ROI	Region of Interest
UMMC	University of Malaya Medical Center
CBC	Complete Blood Count
MGG	May-Grünwald–Giemsa
WHO	World Health Organization
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
SRG	Seeded Region Growing
SE	Structuring Element
SFS	Sequential forward selection
SMOTE	Synthetic Minority Oversampling Technique
RUMA	Relative Ultimate Measurement Accuracy
ME	Misclassification Error
G-mean	Geometric Mean
GLCM	Gray Level Co-occurrence Matrix
Axis Length	AL
CCD	Charge-Coupled Device
ML	Machine Learning

CHAPTER 1

INTRODUCTION

1.1 Preliminary Background

In recent years, image recognition applications have become extremely widespread. They have become tremendously important in several life sectors such as medicine, engineering and science. Vision is the most advanced sense of man's life. However, computerized systems, through the concept of image processing, and machine learning (ML), provides the ability to acquire information about the problem under study in a way that is tough for a human being to obtain. In other words, this information could sometimes be undistinguishable by human vision (Fabijańska & Sankowski, 2009).

The contribution of image processing and machine learning techniques to the field of medicine has been done through the digitized medical images where many phenomena can be analyzed and studied with the aid of the computer. Exponential progress in research and development in the field of image analysis has contributed significantly to the field of medicine. Medical images are considered as a vital tool utilized for the diagnosis and analysis of many diseases, for instance, breast, chest, abdominal illnesses, blood disorder etc. The digital format of the medical images provide an opportunity for further analysis that may lead to a more accurate diagnosis and hence, an optimized patient management. Such images can also be used for research and teaching purposes. The digital medical images that are used in this work are microscopic Peripheral Blood (PB) smear images (Please Refer to Section 2.5.2). The analysis of blood components and its changes is one of the regular diagnostic tests in clinical routine practice.

This research is an attempt to apply digital image processing and ML techniques in the area of medical image analysis and recognition, in particular, Hematology.

The focus of this work is on developing a methodology to diagnose and classify acute leukemia, based on cell morphology, into either Acute Lymphoblastic Leukemia (*ALL*) or Acute Myeloid Leukemia (*AML*).

The PB has been chosen for this research over the Bone Marrow (BM) sample for a number of reasons including:

- 1) The initial leukemia diagnostic process is performed based on the microscope morphological examination of PB slides. Further laboratory tests will be done based on the outcome of the initial diagnosis.
- 2) The PB is usually used for a periodic treatment evaluation, since it is much easier, more economical and less painful to get blood from the vein rather than from the BM.

Leukemia is a blood cancer which affects the White Blood Cells (WBCs); it is one of the most dangerous diseases causing fatality among people, particularly in developed countries (Kothari, R. et al., 1996). Blood is a suspension of millions of cells in a clear liquid. There are three basic types of blood cells namely Red Blood Cells (RBCs/Erythrocytes) (which are responsible for transporting oxygen), White Blood Cells (WBCs/Leukocytes) (responsible for fighting infections) and platelets (specialized cells responsible for blood clotting). They are all made in the factory of blood known as the Bone Marrow (some types of WBCs are also made in the lymph glands) and once they are mature, they are released into the blood stream. In the case of leukemia, WBCs become cancerous for reasons that are still not well understood (Lavelle, 2004).

Leukemia arises in one of the types of WBCs. They may arise in lymphoblasts, which are lymphoid cells in the early stage of development, resulting in a rapid-onset of illness termed Acute Lymphoblastic Leukemia (*ALL*).

Alternatively, when the neoplasm (abnormal rapid reproduction of a cell) (Ciesla, 2007) involves mature cells, it is termed Chronic Lymphocytic Leukemia (**CLL**) and is usually more indolent. In the World Health Organization (WHO) classification, **CLL** is part of *Non-Hodgkin Lymphoma* (NHL) (Swerdlow et al., 2008). Leukemia may also be granulocytic in origin, occurring in either young myeloblastic cells resulting in Acute Myeloid Leukemia (**AML**), or in the mature granulocytes resulting in Chronic Myeloid Leukemia (**CML**). In chapter two, different medical/cellular terms and a full description of blood components will be discussed, in addition, a full description of leukemia characteristics, diagnosis methods, and treatment will be covered.

According to statistics by the American Cancer Society (ACS), leukemia is considered as one of the most common types of cancer, especially in children (American Cancer Society, 2013). New Leukemia cases are diagnosed in about 29,000 adults and 2000 children each year in the United States.

Leukemia affects people of all ages. Approximately 85% of leukemia in children is of the acute type. Based on a study carried out by ACS, it has been reported that leukemia is the second leading death in children aged 1 to 14 years old, after accidents (American Cancer Society, 2013).

According to the (American childhood cancer organization, 2012), the following graph in Figure 1.1, illustrates the distribution of the more common childhood cancers for children from birth to 14 years old in the United States.

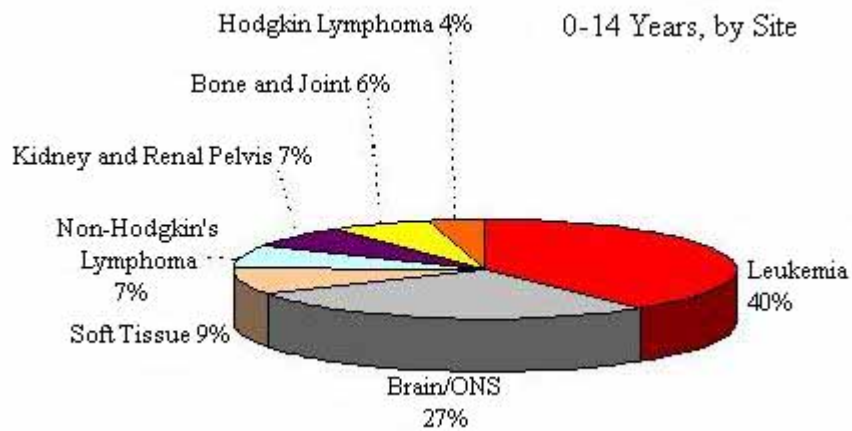


Figure 1.1: Most Common Childhood Cancer in the United States

As this research is conducted in Malaysia, some local statistics about cancer in general and about leukemia in particular is introduced.

Nearly 70,000 new cancer cases were diagnosed among Malaysians in Peninsular Malaysia between 2003 and 2005, according to a report released in early 2008 on the incidence of the disease in West Malaysia. The Cancer Incidence in Peninsular Malaysia 2003-2005 report, published by (Lim et al., 2008), stated that a total of 67,792 new cases were diagnosed among 29,596 males (43.7 per cent) and 38,196 females (56.3 per cent). The annual crude rate for males was 100.2 per cent per 100,000 population, and 132.1 per cent per 100,000 for females. Table 1.1 shows the cancer incidence per 100,000 by gender in peninsular Malaysia (Lim et al., 2008).

Table 1.1 Cancer Incidence per 100,000 population (CR) and Age-Standardize incidence (ASR), by gender, Peninsular Malaysia 2003-2005

Gender	No.	%	CR	ASR
Male	29596	43.7	100.2	136.9
Female	38196	56.3	132.1	156.4
Both Genders	67792	100	116.0	145.6

The National Cancer Registry report published in 2008 (Lim et al., 2008), categorized the most common types of cancer in Malaysia according to the gender. In male, the commonest cancers are (from most frequent to least frequent): large bowel, lung,

nasopharyngeal cancer, prostate gland, *leukemia*, lymphoma, stomach, liver, bladder and other skin cancers. (Please Refer to Figure 1.2)

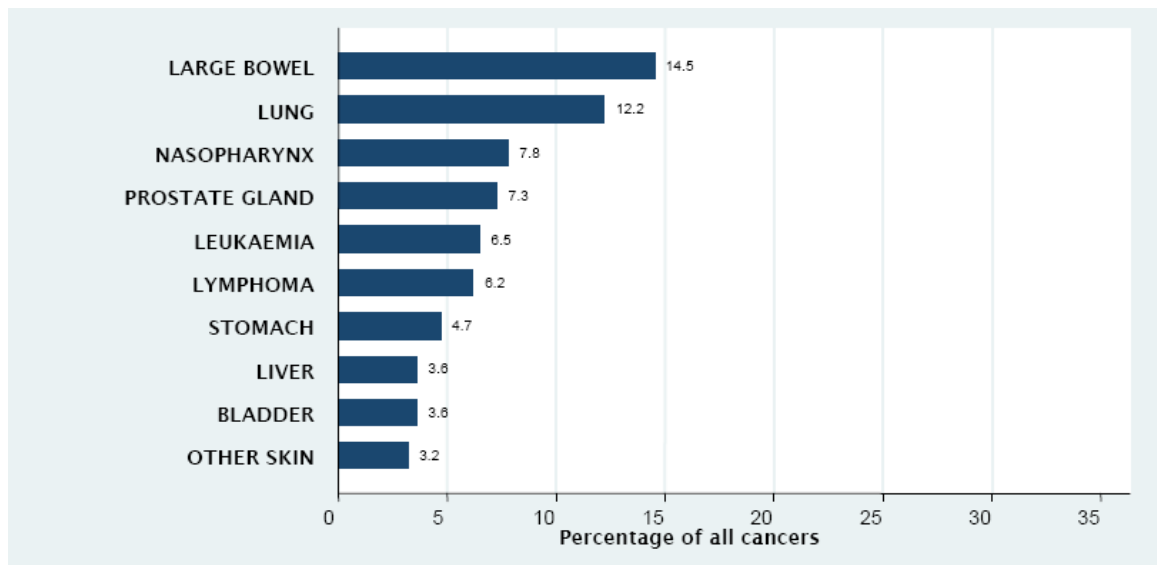


Figure 1.2: The Most Common Types of Cancer in Malaysia in Male

In Female, the commonest cancers are (from most frequent to least frequent): breast, cervix, large bowel, ovary, *leukemia*, lung, lymphoma, corpus uteri, thyroid gland and stomach. (Please Refer to Figure 1.3)

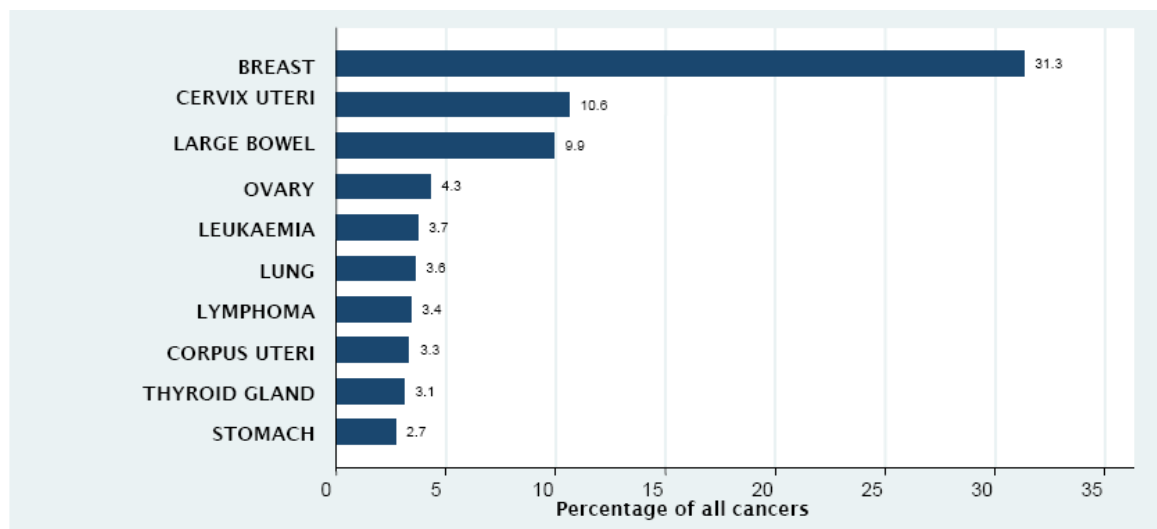


Figure 1.3: The Most Common Types of Cancer in Malaysia in Women

Surprisingly, it was found that leukemia was high in the rank among Malay male cancers, though the fact was consistent with the Kelantan Cancer Registry Report 1999-2003, which

found that leukemia is the third most frequent type of cancer among all males, and second highest among Malay males. In contrast, in the Penang Cancer Registry of the same period, leukemia was found to be the 8th most common cancer among males and females. Table 1.2 shows the leukemia cancer incidence per 100,000 by ethnicity and gender in peninsular Malaysia (Lim et al., 2008).

Table 1.2: Leukemia Cancer Incidence per 100,000 population (CR) and Age-standardized incidence (ASR), by ethnicity and gender, Peninsular Malaysia 2003-2005

Ethnic group	Male		Female					
	No.	%	CR	ASR	No.	%	CR	ASR
Malay	220	67.9	3.6	4	111	55.5	1.8	2
Chinese	86	26.5	3.2	3.3	72	36	2.8	2.6
Indian	18	5.6	2	2.2	17	8.5	1.9	1.9

In the last 40 years, survival rates in leukemia have substantially increased because of the improvements in diagnosis and treatment. In the year 1960, the overall 5-year survival rate for all leukemia was about 14%. However, it is now increased to 70% (Johnson, 2010). Thus, diagnosing the correct type of leukemia is vitally important, since this identifies the treatment options to be given (National Cancer Registry, 2008).

1.2 Problem Background and Problem Statements

Leukemia is the cancer of the BM and the WBCs. Although leukemia is considered as a dangerous type of cancer, the recent advances and development in the diagnostic tools and treatment options have resulted in a cure rate of almost 70% (Priya Johnson, 2010). Generally speaking, there are two types of leukemia; namely acute leukemia and chronic leukemia. Acute leukemia is clinically and biologically different from chronic leukemia. Acute leukemia is characterized by its rapid and aggressive proliferation of immature cells, namely, the blast cells. On the other hand, chronic leukemia progresses slowly over the course of many years.

Chronic leukemia is sometimes monitored over a period of time before treatment is considered in order to ensure maximum effectiveness of the therapy. On the other hand, acute leukemia must be treated immediately (Boundless 2013) otherwise, if left untreated; it can result in death in a matter of a few weeks (Silverstein et al., 2006).

Acute leukemia is a group of heterogeneous diseases that affects all ages (Döhner et al., 2010, Gökbuget & Hoelzer, 2009). The most widely used protocols for acute leukemia classification are the French-American-British (FAB) and the World Health Organization (WHO) classification (Tkachuk et al., 2007).

Basically, both classification protocols categorize acute leukemia as Acute Lymphoblastic Leukemia (*ALL*) and Acute Myeloid Leukemia (*AML*), based on the precursor of the blast cell (Please Refer to Section 2.3). Acute leukemia is very aggressive and requires immediate treatment to be given. Moreover, the treatment of *ALL* is different from that of *AML*. Therefore, it is critically important to determine whether the cell of origin is lymphoid or myeloid as quickly as possible, in order to administer the correct therapy early (Riley & Ben-Ezra, 1999). For this reason, we consider acute leukemia as the current focus of this research.

Clinically, various laboratory tests are used in the diagnosis and differentiation of acute leukemia such as the microscopic morphological examination of PB slides and BM aspiration. The BM is also subjected to immunophenotyping and cytogenetic analysis (Please Refer to Section 2.5.4 and 2.5.5). Microscopic morphological examination of the PB smear is often the first step in the diagnostic process, despite the existence of other advanced diagnostic procedures such as flow cytometry, immunophenotyping, and cytogenetic analysis. This is because PB smear examination is considered as the most economical procedure for initial screening of acute leukemia (Angulo et al., 2006) and it is

usually carried out before exposing the patient to any painful or invasive procedures such as BM biopsy. Another benefit of the PB smear morphological examination in the diagnosis of leukemia is to suggest a likely diagnosis or range of diagnoses, to indicate which more appropriate additional tests are required, and therefore, avoiding sophisticated and unnecessary tests that are difficult to interpret such as immunophenotyping. Hence, a PB smear screening is of particular importance because it facilitates a rapid diagnosis and specific treatment (Bain, 2005). However, the downside of this procedure includes labor-intensive laboratory routines. In addition, it is subject to human error, inter-observer variation (the diagnosis disagreement among different observers) and requires highly trained experts (Scotti, 2005; Le et al., 2008; Briggs et al., 2009; Mohapatra et al., 2013).

Despite the recent momentous improvements in Hematology instruments such as hematology analyzers, these devices can only identify the various types of normal leukocytes circulating into the blood stream without being able to classify abnormal cells, (Bain, 2005; Briggs et al., 2009).

More recently an automated microscope known as CellaVision DM96 (CellaVision AB, Lund, Sweden) was introduced. This instrument scans stained blood slides, identifies potential WBCs and then takes digital images at high magnification. The WBC images are then classified by an artificial neural network based on a database of cells. The user either validates the cell classification if the DM96 has correctly identified the WBCs or manually reclassifies the WBCs in the correct category in case the DM96 misclassify them.

A number of recent studies which investigated the use of DM96 showed that the DM96 was able to detect blast cells. However, according to study carried out by (Billard et al., 2010), the DM96 was only able to classify 74% of the *ALL* and 73% of the *AML*, reflecting a high proportion of cells misclassified by the DM96.

There was still an overestimation of lymphocytes and an underestimation of blast cells. It was recommended from these studies that, laboratory staff should rely upon conventional microscopy in the initial leukemia diagnostic process. Hence, classification of immature and abnormal cells, such as blast cells and atypical lymphocytes, using such analyzers is still unreliable (Billard et al., 2010; Briggs et al., 2009; Cornet et al., 2008).

Computer-aided microscopic morphological examination using image processing and machine learning techniques substantially reduces the time as compared to the manual procedure as it allows scanning larger number of PB slides (Escalante et al., 2012); it also increases the accuracy of the result by eliminating human error, such as error resulting from repetition, fatigue, lack of experience, etc. The computer-aided PB screening for the purpose of acute leukemia diagnosis and classification consists of the following stages after image acquisition: blast cells localization and segmentation, feature extraction/selection, and finally, blast cells classification. This research deals with acute leukemia diagnosis and classification. Thus, all the stages mentioned earlier are included.

From the technical point of view, isolating the cells of interest (blast cells) from the stained blood image background is a key issue in building a computer aided system for hematological malignancy classification. The PB segmentation is crucially important since the accuracy of the succeeding steps, namely, feature extraction and classification are totally dependent on the accurate segmentation of the cells of interest (Liao, Q. & Deng, Y. 2002, Joshi, M. & Karode, A. 2013). Thus, the segmentation stage is considered as the most challenging and difficult problem due to the following reasons:

1. The complex nature of the cells presented in the PB slides (Liao & Deng, 2002). This complexity comes from the diversity in cell shape, size and appearance.

2. Individual cell localization and extraction into a sub-image. Sub-images containing single nucleus per image are essential for feature extraction (Mohapatra, 2011). Accurate cell localization and extraction is affected by the indistinct boundaries between the cell of interest and the background in many cases (Nee el at., 2012).
3. It is almost impossible to obtain the same imaging quality during the acquisition stage (Markiewicz et al., 2005), as this is dependent on the different levels of illumination, lights, staining procedure, and the proficiency of the laboratory staff who prepare the PB smear.
4. Adjacency and superimposition of cells. It is usually challenging to obtain satisfactory segmentation results, especially during the separation of touching or overlapping cells (He & Liao, 2008).

Assuming that all the blast cells are segmented properly, it is a very important to extract proper diagnostic features (Duda., et al, 2012) that describe the blasts through a numerical value. Blast cells are classified as either lymphoid or myeloid based on these features. There are various methods that can be used to generate features for acute leukemia classification. Usually, the features come under three groups, namely, shape, texture, and color (Sinha & Ramakrishnan, 2003). Hundreds of features can be extracted from these three groups. However, not all of them are useful for the classification process. Different blood cells could have similar feature values, for instance, two different cells could have the same area size and thus giving no contribution to the classification process. Thus, the key point is to determine the optimal set of discriminative features, which may lead to the most efficient recognition results (Osowski et al., 2009).

Based on the intensive literature review conducted (Please Refer to Chapter 3), it has been found that there is only small number of scientific work that focused on the problem of acute leukemia diagnosis and classification. Although a number of researchers have attempted to look into this problem, such as (Scotti, 2006; Scotti, 2005; Markiewicz et al., 2005; Supardi et al., 2012; Nasir et al., 2013), there is still a great need for more efforts and research in this field. Since any image analysis system consists of three main stages, namely, segmentation, feature extraction and classification, some researches such as (Sadeghian et al., 2009; Patil et al., 2012; Nee et al., 2012; Madhloom et al., 2012) focused on only one stage namely segmentation. A number of other researchers, including the studies done by (Piuri & Scotti, 2004; Theera-Umpon & Dhompongsa, 2007; Rezatofighi & Soltanian-Zadeh, 2011) focused on differential blood counting of WBCs but not leukemia, while others focused only on **ALL** such as the work by (Scotti, 2005). Chapter 3 will discuss the strengths and weaknesses of the most recent researches that have been conducted in this area.

1.3 Objectives of the Research

This research focuses on developing a diagnostic methodology for acute leukemia blast cells using image processing and ML techniques on PB smear images. In this thesis, we first discuss the relevant image processing and ML techniques in order to identify the most suitable approach for the acute leukemia diagnostic process. The aim of this research is to utilize image processing and ML techniques in order to increase the accuracy of diagnosing acute leukemia for the optimal classification of **ALL** and **AML**. The following objectives have been formulated in order to attain the aim of this research.

1. To apply an image processing algorithm for localization and segmentation of acute leukemia blast cells.

2. To apply ML techniques to select the optimum set of features extracted from blast cell images in order to correctly classify acute leukemia into either *ALL* or *AML*.
3. To evaluate the performance of the proposed approach using real-world PB smears images.

1.4 Research Questions

In order to set the direction of this research, the following research questions have been drawn up:

- a) What are the key points we should include in the proposed segmentation algorithm to solve the issues presented in the existing algorithms?
- b) Can the proposed segmentation algorithm extract the blast cells accurately?
- c) How can unique or discriminative features be extracted from the blast cells?
- d) What are the techniques needed to be integrated in the proposed approach so that it can classify acute leukemia blast cells more accurately?
- e) How can the problems associated with the current methods of diagnosis be solved by the proposed algorithm?
- f) What are the benefits of using a computer-aided diagnosis system over the current available methods?
- g) What evaluation metrics should be performed to confirm the proposed approach can segment and classify acute leukemia blast cells with good accuracy?

1.5 Relationship between Research Objectives and Research Questions

Research questions are sketched to provide the direction of the research. Table 1.3 illustrates the correlation between research objectives and research questions.

Table 1.3: The Relationships between Research Objectives and Research Questions

Objectives	Research Questions
1. To apply an image processing algorithm for localization and segmentation of acute leukemia blast cells.	a) What are the key points we should include in the proposed algorithm to solve the issues presented in the existing algorithms? b) Can the proposed segmentation algorithm extract the blast cells accurately?
2. To apply ML techniques to select the optimum set of features extracted from blast cell images in order to correctly classify acute leukemia into either <i>ALL</i> or <i>AML</i> .	c) How can unique or discriminative features be extracted from the blast cells? d) What are the techniques needed to be integrated in the proposed approach so it can classify acute leukemia blast cells more accurately? e) How can the problems associated with current methods of diagnosis be solved by the proposed algorithm?
3. To evaluate the performance of the proposed approach using a real-world PB smears images.	f) What are the benefits of using a computer-aided diagnosis system over the current available methods? g) What evaluation metrics should be performed to confirm the proposed approach can segment and classify acute leukemia blast cells with good accuracy?

1.6 Research Contribution

In a real-life scenario, a hematologist or laboratory practitioner uses the microscopic morphological examination of PB smear to detect blast cells and determine its type. In many cases, even a skillful operator finds it difficult to manually distinguish the various types of blast cells based on morphology (Please Refer to Table 2.5 and 2.6) (Kawthalkar, 2012). Moreover, the error rate in the manual recognition of blast cells is between 30%-40% depending on the operator's experience (Reta et al., 2010).

As mentioned earlier, the goal of this research is to utilize image processing and ML techniques in order to increase the accuracy of diagnosing acute leukemia for the optimal classification of *ALL* and *AML*.

In order to achieve the intended goal, the research is carried out in four main stages, namely, 1) image acquisition, 2) image segmentation, 3) feature extraction and selection, and finally, 4) classification. These form the four main modules of a typical architecture of a Computer-Aided Diagnosis (CAD) system.

This research extends the work of earlier researchers and makes several key contributions as follows:

- Segmentation of blast cells from other blood components such as RBCs, platelets and plasma as well as segmenting single blast cell into nucleus and cytoplasm.
- An extensive color-channel analysis to determine the most suitable color space and color channels that can lead to the best segmentation quality. For this purpose, two different datasets of PB images are included.
- Objective evaluation of the blast cell segmentation method in PB images against a ground truth of manually segmented PB image. The proposed segmentation algorithm achieves remarkable results of approximately 96% in blast cell extraction and 94% in nucleus/cytoplasm separation.
- Comparative study with two state-of-the-art blast cell segmentation methods, which shows the superiority of the proposed method.
- Guided generation of three different types of features based on shape, texture and color information extracted from the blast cell and its nucleus.
- The proposed approach achieves remarkable results of 96% accuracy in classifying acute leukemia blast cells using two classification engines, namely, the Artificial Neural Network (ANN) and the Support Vector Machine (SVM). The results are remarkably comparable with and outperform the majority of the state-of-the-art methods presented in the literature.

Furthermore, the proposed research outcomes add considerable improvements in the daily routine of the medical laboratory in terms of productivity and quality assurance. It also allows the hematologist or laboratory practitioner to allocate the blast cells automatically where each cell can be reviewed individually on the screen. This feature will reduce the time spent searching for the cells of interest in the whole PB smear. It tremendously reduces the burden of manual screening of PB slides. Moreover, images can be saved for future assessment and comparison. An added advantage is that this system can contribute to the education and training of new laboratory practitioners and act as an efficient learning tool.

Apart from facilitating the laboratory daily routine, the proposed system provides the specialist with substantial assistance when detecting and classifying blast cells. As the initial symptoms of acute leukemia are vague and could resemble other benign diseases such as a viral infection, most patients initially seek medical attention through their general practitioner. The proposed system can be used to alert primary healthcare physicians and general practitioners, who may unwittingly see patients with acute leukemia at the initial presentation. In Malaysia, a country of 13 states and 30 million people, there exist only four tertiary-referral centers for childhood cancer with less than 30 trained pediatric hematologists. Hence having a tool to facilitate the initial screening of children suspected of having acute leukemia would be beneficial to clinicians and laboratories located outside of major hospitals.

1.7 Research Methodology and Proposed Approach

The proposed methodology is carried out systematically to solve the research problem and answers the research questions by logically adopting various stages. It also defines the way in which the data are collected for the research. The diagnosis and classification of acute leukemia generally consist of several stages, these include: image acquisition, image segmentation, feature extraction, feature selection and classification. Figure 1.4 is a diagrammatic representation of the proposed research. The first stage of the research is image acquisition, which is an essential step for the diagnosis of acute leukemia. A prerequisite to efficiently diagnose acute leukemia is to set up a standard methodical procedure under which a large collection of good quality, crisp and well contrasted PB images could be captured. In this research, we collaborated with a highly qualified hematologist, from the University of Malaya Medical Center (UMMC), Kuala Lumpur, Malaysia, in order to establish such a standard and consistent image acquisition procedure. In this context, every advice from the hematologist has been taken into consideration, in order to acquire images under standard magnification and lighting conditions so that the captured images function as good-quality input data to the diagnostic system.

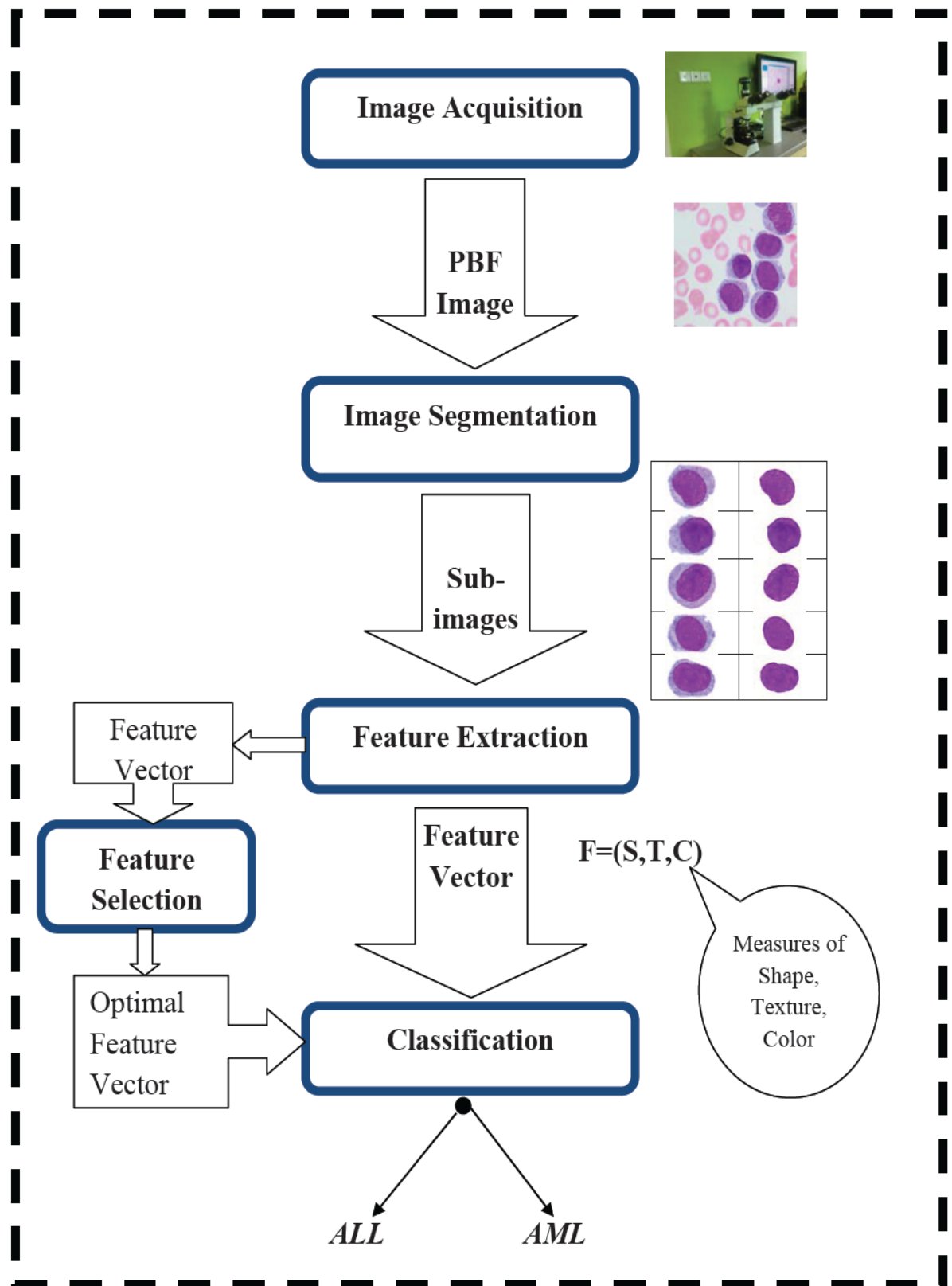


Figure 1.4: Systematic diagram of the proposed research

The second stage of the research is image segmentation. The purpose of the image segmentation stage is to separate the blast cells from the other surrounding blood components such as RBCs, platelets and blood plasma. Furthermore, each blast cell is segmented into the nucleus and the cytoplasm. This stage produces two outputs: (i) a sub-image(s) of the blast cell(s) extracted and placed on a white background, (ii) a nucleus sub-image extracted from the blast cell sub-image. The determined blast cell and its nucleus are the regions of interest (ROI) to be analyzed in the succeeding stages of the research. After the segmentation algorithm proposed in this research have been developed and methodically tested on an adequately large set of PB images, another data-acquisition challenge of this stage, is to obtain a gold standard. The gold standard is the reference manually segmented image needed in order to verify and evaluate the performance of the proposed segmentation algorithm. It was prepared for each image in the dataset by arranging a number of meetings with the hematologist at the UMMC and manually segmenting the blast cells from the acquired PB images using Adobe Photoshop. All the gold standard images were verified by the hematologist from UMMC. Please Refer to Section 4.2.3 for more information about the gold standard images.

The purpose of the next stage namely, feature extraction stage is to extract several features or measurements from the blast cell and its components such as shape, texture and color. These features are later employed as input to the classification engine. Based on the input feature vector, the classifier determines whether the blast cell is *ALL* or *AML*. In many computer vision researches, an intermediary stage is firmly set in place between feature extraction and classification; this is known as feature selection.

The key role of feature selection is to find the optimum subset of features, which gives the highest discrimination power when utilized by the classification engine. (Please Refer to Sections 3.4 and 4.4)

1.8 Thesis Overview

This thesis is logically structured into eight chapters comprising of this introduction chapter and seven further chapters as follows:

Chapter 2 “**Leukemia**” provides background information about healthy blood cells, and leukemia blast cells. It addresses the four main types of leukemia, including *ALL*, *AML*, *CLL* and *CML*. The thesis concentrates on both *ALL* and *AML*.

Moreover, this chapter discusses the current leukemia diagnostic methods used in daily routine. It also demonstrates in detail the two leukemia classification system, namely, FAB and WHO. Towards the end of this chapter, a brief enlightenment on leukemia treatment options and prognosis is provided.

Chapter 3 “**Background and Literature Review**” Throughout this chapter, the key techniques and algorithms that are used in this research to develop the computer-aided diagnostic system are highlighted and explained. This chapter also presents a survey of existing studies on computer-based leukemia diagnostic systems. These studies cover all main components of such systems such as segmentation, feature extraction, feature selection and classification.

Chapter 4 “**Research Methodology**” describes the requirements for designing the proposed acute leukemia diagnosis approach using PB images. First, the design of a proposed approach is introduced. The requirements of image acquisition are then explained.

The requirements of image processing and image segmentation are also discussed, followed by the feature extraction and feature selection processes. The chapter further elaborates on the requirements for classification and recognition of acute leukemia blast cells. Finally, the performance of measurements used to evaluate and test the proposed approach, is elucidated.

Chapter 5 “**Peripheral Blood Smear image Segmentation**” presents two proposed methods for blast cells segmentation in PB images; Blast cells Localization (*BCL*) and Completed Blast Cells Segmentation Algorithm (*CBCSA*), the latter being an enhancement of the former. As a requirement for performing segmentation, both proposed methods apply color-space analysis to determine the most effective and discriminative color channels for detecting blast cells in PB images.

The *BCL* focuses on separating the blast cells from the background components, whereas the *CBCSA* introduces further improvement by addressing various issues presented in PB images segmentation, such as color variation, segregating touching cells and nucleus/cytoplasm separation. Various stages of the development process are covered and the outcome details of each stage are presented.

Chapter 6 “**Feature Extraction, Selection and Classification**” presents the proposed feature extraction method which combines features derived from shape, textural, and color properties of the blast cells. The textural features are derived using first order and second order statistics represented by histogram statistics and Gray Level Co- occurrence Matrix (GLCM) statistics respectively; and the shape features are derived from shape indices, whereas color features are derived from histogram statistics. The chapter then discusses the process of feature selection by applying the sequential feature selection (SFS) method.

Moreover, the application of the two different techniques for classification of acute leukemia blast cells is discussed, namely, the Artificial Neural Network (ANN) and Support Vector Machine (SVM), which are commonly used in blood cells classification related studies.

Chapter 7 “**Results and Discussion**” This chapter presents the discussion and the results of the experiments carried out. The chapter demonstrates how the results of the proposed approach resolve the problems mentioned in the problem statements (Please Refer to Section 1.2).

Chapter 8 “**Conclusion and Future Work**” concludes and summarizes the research contributions made. The achievements and objectives of the research with respect to the experimental results obtained are highlighted along with the key findings and significance of the research. This chapter also discusses the impact and significance of the proposed approach to the hematology community in particular, and to society in general.

CHAPTER 2

LEUKEMIA

2.1 Introduction

Leukemia is a group of heterogeneous blood-related cancers, differing in its aetiology, pathogenesis, prognosis and response to treatment (Bain, 2010). Leukemia is considered as a serious issue in modern society, as it affects both children and adults and even sometimes infants under the age of 12 months. In children, leukemia is considered as the most common type of cancer, while, in adults, the World Health Organization report shows that leukemia is one of the top 15 most common types of cancer (Kampen, 2012). To better understand leukemia, the next sections are dedicated to the discussion of the blood cells lineage, types of leukemia, diagnostic methods currently in use, treatments options as well as prognostic factors.

2.2 Blood and its Components

Blood is a red colored, life-sustaining fluid which circulates through the heart and blood vessels (Veins and Arteries) as shown in Figure 2.1. Blood is vital for life. Blood flows throughout the human body carrying oxygen and nutrients to the tissues and delivers leftover products of metabolism to the lungs, liver and kidneys, where they are then removed from the body (Bain, 2008). Blood comprises of four major elements namely plasma, red blood cells (RBC), white blood cells (WBC) and platelets (Starr et al., 2007; Ciesla, 2007). Table 2.1 demonstrates the four major components of blood.

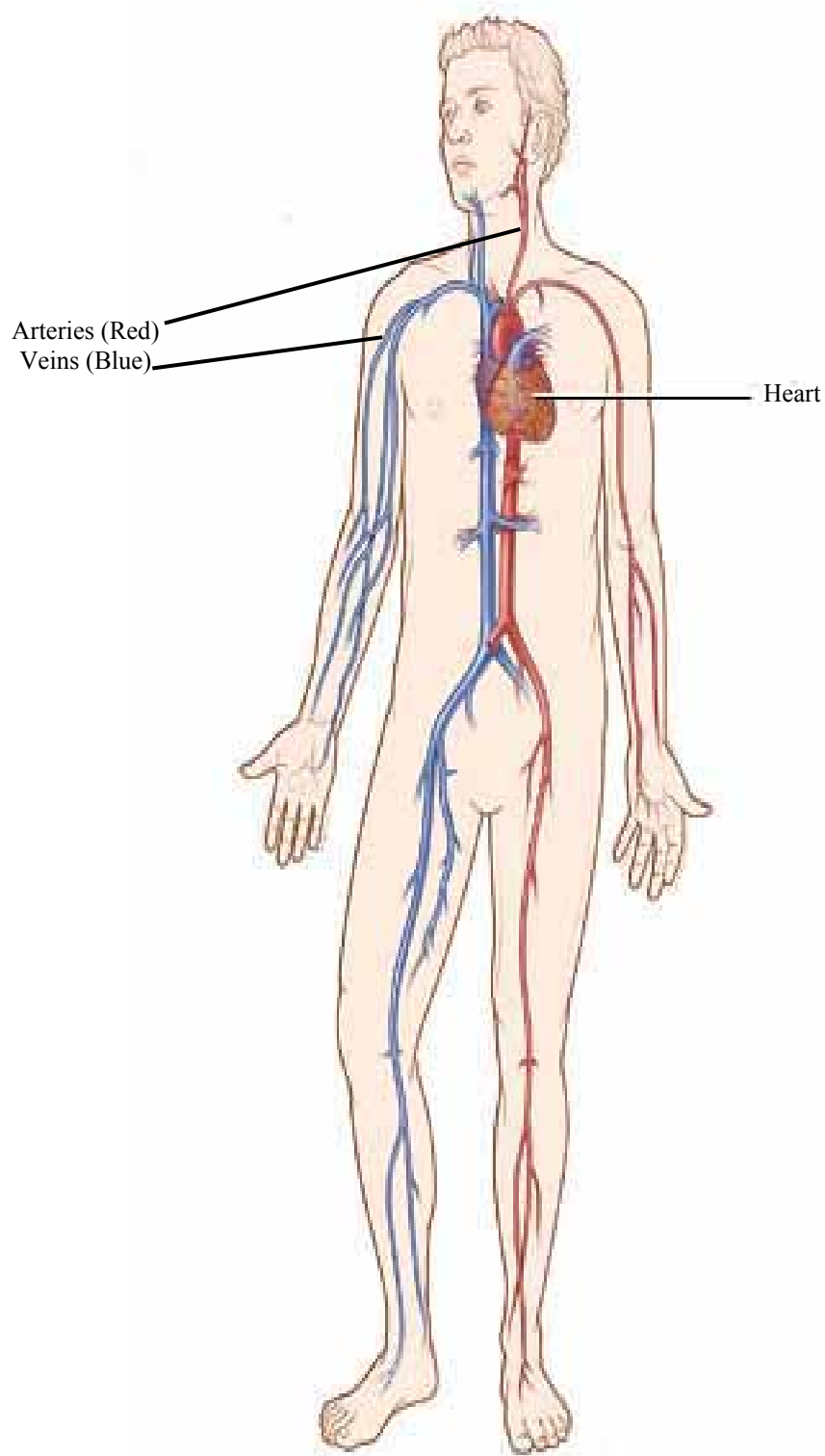
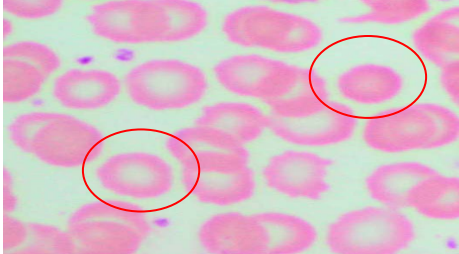
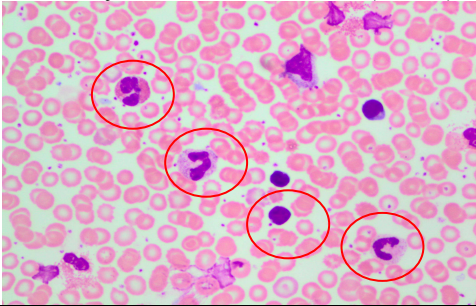
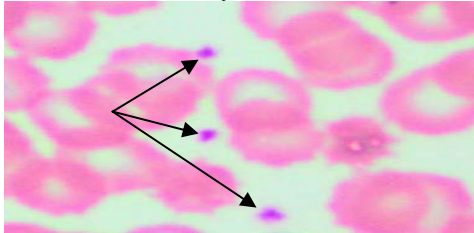
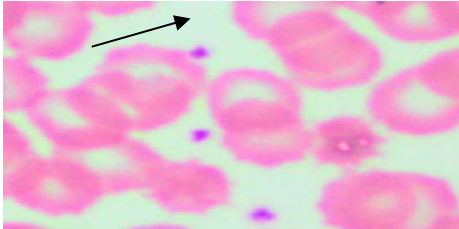


Figure 2.1 Blood Flow System in Human Body (Healthwise Staff, 2014)

Table 2.1: The Four Major Components of Blood

Blood Element	Description
<p>Erythrocytes: Red Blood Cells (RBCs)</p> 	<p>RBCs are responsible for carrying oxygen from lungs to the body tissues and organs and bringing back carbon dioxide to the lung (Paul, 2006).</p>
<p>Leukocytes: White Blood Cells (WBCs)</p> 	<p>WBCs are part of the immune system where they defend the body against both infections and foreign bodies. (Brooks, 2008).</p>
<p>Thrombocytes: Platelets</p> 	<p>Platelets are responsible for aiding in the blood clotting and subsequent wound healing, which occur at a site of injury (Manfred et al., 1999)</p>
<p>Plasma</p> 	<p>Blood plasma carries many important substances such as nutrients, waste, gases, and antibodies. (Aehlert & Vroman, 2011)</p>

All blood cells originate from the BM, growing from the hematopoietic stem cells (lymphoid and myeloid) (Ciesla, 2007). Figure 2.2 shows the maturation path of different blood cells originating from the haematopoietic stem cells including the lymphoid and myeloid stem cells.

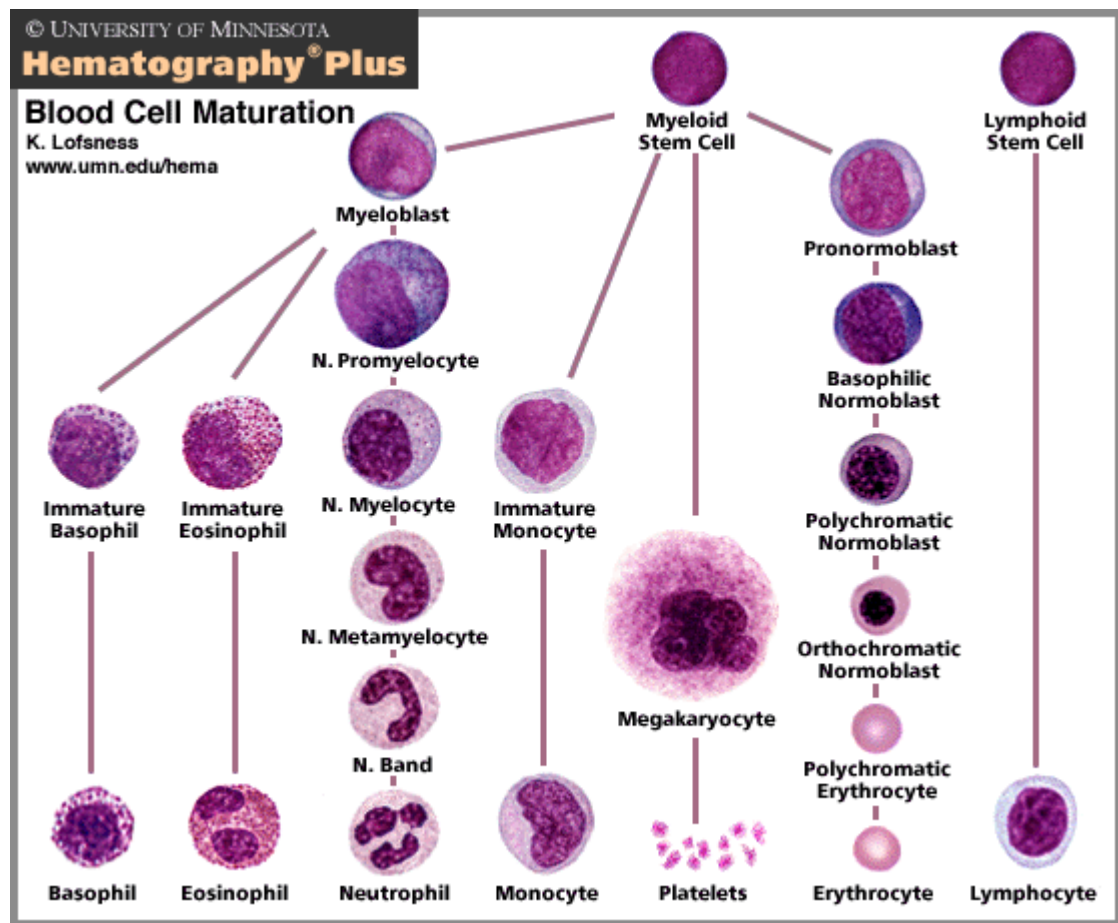


Figure 2.2: Blood Cell lineage and maturation chart (Lofsness. 2008)

2.2.1 White Blood Cells (Leukocytes)

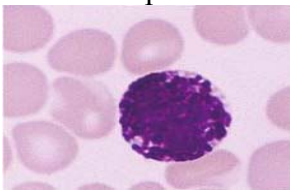
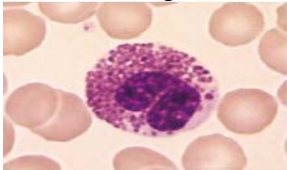
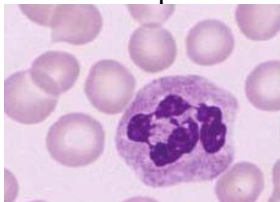

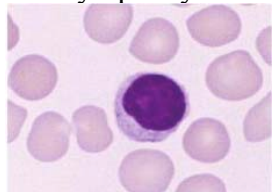
Normally, WBCs are larger in size than RBCs and platelets (Zamani & Safabakhsh, 2006). However, WBCs are the least numerous component of blood cells where each micro liter of blood contains approximately 5000-10000 WBCs (Esteridge et al., 2000) as opposed to 150,000 platelets in the same volume. WBCs are a component of the immune system and provide the first greatest defense against both infections and foreign bodies. (Brooks, 2008).

The human blood comprises of five types of WBCs namely basophil, eosinophil, neutrophil, monocyte, and lymphocytes as shown in Table 2.2. In healthy human blood, each type of WBC has a specific percentage of WBCs as follows: neutrophils 50-

70%, eosinophils 1-4%, basophils 0-1%, monocytes 2-8%, lymphocytes 20.40%.

Calculating the percentage of different type of WBC is known as differential blood count (RI Bijlani & Manjunatha, M. 2010, GK & Pravati, 2006). Section 2.5.1 provides more details about differential blood count.

Table 2.2: White Blood Cells (Basophil, Eosinophil, Neutrophil, Monocyte, Lymphocytes) (Hoffbrand & Moss, 2011, Hoffbrand et al., 2001)

WBCs Type	Description
<p>Basophil</p> 	<p>Basophil cells are only seen in normal peripheral blood. They have many dark cytoplasmic granules, which overlie the nucleus and contain heparin and histamine.</p>
<p>Eosinophil</p> 	<p>Eosinophils are similar to neutrophils in size, nuclear morphology, chromatin pattern and nuclear/cytoplasm ratio. The main difference between them is the presence of uniform, coarse and red granules in the cytoplasm of eosinophils. They provide defense against parasites and help the removal of fibrin formed during inflammation.</p>
<p>Neutrophil</p> 	<p>This cell has a nucleus characteristic consisting of between two and five lobes, and a pale cytoplasm. The granules are divided into primary, which appear at the promyelocyte stage and secondary which appear at the myelocyte stage and predominant in the mature neutrophil. The lifespan of neutrophils in the blood is only about 10h.</p>
<p>Monocyte</p> 	<p>These are usually larger than other PB leucocytes. The main function of monocytes is the defense against bacteria, fungi, viruses, and foreign bodies</p>
<p>Lymphocyte</p> 	<p>These are the immunologically competent cells which assist the phagocytes in the defense of the body against infection and other foreign invasion. Two unique features characteristic of the immune system are the ability to generate antigenic specificity and the phenomenon of immunological memory.</p>

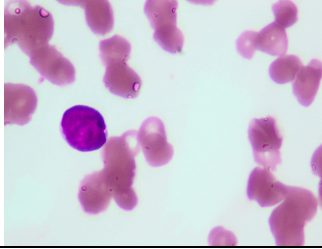
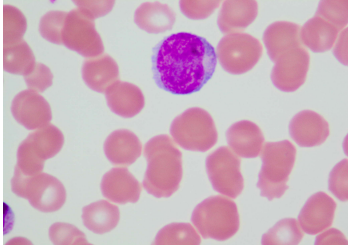
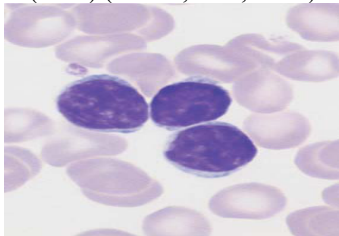
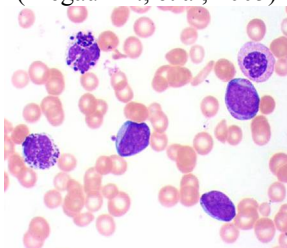
2.3 Types of Leukemia

Figure 2.2 shows the maturation process of the various blood cells and stages of maturation before becoming adult cells and are released into the bloodstream. In the case of leukemia, an interruption in the WBC maturation process occurs where the immature cells (blasts) remain immature, cannot carry out normal function, multiply continuously and eventually invade the BM, replacing all the normal cells.

Leukemia is generally divided into two types known as acute leukemia or chronic leukemia depending on how quickly the immature cells (blasts) proliferate (Bain, 2010). Leukemia can be further grouped based on the type of cell that predominates in the PB and the BM defined according to cell lineage as either myeloid or lymphoid (Please Refer to Figure 2.2) (Ciesla, 2007).

Hence, there are four types of leukemia, namely, acute lymphoblastic leukemia (**ALL**), acute myeloid leukemia (**AML**), chronic lymphocytic leukemia (**CLL**) and chronic myeloid leukemia (**CML**). (Edward, 2002; Bain, 2003; Bain, 2006; Norman, 2009). Table 2.3 exhibits the four types of leukemia.

Table 2.3: The Four Main Types of Leukemia (Hoffbrand et al., 2001)

Progression	Stem Cell	Type	Description
Acute	Lymphoid	Acute Lymphocytic Leukemia (<i>ALL</i>) 	The most common type of leukemia in young children. This disease also affects adults, especially over the age of 65.
	Myeloid	Acute Myeloid Leukemia (<i>AML</i>) 	It develops in both adults and children.
Chronic	Lymphoid	Chronic Lymphocytic Leukemia (<i>CLL</i>) (Theml, et al, 2004) 	It occurs mainly in adults and almost never seen in children
	Myeloid	Chronic Myeloid Leukemia (<i>CML</i>) (Mcgaufflin., et al, 2005) 	Most commonly it affects adults over the age of 55. It sometimes occurs in younger adults, but it is rare in children.

2.4 Initial Symptoms of Leukemia

Leukemia may present in various ways. Many times, the symptoms are non-specific and these include fever, weight loss and loss of appetite. Occasionally, there are symptoms which are related to BM insufficiency e.g. pallor and bruising.

Figure 2.3 shows the common symptoms of acute and chronic leukemia that could affect the, muscles, skin, lungs, bones and joints and spleen (Pauline, 2013). These signs and symptoms are not specific for leukemia and often, when full constellation is absent, can be mistaken for other benign disorders e.g. viral infection.

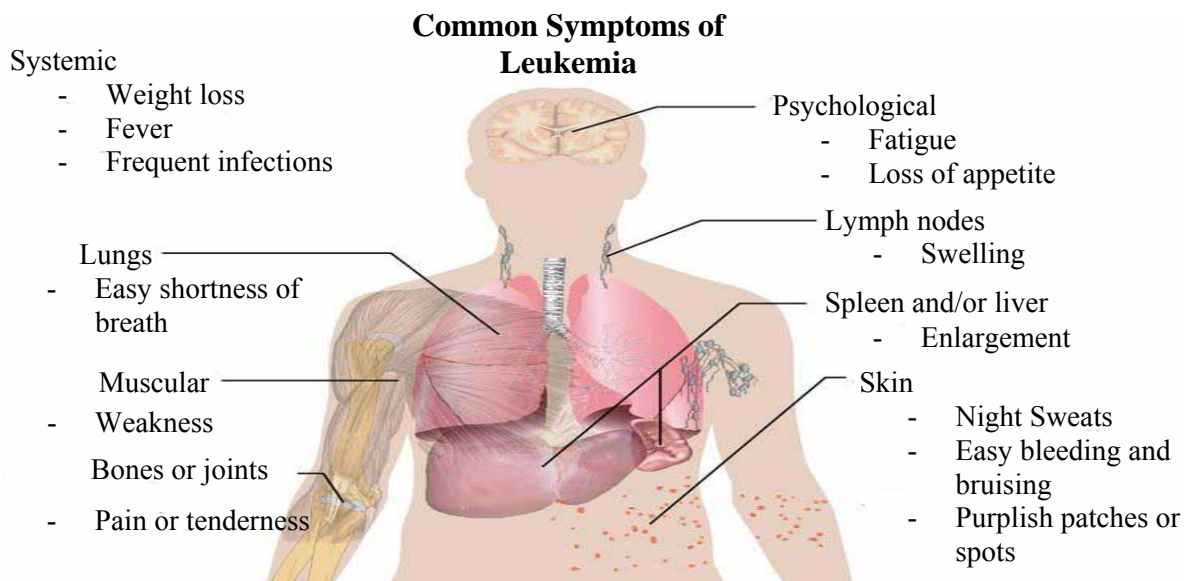


Figure 2.3: Common Symptoms of Leukemia (Häggström, 2009)

2.5 Laboratory Diagnosis of Acute Leukemia

Diagnosis of acute leukemia requires several laboratory tests. Normally, the doctor will go through the medical history to check out how long the symptoms have been present. The patient will go through a routine physical examination to detect abnormalities such as enlarged lymph nodes, and areas of bleeding. If the doctor suspects acute leukemia, microscopic morphological examination of the PB will be requested (Please Refer to Section 2.5.2).

Based on the results of the microscopic morphological examination, BM examination with additional laboratory tests such as BM aspirate morphological examination, immunophenotyping, and cytogenetics analysis (Please Refer to Sections 2.5.3, 2.5.4, and 2.5.5 respectively) would be necessary. (American Cancer Society, 2012).

Figure 2.4 presents the key steps that are required to be taken by a hematologist in order to diagnose a patient with acute leukemia. Table 2.4 provides a more detailed clarification of each step in Figure 2.4.

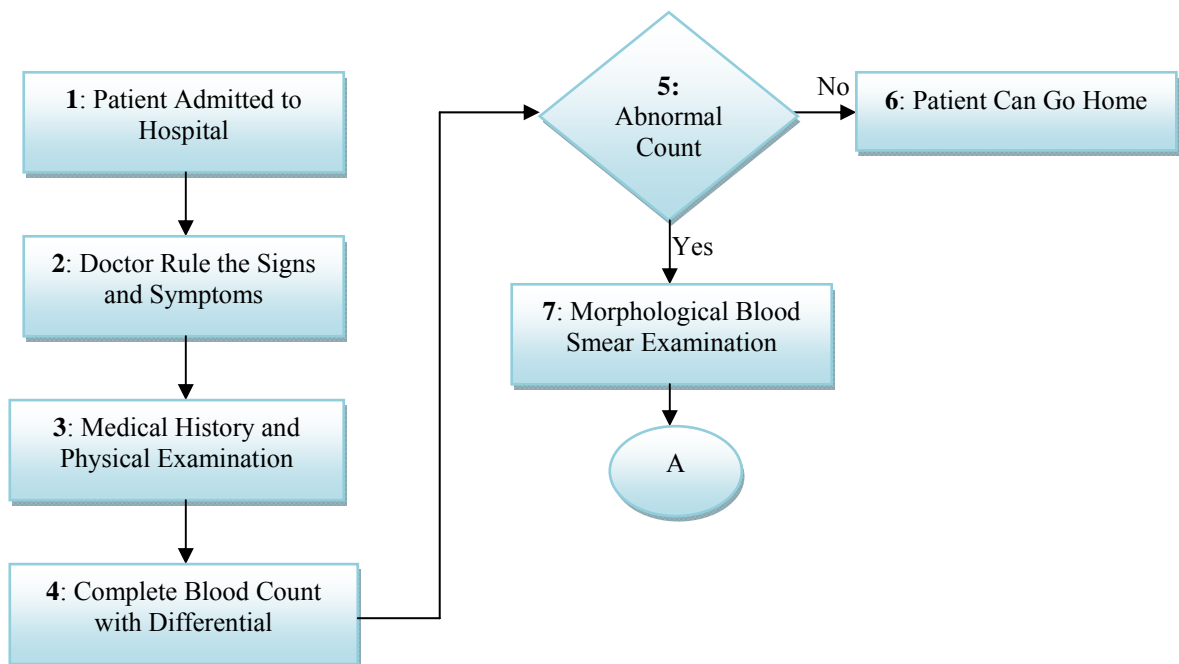


Figure 2.4: Steps to Confirm Acute Leukemia Diagnosis (Part A)

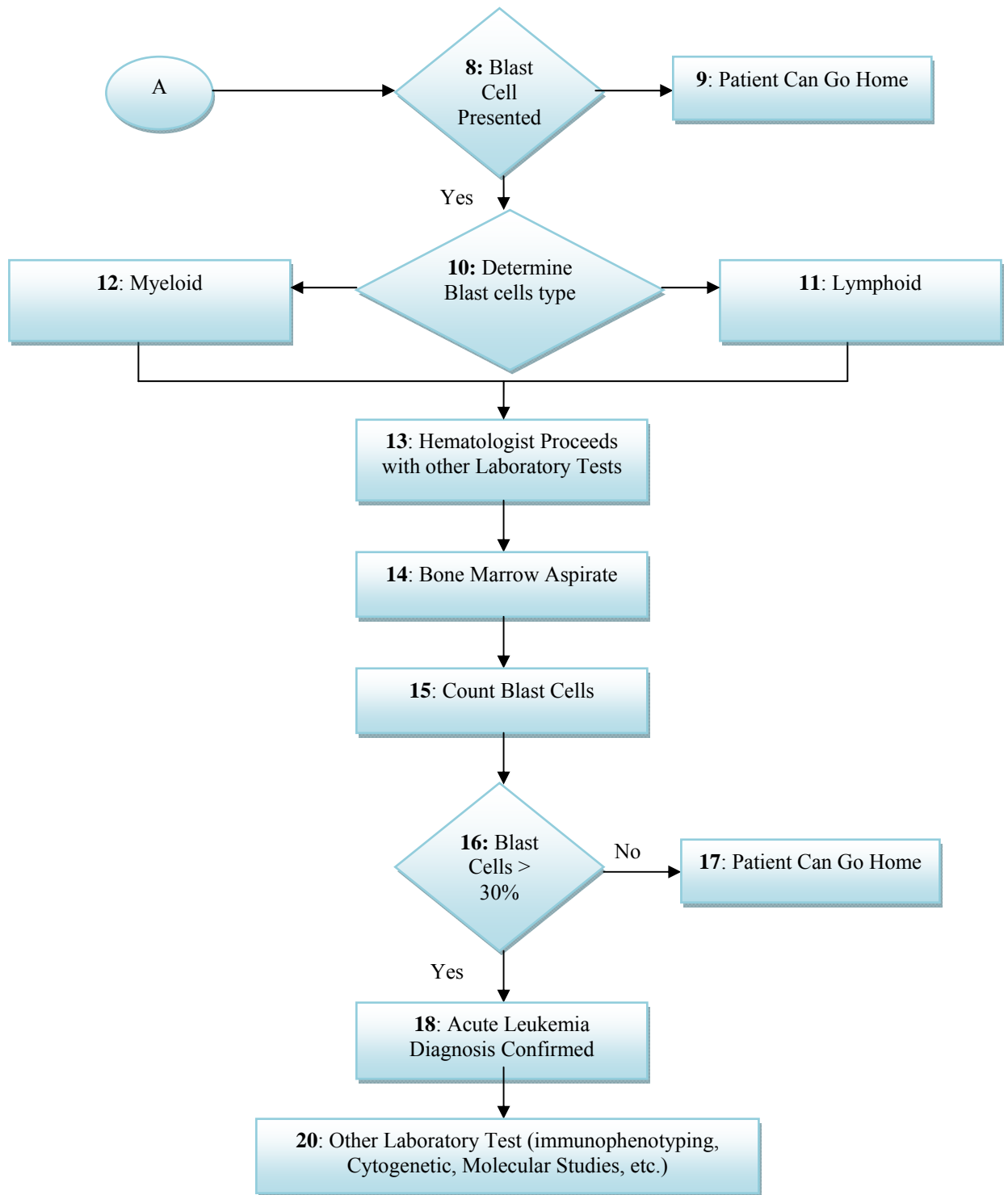


Figure 2.4: Steps to Confirm Acute Leukemia Diagnosis (Part B)

Table 2.4: Description of each step in the acute leukemia diagnosis process

Step	Description
1	Patient admitted to the hospital
2	Doctor will ask about the symptoms that the patient suffer from
3	Doctor will check the medical history of the patient to see for how the patient has been suffering from the presented symptoms. At this step the doctor will examine the patient physically looking for any other signs such as enlarge lymph nodes, area of bleeding, skin rushes, etc.
4	A Complete Blood Count with Differential is required to check if blood count abnormality exists. The count of white blood cells, red blood cell, platelets will be checked. In healthy human being the percentage of WBCs is as follow: <ul style="list-style-type: none"> • Neutrophils 50-70% • Eosinophils 1-4% • Basophils 0-1%, • Monocytes 2-8% • Lymphocytes 20.40%. The blood count is usually performed by the Hematology Analyzer and the blood sample is taken from the vein.
5	The blood count is performed by the Hematology Analyzer, and the count of the five types of WBCs is checked (Please Refer to Table 2.2 (a-b))
6	If the blood count reveals no evidence of leukemia and the blood count gives normal cells counting, other type of sickness needs to be investigated and probably the patient can go home.
7	In case of clinical suspicions and/or Hematology analyzer indicate some abnormalities in blood count, a PB smear is prepared and the slide is referred to a hematologist or laboratory practitioner for microscope morphological examination of PB smear.
8	The microscope morphological examination is performed to check if any blast cells are present in the smear.
9	If the blood count reveals no evidence of blasts in the PBF, other type of sickness need to be investigated and probably the patient can go home.
10	If blast cells are present in PB smear, the blast cells lineage should be determined
11	Lymphoid: Planning and management for ALL treatment
12	Myeloid: Planning and management for AML treatment
13	Based on the result of PB smear morphological examination, doctor proceeds with other laboratory test to confirm the initial diagnosis.
14	Bone Marrow biopsy is taken from the patient
15	Count the percentage of blast cells in the bone marrow
16	To confirm the presence of leukemia, blasts should account for about 30% of cells, based on the WHO classification. Less than 5% is considered normal.
17	If the percentage of the blasts cells is within the normal value recommended by WHO, other types of sickness need to be investigated and probably the patient can go home.
18	If the percentage of the Blast cells are more than 30%, the diagnosis is confirm and then the type of acute leukemia is determined in order to give the right treatment.
19	Other Laboratory tests are required (immunophenotyping, Cytogenetic, Molecular Studies, etc.)

2.5.1 Complete Blood Count (CBC)

CBC is one of the most valuable laboratory tests used in daily medical practice. It measures the cellular components of blood such as WBCs, RBCs, platelets, hemoglobin, differential WBC count, etc. (Sormunen, 2009). In many laboratories, CBC is carried out using an automated machine, namely, the hematology analyzer (Das, 2013). Figure 2.5 shows a picture of the automatic hematology analyzer from Sysmex (Diamond Diagnostics, 2013).

Automatic hematology analyzers utilize the “flow cytometry” principle which forces the blood samples to flow through a small aperture using a specific pressure and aperture setup. The cells and particles in the blood sample pass the aperture one by one. As cells flow through the aperture, several types of signals are recorded and then digitized for transformation, counting, histogram accumulation and further analysis (Qian, 2004). Pulse signal are also recorded, in which the number of pulses manifests the number of blood cells in a sample volume of blood, and the magnitude of each pulse manifests the size of each blood cell. However, sometimes automated counters are inadequate for counting very low numbers of white cells (Bain, 2006).



Figure 2.5: Sysmex KX21N hematology analyzer (Diamond Diagnostics, 2013)

One of the major drawbacks of automated cell counters is that they produce very limited information related to cell morphology (the shape, structure, form, and size of cells). Furthermore, they are unable to reliably classify blast cells. However, when abnormalities are present in the blood sample, results are automatically flagged. In this case, a microscopic morphological examination (Please Refer to Section 2.5.2) of the pathological cells is required (Ceelie et al., 2007).

2.5.2 Peripheral Blood Smear Morphological Examination

Microscopic morphological examination of stained PB and BM aspirate smears remains fundamental to the diagnoses of acute leukemia, and provides information for the separation of *ALL* and *AML*. Morphological examination of PB smear under the microscope has a major contribution to the diagnosis of almost any disease.

Morphological Examination of the PB smear is considered inexpensive. It is a very powerful diagnostic tool in both children and adults. It provides fast and reliable access to information about a variety of hematologic disorders. The smear offers a window into the functional status of the BM. Review of the smear is a prominent adjunct to other clinical data; in some cases, the PB smear alone is sufficient to establish a diagnosis (Bain, 2005).

Figure 2.6 illustrates the process of acquiring the blood sample from a patient and preparing the PB smear, later the PB smear is given to hematologist for microscopic morphological examination. Generally, the blood samples for acute leukemia diagnosis are taken from the vein. A drop of blood is placed on a glass microscope slide which is stained with a dye (Riley et al., 2012).

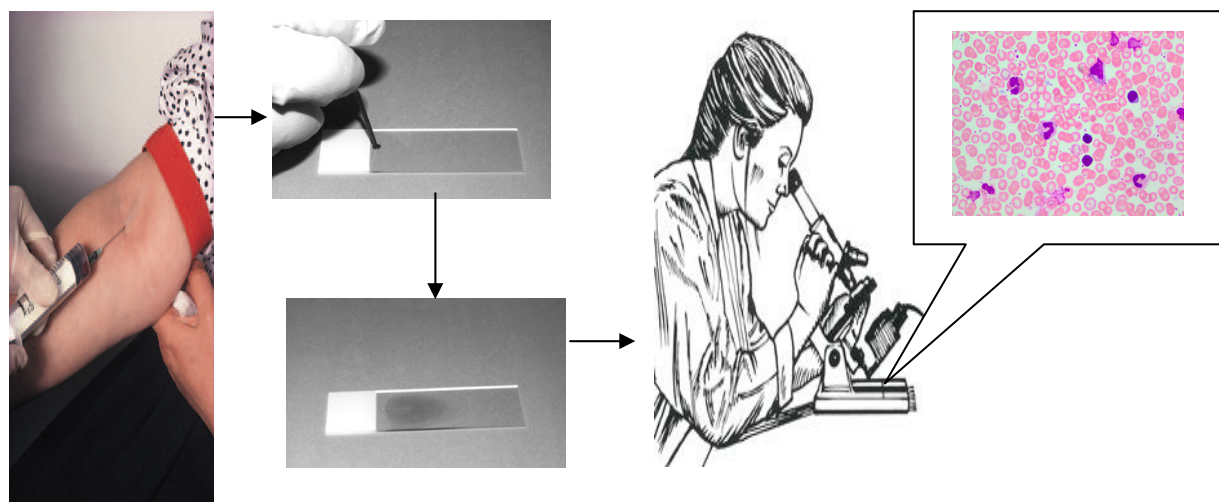


Figure 2.6: Illustration of PB smear preparation and examination

There are several different stains used in the preparation of a PB smear. Although, there is diversity amongst laboratories around the world, most stains used are based on the Romanowsky stain, developed by the Russian protozoologist in the late 19th century (Wittekind, 1979). Romanowsky used a mixture of old methylene blue and eosin to stain the nucleus of a malarial parasite purple and the cytoplasm blue. Subsequently, Giemsa modified the stain, combining methylene azure and eosin. The stain most commonly used in the UK is a combination of Giemsa's stain with May-Grünwald stain (eosin methylene blue); it is therefore designated the May-Grünwald–Giemsa (MGG) stain which is the same staining methods used in Malaysia. The stain most commonly used in North America is Wright's stain, which contains methylene blue and eosin; the methylene blue has been heated, or 'polychromed', to produce analogues of methylene blue. Sometimes this is combined with Giemsa's stain to give a Wright–Giemsa stain, which is generally held to give superior results (Bain, 2006).

Microscopic morphological examination provides a direct visualization to certain WBC features such as cell size, shape, nucleus chromatin structure, where the hematologists can use these features to distinguish between normal WBCs (Please Refer to Table 2.2 (a-b))

and blast cells (Please Refer to Table 2.5 , 2.6). Furthermore, these morphological features can be used to classify blast cells into either *ALL* or *AML*.

Manual microscopic morphological examination of PB smear may be greatly hampered by the following drawbacks:

- Poorly prepared or stained blood smears.
- Time consumption. Even though the PB smear examination is performed by an experienced specialist, it cannot be considered as a rapid process, where the operator has to do a careful study on the blast cells morphology (size, shape, nucleus chromatin structure) in order to come out with correct diagnosis and ensure that the right treatment will be given.
- The procedure is open to human error.

The need for microscopic morphological examination of PB smear may outstrip the availability of trained personnel to interpret them; hence leading to backlog and delay.

The contribution of the microscopic morphological examination of PB smear can be summarized as follows:

- It gives information about the patient condition where the count of each WBC type can indicate certain condition. For example, a blast cell count that is more than 30% indicates that the patient has leukemia (Sheikh et al., 1996).
- It provides data for selection of further pertinent tests in order to establish a diagnosis. For example, if blast cells are found in the PB smear, then the physician may need to do a cytogenetic test in order to estimate the patient's prognosis (Sheikh et al., 1996).

- It acts as a guide to therapy. In leukemia patients, the PB smear has to be examined on the 8th day of therapy where the numbers of circulating blast cells are determined. The patient with more than 1000 lymphoblasts per Microliter are regarded as “poor-risk” and will receive more intensive chemotherapy whereas those who are found to have less than 1000 lymphoblasts per Microliter of circulating blasts are considered “good-risk” and will have less intensive therapy (Ariffin, 2012, Madhlom et al., 2012b).
- It is used as an indicator of harmful effects of chemotherapy and radiotherapy. Chemotherapy may lead to the destruction of healthy cells such as RBCs. For that reason, the physician should carefully monitor this condition through a PB smear morphological examination (Abou-Alfa & DeMatteo, 2011).

2.5.3 Bone Marrow Aspirate Morphological Examination

BM is a special fatty tissue containing stem cells (Please Refer to Figure 2.2), located inside a few large bones. These stem cells can transform into WBCs, RBCs and platelets that have various roles. Inside this special tissue, immature stems cells reside, along with extra iron. Stem cells remain undifferentiated until abnormal, weakened, or damaged cells need to be replaced. This is the only process through which cells get replaced to maintain a healthy body (Orazi et al., 2006). Figure 2.7 shows an example of BM cells corresponding to *AML*. Figure 2.8, graphically depicts the process of taking a BM sample.

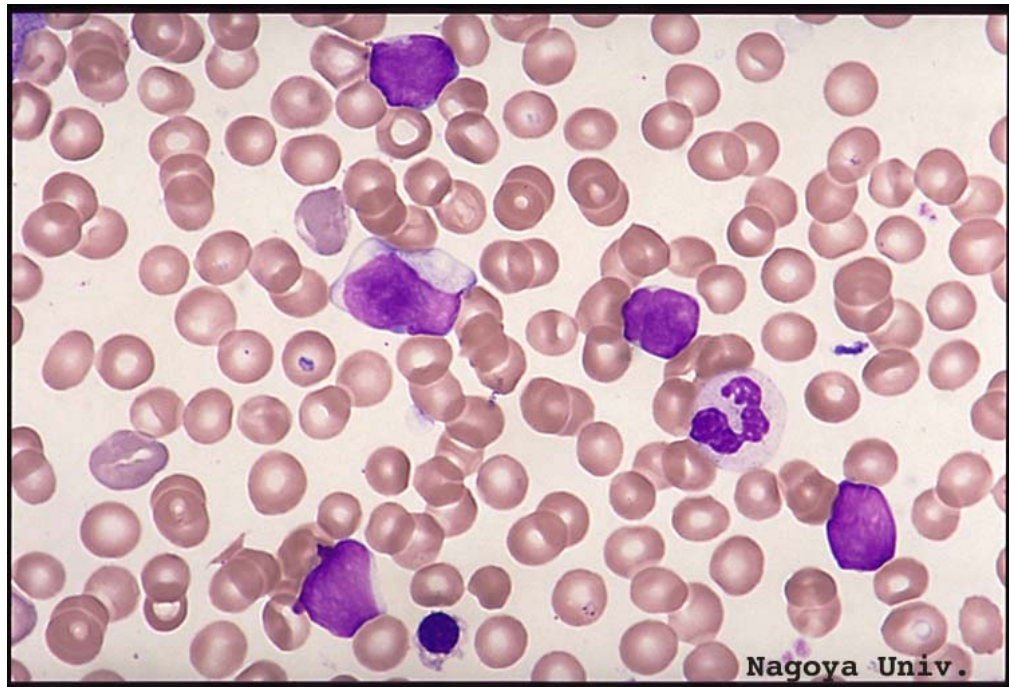


Figure 2.7: Bone Marrow sample (Ichihashi, et al, 2013)

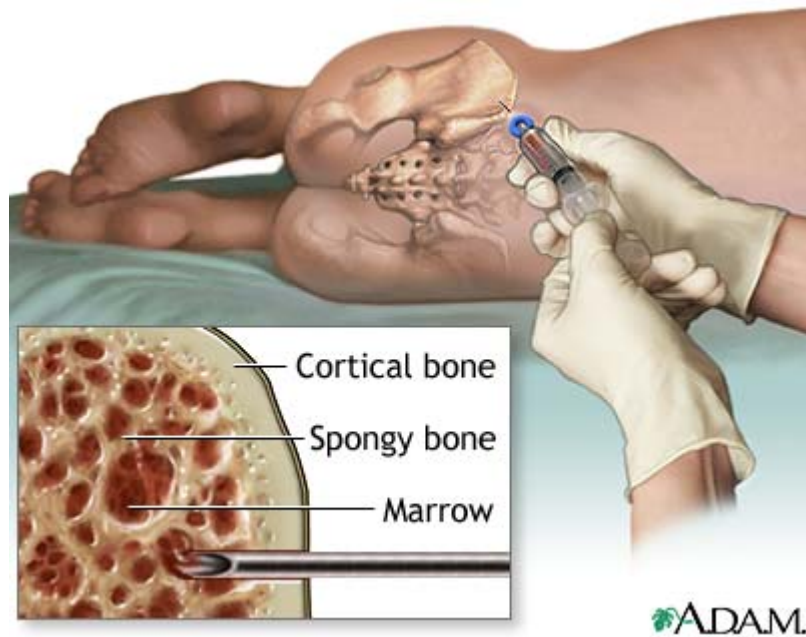


Figure 2.8: Blood taken from Bone Marrow (Dugdale,. 2010).

2.5.4 Immunophenotyping

Immunophenotyping is the process used to identify cells, based on the types of antigens or markers on the surface of the cell. By choosing appropriate antigens, the differentiation of leukemia cells can be determined (Wang, 2014). Immunophenotyping is performed using sophisticated medical equipment named flow cytometer. Flow cytometric techniques are expensive and require the expertise of highly trained personnel. In resource-limited countries ready access to technical support and quality assurance programs for flow cytometry are often not readily available (Zijenah et al, 2006). Nevertheless, the modern standard of care requires all acute leukemia cases to undergo immunophenotyping for proper classification (Jaroszeski & Heller, 1998).

2.5.5 Cytogenetics

In addition to morphological examination and immunological tests, cytogenetics analysis is another type of laboratory test used for leukemia diagnosis (Reaman, 2011).

Cytogenetic examination is usually performed by examining chromosomes (pieces of DNA) under a high-powered microscope to detect any changes. Normal human cells contain 23 pairs of a chromosome, each of which are of a certain size and stains a specific way. In some types of leukemia, chromosome changes may be seen. For instance, sometimes two chromosomes swap some of their genetic material, leaving one longer than normal and one shorter than normal. This change is called a translocation, and can usually be seen under a microscope. Recognizing these translocations can help identify certain types of *ALL* and *AML* and can help determine the prognosis (outlook) (American Cancer Society, 2012). More details regarding the classification of acute leukemia is described in the next section.

2.6 Classification of Acute Leukemia

The current acute leukemia classification systems are based on cytomorphology, cytochemistry, immunophenotyping, immunogenetics and molecular cytogenetics. However, the initial classification of acute leukemia is generally performed based on morphology (Szczepański et al, 2003). Although the best criteria for categorizing a case of acute leukemia as myeloid or lymphoid may be disputed, the importance of such categorization is beyond doubt. Not only does the natural history differ but the best current modes of treatment are still sufficiently different for an incorrect categorization to adversely affect prognosis (Bain, 1991).

Currently, there are two classification systems which use laboratory hematology to categorize acute leukemia into either lymphoid or myeloid. The following sections describe the characteristics of each classification type.

2.6.1 The French-American-British (FAB) Classification System

The FAB Cooperative Group has defined standardized criteria to establish the nature of acute leukemia and to categorize it into further subtypes. The FAB classification is solely based on morphological and cytochemical criteria of both PB and BM smears (Bennett et al., 1976, 1980, 1981, 1985, 1985a, 1991).

The development of the FAB classification of acute leukemia by a collaborating group of French, American and British hematologists was a major advance in leukemia classification, permitting a uniform classification of these diseases over two decades.

The FAB classification categorizes **ALL** into three subtypes (L1-L2-L3) and categorizes **AML** into eight subtypes (M0-M1-M2-M3-M4-M5-M6-M7). Table 2.5 and 2.6 summarize the morphological features of cells belonging to both types (**ALL**, **AML**) respectively (Cairo & Perkins, 2012; Erber, 2010; Tkachuk et al., 2007; Schrier, 2007).

Table 2.5: Morphological features of **ALL** subtypes based on FAB classification system

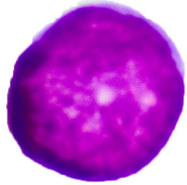
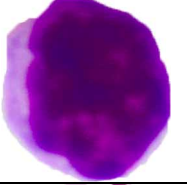
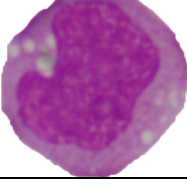
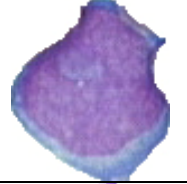
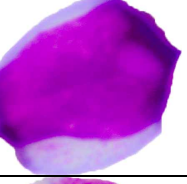
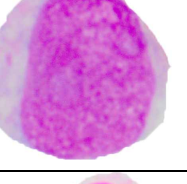
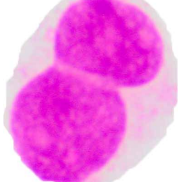
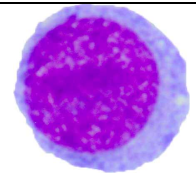
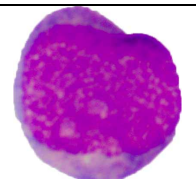
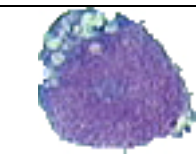
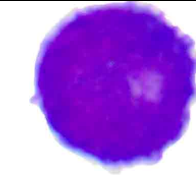
Acute Lymphoblastic Leukemia (ALL)					
Morphology	Classification	Description	Nucleus	Chromatin	Cytoplasm
	L1	Small blasts with little cytoplasm, little cell-to-cell variation	Round, homogenous	Slightly reticulated with perinucleolar clumping	Scant blue
	L2	Larger cells with greater amount of cytoplasm, greater cell-to-cell variation; irregular nuclei with multiple nucleoli	Irregular, inhomogeneous	Fine	Moderate pale
	L3	Large cells, strongly basophilic cytoplasm; often with vacuoles; nucleoli often multiple	Round to oval, homogenous	Coarse with clear parachromatin	Moderate blue prominently vacuolated

Table 2.6: Morphological features of **AML** subtypes based on FAB classification system

Acute Myeloid Leukemia (AML)					
Morphology	Classification	Description	Nucleus	Chromatin	Cytoplasm
	M0	Undifferentiated acute myeloblastic leukemia	Round to oval	Fine to coarse	Scant non-granulated
	M1	Acute myelocytic leukemia: cells very undifferentiated with only occasional granules	Round to oval	Fine	Scant, variably granulated
	M2	Acute myelocytic leukemia: cells more differentiated with granules, and often with Auer rods	Round to oval	Fine	Moderate azurophilic, granules with or without auer rods
	M3	Acute promyelocytic leukemia: hypergranular promyelocytes	Round to indented to lobed	Fine	Prominent azurophilic, granules and/or multiple auer rods

	M4	Acute myelomonocytic leukemia: both monocytes and myelocytes predominate	Round to indented folded	Fine	Moderate blue to gray, may be granulated
	M5	Acute monocytic leukemia: monoblasts with relatively agranular cytoplasm	Round to indented folded	Variable lacy or rosy	Scant to moderate gray-blue, dustlike lavender granules
	M6	Erythroleukemia: red blood cell precursors predominate, but myeloid blasts may also be seen	Single to bizarre	Open megaloblastoid	Abundant red to blue
	M7	Megakaryocytic leukemia: extremely variable morphology; may be diagnosed with monoclonal antibodies to platelets	Round to oval	Slightly to moderately reticulated	Scant to moderate gray-blue with blebbing

2.6.2 The World Health Organization (WHO) Classification System

Cytology and cytochemistry are fundamental to the acute leukemia diagnosis process but important and often essential information is also gained from immunophenotyping, cytogenetic analysis and molecular genetic (DNA or RNA) analysis. (Bain B.J., 2008). Recently, the World Health Organization (WHO), in conjunction with the Society for Hematopathology and the European Association of Hematopathology, published a new classification system for acute leukemia. The concepts that underlie this classification were derived from numerous published clinical and scientific studies and from the experience of more than 100 pathologists, clinicians, and scientists from around the world who collaborated to develop this consensus classification (Harris et al., 1999).

As mentioned earlier in section 2.6.1, the FAB classification system evaluates the blast cells mainly based on the morphological characteristics. On the other hand, the WHO classification system requires the additional evaluation of the blast cells based on molecular analysis and flow cytometry (Harris, 1999; Sachdeva et al., 2006; Falini et al., 2010;

Angelescu et al., 2012). Table 2.7 and Table 2.8 summarize the new classification of *ALL* and *AML* respectively as proposed by WHO.

Table 2.7: WHO classification system of *ALL*

Precursor B-cell ALL/LBL
Cytogenetic subgroups
t(9;22)(q34,q11),BCR/ABL
t(v;11q23);MLL rearranged
t(1;19)(q23;p13);PBX1/E2A
t(12;21)(p13;q22);TEL/AML1
Hypodiploid
Hyperdiploid, >50
Precursor T-cell ALL/LBL
Mature B-cell leukemia/lymphoma

Table 2.8: WHO Classification System of *AML*

Acute Myeloid Leukemia (AML) and Related Precursor Neoplasm
AML with recurrent genetic abnormalities
AML with t(8:21)(q22;22q); RUNX1-RUNX1T1
AML with inv(16)(p13.1q22)or t(16;16)(p13.1;q22);CBFB-MYH11
Acute promyelocytic leukemia with t(15;17)(q22;q12);PML-RARA
AML with t(9;11)(p22;q23);MLLT3-MLL
AML with inv(3)(q21q26.2) or t(3;3)(q21;q26.2);RPN1-EVI1
AML with mutated NPM1
AML with mutated CEBPA
AML with myelodysplasia-related changes
Therapy-related myeloid neoplasms
Myeloid sarcoma
Myeloid proliferations related to Down syndrome
Transient abnormal myelopoiesis
Myeloid leukemia associated with Down syndrome
Blastic plasmacytoid dendritic cell neoplasm

2.7 Leukemia Treatment Options

The treatment options of acute leukemia are dependent on several important factors. The most significant is the type or subtype of acute leukemia. Additional to this, there are other factors that should be taken into consideration to establish a treatment plan including cytogenetic abnormalities of the blasts as well as clinical features e.g. the patient's age and involvement of the central nervous system (CNS) (Pui, 2003). The mainstay of treatment is chemotherapy and in some protocols, children with CNS diseases are given cranial irradiation.

2.8 Leukemia Prognosis

Webster's new world medical dictionary defines prognosis as the likelihood of cure (WebMD, 2008). Prognosis also refers to the likely course and outcome of a disease (Celik et al., 2006).

Physicians normally use the 5-year survival rates to measure disease outcome. Survival rates include patients who survive 5 years after diagnosis, whether in remission, i.e. in a state during which the symptoms of the disease are abated.

Common factors in acute leukemia prognosis are as follow:

1. Age of the patient
2. Gender
3. WBCs count at presentation
4. Cancer spread to other body organs such as brain
5. Morphological, immunological, and genetic subtypes
6. Initial response to the treatment
7. Cytogenetic abnormalities

ALL is the most common form of cancer in children, one-fourth of all cancers in children belong to this type. It has a high incidence rate among adults, older than 45 years in age. Chemotherapy is the established treatment method for this disease. In the absence of chemotherapy and other cancer cure methods such as radiation therapy, a patient with **ALL** could survive for 4 months at the most. However, thanks to modern treatment methods, about 80% of the affected children are completely cured (Tecklin, 2008). Adults have been seen to have a 40% chance of complete cure (Greer et al., 2013). Acute leukemia prognosis will vary, depending on the disease progression, but children in the age group of 3 to 7 seem to have the highest chance of complete recovery (Pilgrim, 2010).

In contrast, **AML** is rare in children and is characterized by poorer outcomes. The 5-years event-free survival of childhood **AML** has been reported to be between 50-60% (Pui, 2012). Patients with **AML** require more intensive chemotherapy compared to those with **ALL**; however the treatment duration is shorter.

2.9 Summary

This chapter discussed blood and leukemia. General information about the blood components and the role of each component was initially presented. Next, the background study regarding leukemia and its types were discussed. The symptoms of leukemia and the current diagnosis methods in details with the advantages and disadvantages were subsequently presented together with the two acute leukemia classification systems (FAB and WHO). Towards the end of this chapter, the treatment options of leukemia and the prognosis was discussed.

CHAPTER 3

BACKGROUND AND LITERATURE REVIEW

3.1 Introduction

The purpose of this chapter is to highlight the key techniques and algorithms used in this research. The fundamental concepts of digital image processing and ML techniques used in acute leukemia diagnosis applications are also discussed. This is followed by a review of similar works in the area, including works that involved image processing and ML based methods.

3.2 Fundamental of Image Processing

In order to adequately grasp the methods discussed in this thesis, the concept of digital image processing is certainly needed. Particular emphasis is granted to image segmentation techniques.

3.2.1 Representation of Microscopy Blood Digital Images

Generally, digital image is defined as a discrete two-dimensional function $f(x, y)$ where x and y are spatial coordinates. The value of f at any coordinates (x, y) is called picture element (pixels). The amplitude of each pixel represents the intensity at that particular pixel (Gonzalez R et al, 2003). The resolution of a digital image is determined by the number of pixels in the image. In digital image, the spatial coordinate $f(0,0)$ represents the top left corner of the image and the spatial coordinate $f(x - 1, y - 1)$ represents the bottom right corner. Figure 3.1 shows a sample of microscopic PB image on a grid highlighted with the top left corner and the bottom right corner.

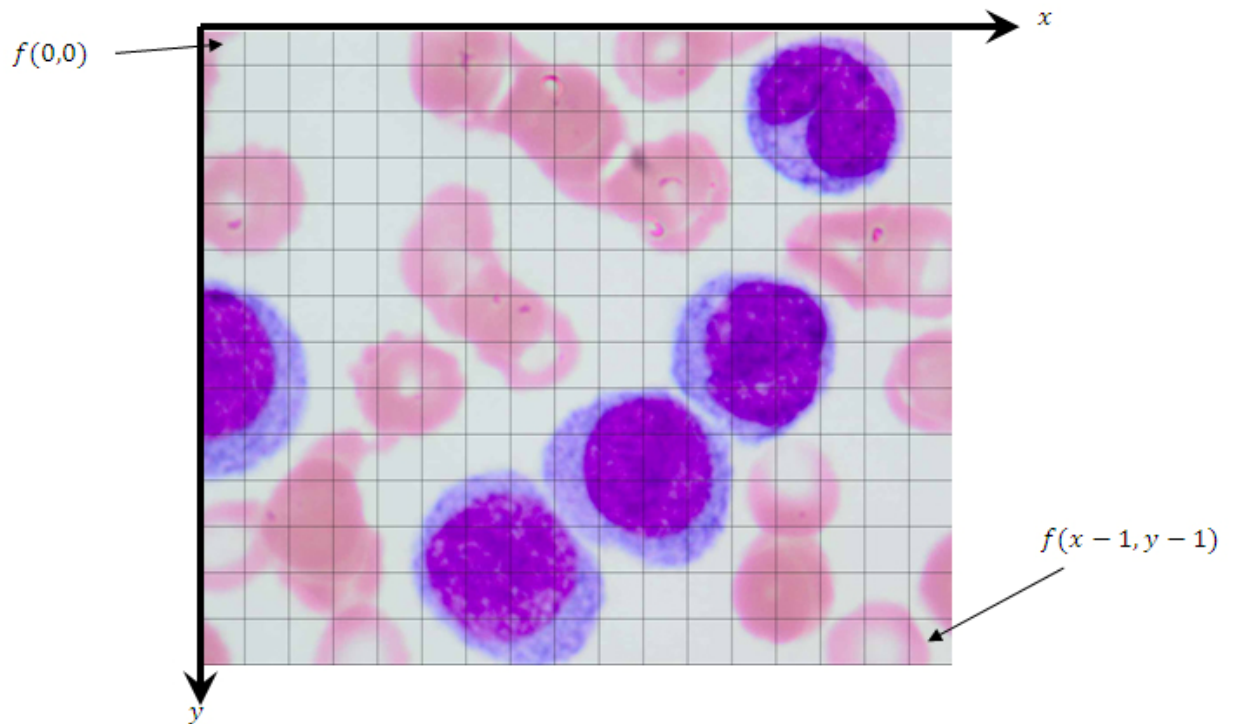


Figure 3.1: Representation of Microscopic PB digital image

Basically, there are several types of digital images. The most frequently used types are Color, Grayscale and Binary (Umbugh, 1998).

A *Binary image* consists of two pixels color either black where the pixel value is 0 or white where the pixel value is 1. The pixels having the value of 0 represent the background while the pixels having the value of 1 represent the foreground. A group of connected foreground pixels forms an object in the binary image (Wu et al., 2010). Figure 3.2 (a) shows the Mona Lisa image in binary format. On the other hand, *Grayscale image* measures the light intensity of brightness of an object shown at coordinates (x, y) of the image and is represented by a number called “gray level”. The gray level range can be between 0-255. The higher the gray level value the brighter the image will be at coordinate point (x, y) . The maximum value on the range of gray level represents a completely bright pixel while a pixel within the gray level 0 is completely dark (black). The pixels that are neither dark nor bright get a gray level value between 0 and the maximum value of brightness (255).

Figure 3.2 (b) depicts the Mona Lisa image in grayscale format (Najarian & Splinter., 2012, Gonzalez & Woods., 2002).

A *color image* has three values per pixels and they measure the intensity and chrominance of light. Each pixel of this plane is considered as a 3-D vector $x = \{x_R, x_G, x_B\}$. Figure 3.2 (c) shows the Mona Lisa image in color format. Color image can be modeled as three-channel grayscale image, where each band of data corresponds to a different color. The most common color spaces are the RGB (Red, Green, Blue), HSV (Hue, Saturation, Value), and the Lab (Luminance, chromatic components) (Jayaraman et al., 2011).

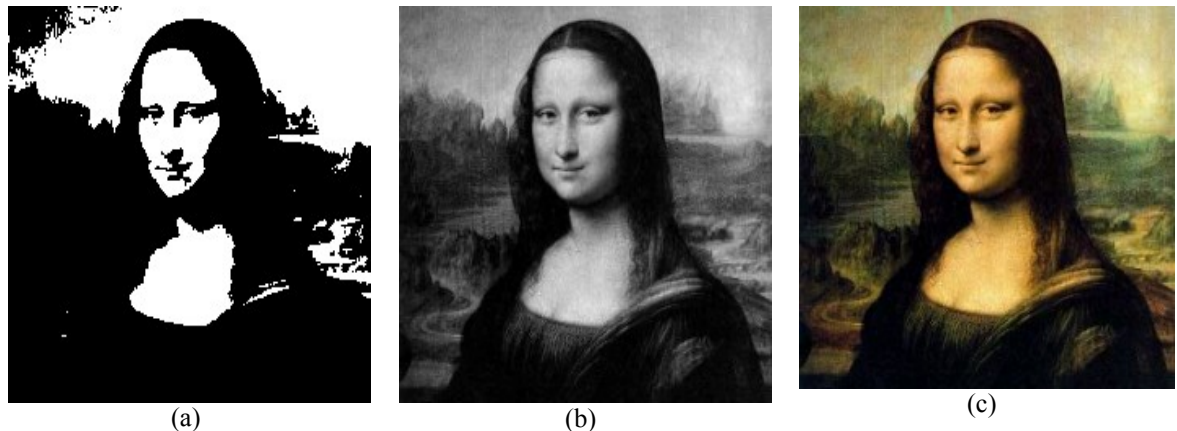


Figure 3.2: Digital image types (a) Binary image (b) Grayscale image (c) Color image (Mona Lisa, 2012)

3.2.2 Color Spaces in Microscopic Blood Images

Most of the medical imaging technologies developed at the early stages of the 20th century produce grayscale images such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and X-rays. Consequently, the color features are neglected in medical image processing applications. Nowadays, there are many medical instruments that produce a medical imaging with color information such as microscopy. Although, color is an effective feature that often eases the process of ROI detection and extraction in an image (Gonzalez et al., 2008). However, as compared to the grayscale image processing, the color image processing could be considered as a new area.

For that reason, colored medical image processing present new challenges for researchers where most of the techniques developed for binary image are often unsuitable for color image. A color space is an abstract mathematical model describing the way color can be represented as tuples of numbers, typically as three components (Pise et al., 2010). The careful selection of appropriate color space is significantly important for the performance of image segmentation (Tian et al., 2010). Skarbek et al. (1994) concluded that a highly efficient segmentation result can be obtained from color images rather than grayscale images. Given that, color images contain rich details, and it can provide a better description and differentiation among various image regions as compared to grayscale image (Kumar et al., 2006). PB images are usually captured in RGB color space, however, many works in the literature discovered that other color spaces such as HSV, Lab (Madhloom et al., 2012; Harun et al., 2010; Mohapatra et al., 2013) could be more useful than RGB in the extraction of blast cells. In the following sections, the most usable color spaces in blast cell detection process are discussed.

3.2.2.1 RGB Color Space

RGB is a convenient color model for computer graphics because the human visual system works in a way that is similar (though not quite identical) to an RGB color space (Soloman, 2009). It can be represented by a combination of three colors namely Red, Green, and Blue. Most of the cameras and emissive color displays represent pixels as a triple of intensities of the primary colors in the RGB color space. Hence, The RGB color space takes the form of cube of unit length as shown in Figure 3.3.

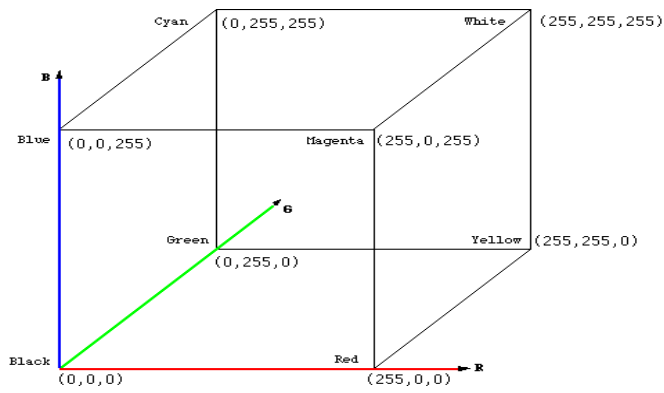


Figure 3.3: RGB Cube (Bourke, 1995)

The disadvantage of the RGB representation is that the channels are much correlated, as all of them include a representation of light (Jack, 2005; Tsagaris & Anastassopoulos, 2004), such as when the intensity of the light source is changed, then the three RGB components will simultaneously change. (Haifeng & Lanlan, 2010).

3.2.2.2 HSV Color Space

The HSV color model represents every color in three components namely Hue (H), Saturation (S), Value (V). It strongly represents colors in a way that is very similar to how the human eye senses color. The HSV is a very popular color space because it separates the pure color aspects from the brightness. Figure 3.4 depicts the HSV color space hexcone, where the Hue band is the angle around the vertical axis corresponding to the spectral frequency and it is arranged on a circle encoded from 0° to 360° . Saturation expresses how pure the color is, the more saturated a color is, the more vibrant and rich it will appear and it is represented by the distance from the central axis. While the Value represents the distance along the vertical axis, and denotes the color brightness.

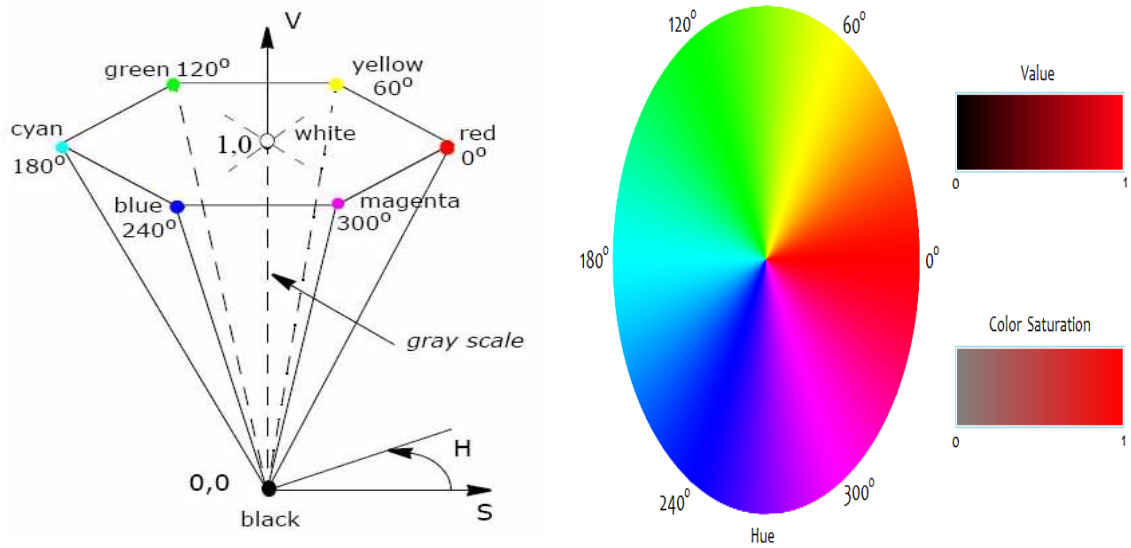


Figure 3.4: HSV Color Space (Intel Developer Zone, 2012)

The advantage of HSV over the RGB is that it appears more intuitive about color in terms of brightness and spectral name rather than the mixture coefficients of R, G, and B (Smith, 1978; Jack, 2001; Rüger, 2010).

3.2.2.3 Lab Color Space

Currently, the Lab color space is considered as one of the most popular uniform color spaces (Forsyth & Ponce, 2003). As shown in Figure 3.5, the Lab color space consists of three color components. The **L** represents lightness and it can take a value from 0 (black) to 100 (white). The **a** value represents the variation from $-a$ (greenness) to $+a$ (redness) while the **b** value represents the variation from $-b$ (blueness) to $+b$ (yellowness) (Sahin & Sumnu, 2006).



Figure 3.5: Lab color space (COLORROTATE, 2012)

The advantage of using the Lab color space is that it yields perceptually uniform spacing of colors, as the Lab is linear with visual perception, while the RGB is none linear. Another advantage is that the luminance factor L of the Lab color space could be discarded, as the luminance should be nearly constant for all pixels in the image. This could reduce the dimension of the data from 3 to 2, reducing data size and computation time (Sigurdsson et al., 2003). More details about the conversion from RGB to Lab color space can be found in (Hagen et al., 2008).

3.2.3 Image Segmentation

Image segmentation is the process of partitioning an image into disjoint and homogeneous meaningful regions with respect to some characteristics such as color, texture, etc. (Sonka et al., 2014). Errors in the segmentation process almost certainly lead to inaccuracies in any subsequent analysis (Wu et al., 2010). Image segmentation is a mandatory step in the development of any computer vision system (Pandey & Singh, 2010). Correspondingly, it is the primary step that significantly contributes to the analysis and evaluation of medical images in a computer-based diagnostic system (Arslan et al., 2014). The purpose of segmentation is to detect the border of the blast cells, in order to separate them from the background components such as plasma and RBCs. It is considered as the most difficult stage in the development of a computer-based acute leukemia diagnosis (Mao-jun, et al., 2008; Rezaatofghi & Soltanian-Zadeh; 2011, Patil, et al., 2012). The accuracy of the detected blast cells is crucial, as exclusion of any part of the blast cell may lead to loss of shape, color, and texture-based information. As shown in Table 2.5 and Table 2.6, the shape and the structural properties of the blast cells have great diagnostic importance; this is because blast cells from *ALL* type are different in shape than *AML* (Cairo & Perkins, 2012).

There are numerous factors that make the blast cells segmentation a challenging process such as: Low contrast between the blast cells and the surrounding background, irregularity and blurriness of the blast cells border, image artifacts such as excessive stain, microscope illumination, touching cells, color variation, etc.

Various image features such as color, gray level intensity, shape, and texture can be used to perform blast cells segmentation. Hence several approaches have been developed for blast cells extraction using PB images. Existing blast cells segmentation algorithms can be categorized into the following set of techniques: (1) Pixels-based threshold, which involves the determination of one or more threshold values that separate the ROI as a foreground and the rest regions as background, used in works by (Scotti, 2005; Sadeghian, et al, 2009; Nasir, et al., 2009; Aimi Salihah et al., 2010; Harun, et al., 2010; Halim, et al., 2011; Madhloom et al., 2012) (2) Edge-Based methods, which apply edge operators to determine the edges between the background and foreground regions applied by (Scotti, 2005; Sadeghian et al., 2009) (3) Region-Based methods which use region-merging and region-splitting algorithms to group the pixels into homogeneous regions (Osowski et al., 2004; 2009; Markiewicz et al., 2005; Siroic et al., 2007) (4) color-clustering methods which partition the color space into homogeneous regions using unsupervised clustering algorithms, applied by (Sabino et al., 2003; Mohaptra, et al., 2010; 2010b, 2011, 2011b, 2011c, 2013). (5) Morphological methods which use a predetermined seed and apply dilation, erosion or combination of both in order to detect the ROI, used in works by (Scotti, 2005; Khashman, & Al-Zgoul; 2009, Madhloom et al., 2012). (6) Active-contour methods such as snakes which use curve evolution techniques to determine the contours of the shape, as applied in the works by (Sadeghian et al., 2009). It can be seen that most of the previous works can be listed under several segmentation techniques.

This fact indicates that many of the previous studies have combined the results of different segmentation techniques. For example, Scotii (2005) developed a blast cell segmentation algorithm combining edge detection and mathematical morphology applied on grayscale PB images to extract leukocytes and pixel-based threshold to separate the nucleus from the cytoplasm.

In our early investigation, we combined mathematical morphology and pixels-based threshold applied on HSV color space to extract lymphoblast the from PB images that contain many blast cells (Madhloom et al., 2012). More details about the segmentation techniques used in the literature for the purpose of blast cells segmentation are discussed intensively in Section 3.4.2 of this chapter.

3.2.3.1 Selected Image Segmentation Techniques

Blast cells segmentation can be typically performed using various segmentation methods. This section discusses the segmentation techniques applied in this research for the purpose of blast cells segmentation.

3.2.3.1.1 Otsu Threshold

Otsu method is one of the most significant techniques for pixels-based threshold invented in 1979 by Nobuyuki Otsu (Otsu, 1979). It assumes that the image has two classes $\{C_1, C_2\}$ of pixels, namely, foreground and background then select the global optimal threshold by maximizing the between-class variance. Let I be an image represented with G gray levels $[0, 1, 2, \dots, G - 1]$. The number of pixels at gray level i denoted by n_i and the total number of pixels is represented by $N = n_1 + n_2 + \dots + n_G$. The probability of gray level i is denoted by (Otsu, 1979):

$$P_i = \frac{n_i}{N}, P_i \geq 0, \sum_{i=1}^G P_i = 1 \quad (3.1)$$

Figure 3.6 depicts a typical histogram of a bi-level image, where the image has two classes $\{C_1\}$ with gray level $[0, 1, \dots, T]$ and $\{C_2\}$ with gray level $[T + 1, T + 2, \dots, G - 1]$.

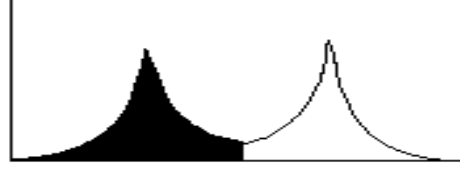


Figure 3.6: Typical histogram of a bi-level image

The gray level probability distributions of $\{C_1\}$ and $\{C_2\}$ can be obtained by equation 3.2 and 3.3 respectively

$$\omega_1 = \Pr(C_1) = \sum_{i=0}^T P_i \quad (3.2)$$

$$\omega_2 = \Pr(C_2) = \sum_{i=T+1}^{G-1} P_i \quad (3.3)$$

Then the means of the two classes can be calculated by equation 3.4 and 3.5 respectively

$$\mu_1 = \sum_{i=0}^T iP_i / \omega_1 \quad (3.4)$$

$$\mu_2 = \sum_{i=T+1}^{G-1} iP_i / \omega_2 \quad (3.5)$$

The total mean of the gray level is denoted by

$$\mu_T = \omega_1 \mu_1 + \omega_2 \mu_2 \quad (3.6)$$

The class variances are

$$\sigma_1^2 = \sum_{i=0}^T (i - \mu_1)^2 P_i / \omega_1 \quad (3.7)$$

$$\sigma_2^2 = \sum_{i=T+1}^{G-1} (i - \mu_2)^2 P_i / \omega_2 \quad (3.8)$$

The within-class variance is

$$\sigma_{within}^2 = \sum_{K=1}^M \omega_K \sigma_K^2 \quad (3.9)$$

The between class variance is

$$\sigma_{Between}^2 = \omega_1(\mu_1 - \mu_T)^2 + \omega_2(\mu_2 - \mu_T)^2 \quad (3.10)$$

The total variance of gray level is

$$\sigma_{Total}^2 = \sigma_{within}^2 + \sigma_{Between}^2 \quad (3.11)$$

Otsu method (Otsu, 1979) chooses the optimal threshold T by maximizing the between-class variance, which is equivalent to minimizing the within-class variance, since the total variance (the sum of the within-class variance and the between-class variance) is constant for different partitions (Otsu, 1979).

$$T = \arg \left\{ \max_{0 \leq T \leq G-1} \{ \sigma_{Between}^2(T) \} \right\} = \arg \left\{ \min_{0 \leq T \leq G-1} \{ \sigma_{within}^2(T) \} \right\} \quad (3.12)$$

3.2.3.1.2 Seeded Region Growing (SRG)

In 1994 Adams and Bischof introduced segmentation algorithm which is robust, rapid and free of tuning parameters known as Seeded Region Growing (SRG) (Adams & Bischof, 1994).

The essential idea behind SRG is that, the observation of the pixels belonging to one element of the object can possess similar properties, such as, the gray level value. Therefore if the considered pixel has gray level value that is near the common gray value of the region, this pixel can be associated into this region.

SRG is an iterative process initiated in a pixel from the set of seeds S_1, S_2, \dots, S_n . Pixels at the seed's border are subsequently labeled whether or not they are part of the same region as the seeds (Hirschmugl et al., 2007). The seeds are either chosen automatically based on some feature presented in the image or interactively according to the user opinion.

The SRG process develops inductively from the choice of seeds selected, known as, the initial state of the sets S_1, S_2, \dots, S_n .

In SRG, each step of the process performs addition of one pixel to any on the above sets. Then considering the state of the set S_i after m steps, let T be the set of all unallocated pixels (none labeled), bordering at least one of the regions S_i such that (Adams & Bischof, 1994):

$$T = \left\{ x \notin \bigcup_{i=1}^n S_i \mid N(x) \cap \bigcup_{i=1}^n S_i \neq \phi \right\} \quad (3.13)$$

Where $N(x)$ is the second-order neighborhood of the pixel of interest x as shown in Figure 3.7

$(x-1, y-1)$	$(x, y-1)$	$(x+1, y-1)$
$(x-1, y)$	(x, y)	$(x+1, y)$
$(x-1, y+1)$	$(x, y+1)$	$(x+1, y+1)$

Figure 3.7: The second-order neighborhood $N(x, y)$ of current testing pixel at (x, y)

If for, $x \in T$ we have that $N(x)$ meet just one of the S_i , then we can define $\varphi(x) \in \{1, 2, \dots, n\}$ to be that index such that $N(x) \cap S_{\varphi(x)} \neq \phi$ and $\delta(x)$ is a measure of how (x) is different from the region it joins. The simplest definition of $\delta(x)$ is (Adams & Bischof, 1994):

$$\delta(x) = \left| (g(x) - \text{mean}_{y \in S_{\varphi(x)}} [g(y)]) \right| \quad (3.14)$$

Where $g(x)$ is the gray level intensity of the image pixel x . If $N(x)$ meets two or more of the S_i , $\varphi(x)$ is taken to be the value of i such that $N(x)$ meets S_i and $\delta(x)$ is also minimized. In these circumstances, it is desirable to classify the pixel x as the boundary pixel and append it to set B , which is a set of already-found boundary pixels.

We then take $z \in T$ such that (Adams & Bischof, 1994):

$$\delta(z) = \min_{x \in T} \{\delta(x)\} \quad (3.15)$$

And append z to $S_i(z)$. This process completes step $m + 1$. This entire process is iteratively repeated until all pixels are allocated. In SRG, the process starts with each S_i being one of the seed sets. Thus, the definition of $\delta(x)$ in equation (3.13) and (3.14) ensures that the final segmentation is as homogenous as possible.

Practically, the criteria to choose the seed depends on the nature of the problem. For instance, if the targeted region needs to be detected using infrared images, the brightest pixels are chosen. The pixels homogeneity can be traced based on any characteristic of the ROI in the image such as texture, color, average intensity, etc.

3.2.3.1.3 Mathematical Morphology

Mathematical morphology is a non-linear process, which is considered as the basic foundation for many image processing algorithms. It can be used to investigate the geometrical structure in image by manipulating the original image with another image known as Structuring Element (SE) (Serra, 1982; Shih et al., 1995).

It has been proven that this technique is very useful for the analysis of biological and medical images (Wu et al., 1995). This processing technique has also proves to be a powerful tool for many computer-vision tasks in binary and gray scale images, such as image enhancement, noise suppression, edge detection, skeletonization, etc. (Ortiz et al., 2002).

Mathematical Morphology is based on simple mathematical concepts from set theory. Morphological operators are originally developed for binary images. However, it can also be used for gray level images.

It views binary images as assets of its foreground (1-valued) pixels, and set operations such as set union and intersection can be applied directly to sets of binary image (Gonzalez et al., 2004). The two fundamental mathematical morphology operators are *Dilation* and *Erosion*. *Dilation* is used to grow or thicken regions in a binary image, while in the gray level image; *Dilation* is used to brighten small dark areas, and to remove small dark "holes". According to (Gonzalez et al., 2004), *Dilation* on an image I by a structure element SE is denoted by $I \oplus SE$ and it is represented by the following Equation:

$$I \oplus SE = \{z \mid (\widehat{SE})_z \cap I \neq \emptyset\} \quad (3.16)$$

Where \widehat{SE} is the reflection of SE . It means that dilation of I by SE is done by reflecting SE and then shifting SE over I by z . On the other hand, erosion is used to shrink or thins region in binary image, while in a gray level image, erosion darkens small bright areas, and remove very small bright areas like noise spikes or small spurs. Erosion is represented by the following Equation (Gonzalez et al., 2004):

$$I \ominus SE = \{z \mid (SE)_z \subseteq I\} \quad (3.17)$$

The two basic morphological processes can be combined together to produce two more interesting operators, namely, *Opening* and *Closing*. The morphological opening (equation 3.18) is simply an erosion of I by SE followed by dilation of the result by SE (Gonzalez et al., 2004).

$$I \circ SE = (I \ominus SE) \oplus SE \quad (3.18)$$

Morphological opening is generally used to smooth region boundaries, break thin connection, and remove thin protrusions in images. On the other hand, morphological closing (equation 3.19) is performed by dilating I by SE and then eroding the result by SE (Gonzalez et al., 2004).

$$I \bullet SE = (I \oplus SE) \ominus SE \quad (3.19)$$

Unlike opening, morphological closing tends to join narrow breaks, fill long thin gulfs, and fill holes smaller than *SE* (Gonzalez & Woods, 2002; Gonzalez et al., 2004).

3.2.3.1.4 Watershed Segmentation

One of the most challenging problems in microbiological image processing is separating touching cells (Wilkinson & Schut, 1998). There are a number of factors that can lead to this type of problem during the process of PB smear preparation such as (i) the size of the drop of blood, (ii) the angle of the spreader slide and (iii) the speed at which the smear is made (Estridge & Reynolds, 2011).

The method that is usually preferred for separating touching, but mostly convex, features in an image is known as the watershed segmentation (Beucher & Lantéjoul, 1979; Lantéjoul & Beucher, 1981; Sun & Luo, 2009).

The watershed transform can be classified as a region-based segmentation approach. The intuitive idea underlying this method comes from geography: it is that of a landscape or topographic relief which is flooded by water, watersheds being the dividing lines of the domains of attraction of rain falling over the region. An alternative approach is to imagine the landscape being immersed in a lake, with holes pierced in local minima. Basins (also called 'catchment basins') will fill up with water starting at these local minima, and, at points where water coming from different basins would meet, dams are built. When the water level has reached the highest peak in the landscape, the process is stopped. As a result, the landscape is partitioned into regions or basins separated by dams, called watershed lines or simply watersheds. A simulation of the watershed transform is shown in Figure 3.8.

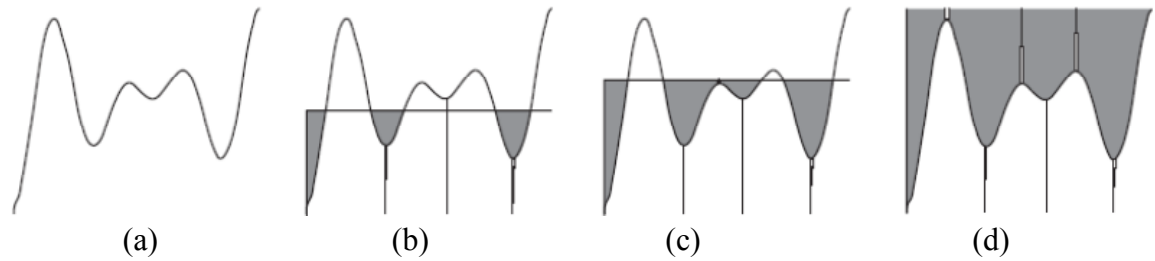


Figure 3.8 Simulation of the watershed transform. (a) Input image. (b) Punched holes at minima and initial flooding. (c) A dam is built when waters from different minima are about to merge. (d) Final flooding, with three watershed lines and four catchment basins. (Wu et al., 2010)

Advantages of the watershed transform include the fact that it is a fast, simple and intuitive method. More importantly, it is able to produce a complete division of the image in separated regions even if the contrast is poor, thus there is no need to carry out any post-processing work, such as contour joining, thus the watershed segmentation technique has been widely used in medical image segmentation (Ng et al., 2008), such as the segmentation of blood cell images (Nemane & Chakkarwar, 2012; Sharif et al., 2012), MRI brain images (Ng et al., 2006), Pap smear images (Plissiti et al., 2010; Orozco-Monteagudo et al., 2013), Colonoscopy images (Hwang et al., 2007) and many others.

3.3 Feature Extraction and Analysis

A major issue in any pattern classification system is the extraction of proper features that effectively differentiate various patterns (Osowski et al., 2009). In the last few decades, various feature extraction schemes were developed. Usually in feature extraction, the visual information of an image is analyzed in order to produce features such as shape, texture and color as summarized in Figure 3.9.

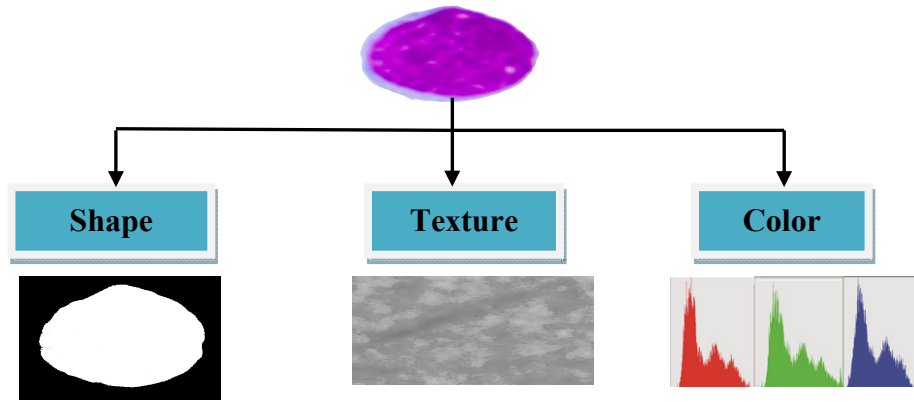


Figure 3.9: Image features description

It is widely agreed that there is no general powerful set of features that is suitable for different applications (Esposito & Malerba, 2001). Determining which features are the most effective is, of course, directly dependant on the application and the problem under study. Nevertheless, to improve the accuracy of detection, one or more features are often combined (Akilandeswari et al., 2012).

For the problem of acute leukemia diagnosis and classification as discussed in Chapter 2, acute leukemia is categorized into two main types (*ALL* and *AML*). According to the FAB classification, each acute leukemia type is classified based on certain morphological characteristics as illustrated in Table 2.5 and Table 2.6. Finding a quantitative measurement that can mimic the visual features used by hematologist to distinguish between the two types is not straightforward. It requires the use of various kinds of features such as shape, texture and color. However, it is worth mentioning that the usefulness of various feature sets cannot be assessed without considering the entire system, hence, it is necessary to try various kind of features that seem to be close to those that human experts would choose and select the best set of features based on the performance evaluation at the system level. Subsequently, in this research, three sets of features were used, namely,

shape, texture and color. The following sections discuss the theoretical details regarding each feature group.

3.3.1 Shape-Based Features

Shape is a prominent visual feature; it is considered as one of the fundamental features for object recognition. However, shape description is a very tough task, because it is difficult to define relevant shape features, and measure the similarity between look alike shapes. Furthermore, shape is often affected by noise, defection and occlusion (Zhang & Lu, 2004). Shape representation and description techniques are divided into two major categories namely *Boundary-based* representation and *Region-based* representation. Figure 3.10 shows the hierarchy of shape categories.

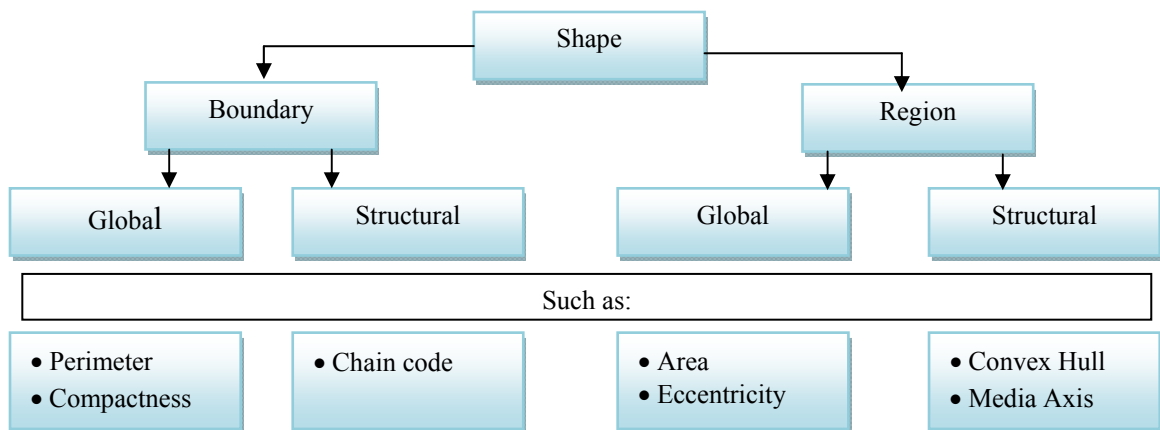


Figure 3.10: Classification of shape representation and description techniques.

Boundary-based approach exploits only the contour information of the ROI, and completely ignores the interior details such as, perimeter, circularity, eccentricity, etc. On the other hand; the *Region-based* approach takes into account the interior details as well as the boundary details of the ROI such as, area size, major and minor axis length (Pavlidis, 1978).

Under each class, the different methods are further divided into global approaches and structural approaches. This sub-class is based on whether the shape is represented as a whole or represented by segments/sections (primitives) (Zhang & Lu, 2004).

Although such features are rarely decisive for discrimination purposes, they are useful in distinguishing between various types of cells (Rodenacker & Bengtsson, 2003). The reader is referred to (Costa & Cesar, 2000) for further information on shape and size features. The blast cells are different in size, shape, the amount of cytoplasm, shape and amount of nucleus and the constituent in the cytoplasm (Ismail et al., 2010). In this research, the focus has been given to a set of shape features from both categories mentioned earlier in order to find out how much these shape features can contribute to the problem of acute leukemia blast cells recognition.

3.3.2 Texture-Based Features

Texture refers to the arrangement of the basic constituents of a material. In digital image the texture is depicted by the interrelationships between spatial arrangements of the image pixels. They are seen as changes in intensity patterns, or the gray tones (Osowski et al., 2004). There are two main types of texture, namely, tactile textures and visual textures. Tactile textures are related to the sense of touch such as, the feeling when we touch smooth or rough surface. The visual textures describe the visual perception that texture gives to a human viewer, and they are related to local spatial variations in terms of color, orientation and intensity in an image (Wilson & Moore, 2010). Figure 3.11 illustrates some patterns of different texture features.

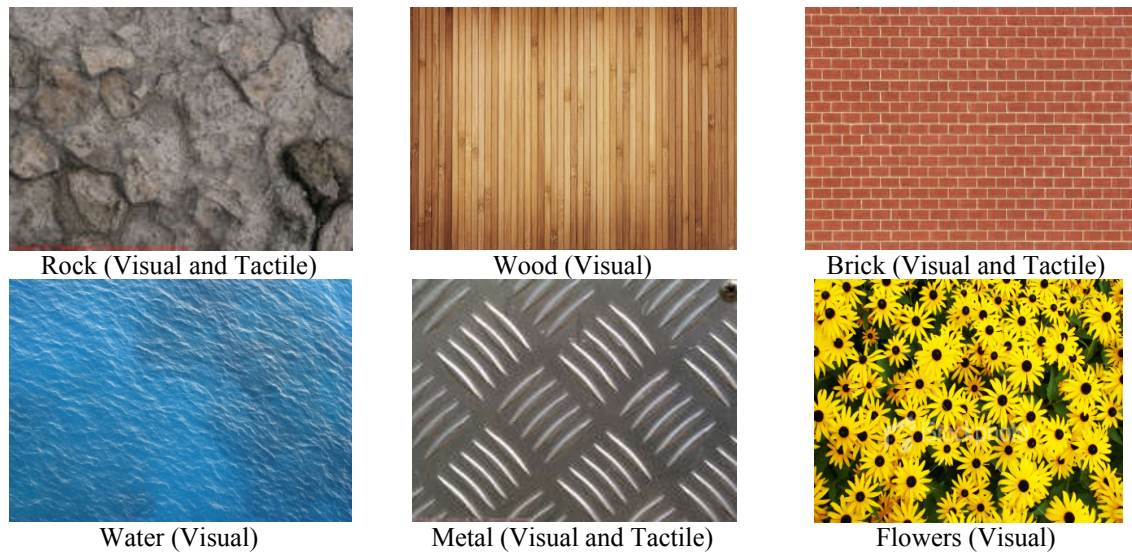


Figure 3.11: Samples of textures

Tuceryan & Jain (1993) divided texture feature extraction techniques into four main categories, namely: structural, model, transform, and statistical. In this research, two statistical based feature extraction methods are used namely, Histogram-based statistics and Gray Level Co-occurrence Matrix (GLCM).

Statistical-based methods represent texture indirectly by the non-deterministic properties that govern the distribution and relationship between the gray levels of an image. By computing local features at each point in the image and deriving a set of statistics from the distribution of the local features, statistical methods can be used to analyze the spatial distribution of gray values. Based on the number of pixels defining the local feature, statistical methods can be classified into first-order (one pixel), second-order (pair of pixels) and higher-order (three or more pixels) statistics. The difference between these classes is that, the first-order statistics estimate properties (e.g. average and variance) of individual pixel values by waiving the spatial interaction between image pixels, but in the second-order and higher-order, statistics estimate properties of two or more pixel values occurring at specific locations relative to each other.

The most popular second-order statistical features for texture analysis are derived from the co-occurrence matrix (Haralick, 1979). Methods based on second-order statistics (i.e. statistics given by pairs of pixels) have been shown to achieve higher discrimination rates than the power spectrum (transform-based) and structural methods (Weszka 1976; Castellano et al., 2004).

Generally speaking, medical images hold a significant amount of texture information which could be useful for clinical diagnosis. Texture information has been successfully used in the classification of various pathological tissues such as liver, thyroid, breasts, kidneys, prostate, heart, brain, lungs as well as in acute leukemia classification. (Bernasconi et al., 2001; Bonilha et al., 2003; Chen et al., 2002; Caselato et al., 2003; James et al., 2001; Ji et al., 2000; Sinha et al., 1997; Osowski et al., 2009; Mohapatra., et al, 2013).

Statistical-based methods are the most widely used texture-based analysis in medical images (Holli, et al, 2010). Moreover, Statistical-based methods prove to be superior and have achieved higher discrimination indexes compared to other texture analysis approaches such as structural or transform methods (Castellano et al., 2004). Therefore, in this research, two different statistical-based methods are selected for texture feature extraction, namely Histogram-based approach and Gray level Co-occurrence Matrix (GLCM). The next sections describe the theoretical background related to the selected texture analysis methods.

3.3.2.1 Histogram-Based Approach

The histogram $\{h\}$ of an image is calculated based on the frequency occurrence of each individual gray-level intensity value in the image. Therefore, the histogram contains the first-order statistical information about the image (Srinivasan & Shobha, 2008; Selvarajah & Kodituwakku, 2011).

Figure 3.12 (a-b) depict a histogram of an image with 16 gray level intensity values. The indices of the histogram element $i = (0 \dots 15)$ represent gray level intensity values. Each bin of the histogram represents the number of pixels at a particular intensity value. For instance, the gray level $i = 2$ in Figure 3.12 (a) contains 10 pixels as shown in Figure 3.12 (b). Other bins can be obtained from the histogram as shown in Figure 12 (a-b).

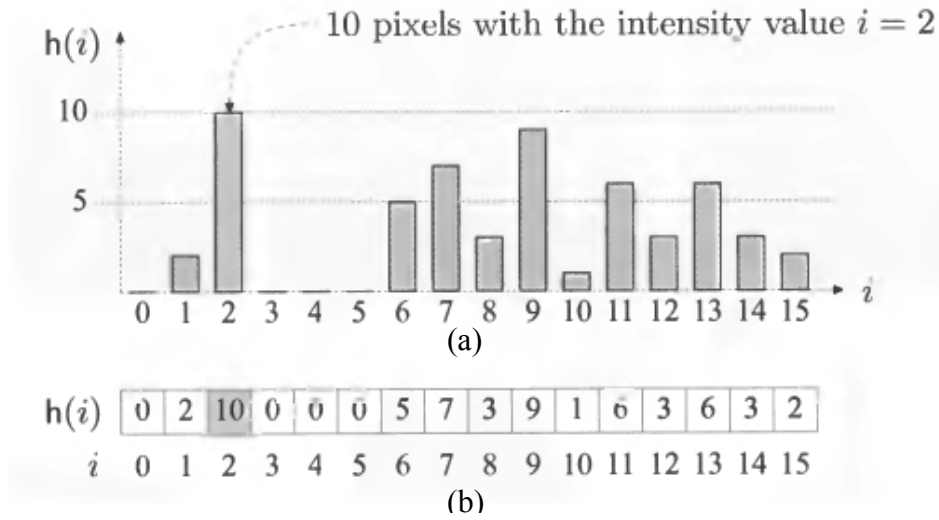


Figure 3.12: Histogram of image with 16 gray-level intensity (a) Histogram bins
(b) Number of pixels in each gray level intensity (Burger & Burge, 2009)

Several texture features can be extracted based on the histogram statistics such as mean, standard deviation, average energy, entropy, skewness and kurtosis (Suematsu et al., 2002). The mean describes the average level intensity of the image, whereas the standard deviation measures the dispersion of the histogram. The skewness and the kurtosis measure the dissymmetry and flatness of the gray level distribution (Díaz & Manzanera, 2011). The entropy is a measure of uniformity of the distribution, while the energy value measures how the pixels are distributed along the gray level range, histogram with many gray levels will have lower energy and vice versa (Marques, 2011). The texture features extracted from image histogram were proved to be very useful in many medical applications, especially in cancer diagnoses such as cervical cancer (Downey et al., 2013) breast cancer (Nithya &

Santhi, 2011), lung cancer (Shah et al., 2005), and acute leukemia (Nasir et al., 2013). The texture features based on the image histogram can be computed as shown in Table 3.1.

Table 3.1: Texture features extracted from gray level histogram

No.	Histogram Feature	Equation	Equation No.
1.	Mean	$\mu = \sum_{i=0}^{N-1} i h(i)$	(3.20)
2.	Standard deviation	$\sigma = \sqrt{\sum_{i=0}^{N-1} (i - \mu)^2 h(i)}$	(3.21)
3.	Energy	$eng = \sum_{i=0}^{N-1} h^2(i)$	(3.22)
4.	Entropy	$ent = - \sum_{i=0}^{N-1} \log_2(h(i)) h(i)$	(3.23)
5.	Skewness	$\mu_3 = \frac{\sum_{i=0}^{N-1} (i - \mu)^3 h(i)}{\sigma^3}$	(3.24)
6.	Kurtosis	$\mu_4 = \frac{\sum_{i=0}^{N-1} (i - \mu)^4 h(i)}{\sigma^4} - 3$	(3.25)

3.3.2.2 Gray Level Co-occurrence Matrix (GLCM)

Gray Level Co-occurrence Matrix (GLCM) (Haralick et al., 1973) is one of the most powerful and popular statistical texture analysis methods for extracting texture information from an image (Nikoo et al., 2011; Kumar, 2008). The GLCM is computed based on the estimation of second-order joint conditional probability density functions $P(i, j; d, \theta)$. As illustrated in Figure 3.13, each $P(i, j; d, \theta)$ is the probability that two neighboring pixels with grey levels i and j occur for a given distance d and direction θ . This yields a matrix of dimensions equal to the gray levels in the image, for each distance and orientation (d, θ) . Hence, the parameters required for computing GLCM are, the number of gray levels N_g , the distance between Pixels (d) and the angle (θ).

The number of gray levels is an important factor in computing the GLCM to represent a set of textures.

The more levels are included in the GLCM, the more is the computational cost of the texture statistics. The quantization merges similar gray levels within the image and thus reduces the noise-induced effects to some degree (Soh & Tsatsoulis, 1999). This is important to remark because if the texture patterns come from noise or artifacts, then the texture data could not be represented adequately by the GLCM. The distance (d) is used to specify the distance between pair of pixels. Normally the pair of pixels and neighbors; however, the matrix could also be computed for non-consecutive pixels. Similar to the distance parameter, the direction of the analysis is also another important parameter. The most common directions are 0° , 45° , 90° and 135° .

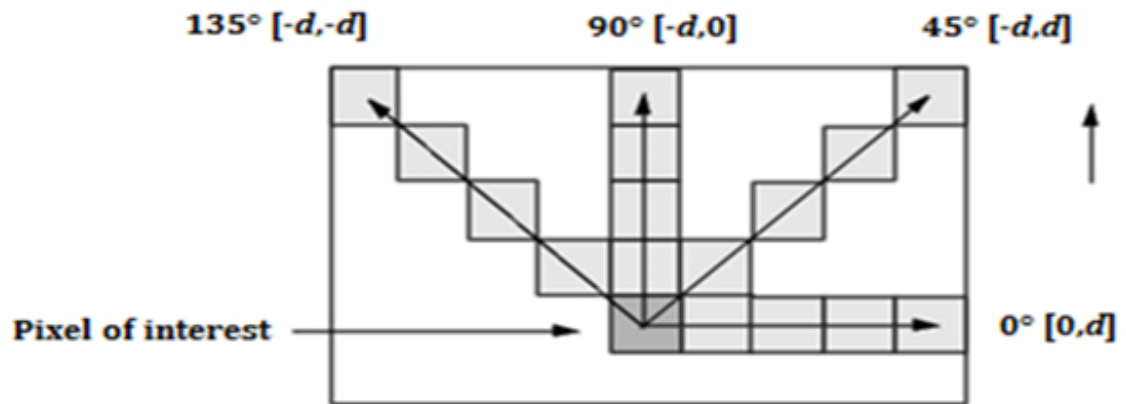


Figure 3.13: Spatial relationships of pixels defined by offsets, where d is the distance from the pixel of interest (Image Processing Toolbox (R2014b))

Figure 3.13 demonstrates the spatial relationship of pixels with respect to the distance (d) and the angle (θ). Figure 3.14 illustrates the way how the GLCM is computed, where the matrix on the left side of the figure is the input image and matrix C on the right side is the GLCM.

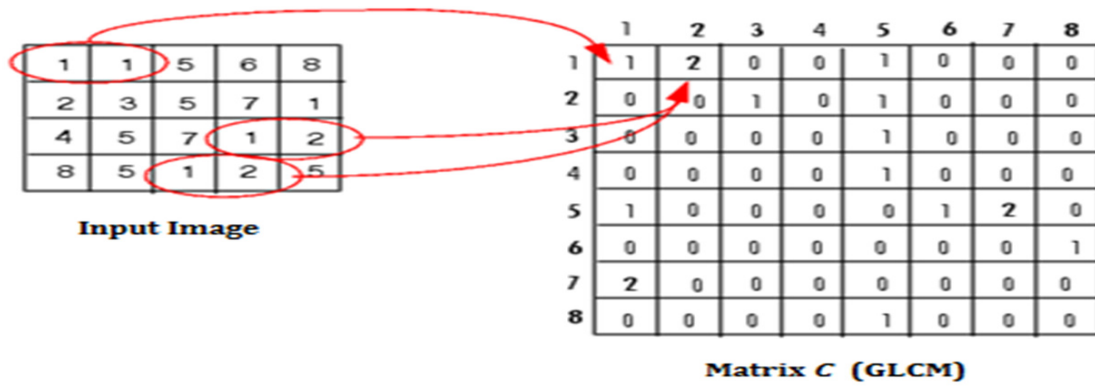


Figure 3.14: Illustration of the GLCM computation process (Image Processing Toolbox (R2014b))

As an example, consider the computation of the first two value of Matrix C in Figure 3.14. The element (1,1) contain the value 1 because there is only one occurrence in the input image where two horizontally adjacent pixels have the value 1 and 1, respectively. In a similar way, the element (1,2) in Matrix C contains the value 2 as there are only two occurrences where two horizontally adjacent pixels in the input image have the value 1 and 2, respectively. Through this process, the input image will be completely scanned for other occurrences of pixels (i,j) and the sum are placed in the corresponding element in the co-occurrence matrix. (Haralick et al., 1973) illustrated the applications of textural features based on GLCM on three different kinds of image data: photomicrographs of different kinds of sandstones (Haralick & Shanmugam, 1973) panchromatic aerial photographs of land-use categories, and earth resources technology satellite (ERTS) multispectral imagery containing land-used categories. Fourteen texture features were extracted in Haralick study namely: 1) Angular Second Moment, 2) Contrast, 3) Correlation, 4) Sum of Square Variance, 5) Inverse Difference Moment, 6) Sum Average, 7) Sum Variance, 8) Sum Entropy, 9) Entropy, 10) Difference Variance, 11) Difference Entropy, 12) Information Measure of Correlation I, 13) Information Measure of Correlation II and 14) Maximum Correlation Coefficient.

In more recent studies such as the work by (Soh & Tsatsoulis, 1999) who tried to classify seven different sea ice textural contexts based on synthetic aperture radar (SAR) imagery, concluded that, not all of the Haralick features could give the best classification result where only four features were used. However, (Soh & Tsatsoulis, 1999) proposed six new texture features based on GLCM which were found to be more useful for classification, namely, 1) Autocorrelation, 2) Cluster Prominence, 3) Cluster Shade, 4) Dissimilarity, 5) Homogeneity, and 6) Maximum Probability. Later (Clasui, 2002), who also addressed the problem of sea ice textual classification, suggested some improvement on one of the Haralick features. This is the Inverse Difference Moments where (Clasui, 2002) derived two new features from the Inverse Difference Moments, namely, 1) Inverse Difference Normalized and 2) Inverse Difference Moment Normalized.

The Matlab image Processing Toolbox provides a function, namely *graycoprops* which can be used to calculate four different GLCM feature, namely, 1) contrast, 2) correlation, 3) Homogeneity and 4) Energy. It is noticed that the computational procedure of the Contrast and Energy in the Matlab Image Processing Toolbox is similar to the ones found in (Haralick, 1973). However, the other two features (Correlation and Homogeneity) are calculated differently. The computation of all the GLCM texture features proposed by (Haralick, 1973; Soh & Tsatsoulis, 1999; Clasui, 2002 and Matlab Documentation, 2014) is shown in Table 3.2.

Table 3.2: GLCM Texture Features

No.	Ref.	GLCM Feature	Equation	Equation No.
1.	(Haralick et al,1973)	Angular Second Moment	$\sum_i \sum_j \{p(i,j)\}^2$	(3.26)
2.		Contrast	$\sum_i \sum_j i - j ^2 p(i,j)$	(3.27)
3.		Correlation	$\sum_i \sum_j \frac{(i,j)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$ <p>Where μ_x, μ_y, σ_x, and σ_y are the means and standard deviation of p_x and p_y, the partial probability density function</p>	(3.28)
4.		Sum of Square Variance	$\sum_i \sum_j (i - \mu)^2 p(i,j)$	(3.29)
5.		Inverse Difference Moment	$\sum_i \sum_j \frac{p(i,j)}{1 + (i - j)^2}$	(3.30)
6.		Sum Average	$\sum_{i=2}^{2N_g} i p_{x+y}(i)$ <p>Where x and y are the coordinates (row and column) of any entry in the co-occurrence matrix and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x + y$</p>	(3.31)
7.		Sum Entropy (S_{ent})	$-\sum_{i=2}^{2N_g} p_{x+y}(i) \log \{p_{x+y}(i)\}$	(3.32)
8.		Sum Variance	$\sum_{i=2}^{2N_g} (i - S_{ent})^2 p_{x+y}(i)$	(3.33)
9.		Entropy	$\sum_i \sum_j p(i,j) \log (p(i,j))$	(3.34)
10.		Difference Variance	$\sum_{i=0}^{N_g-1} i^2 p_{x+y}(i)$	(3.35)

11.		Difference Entropy	$-\sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$	(3.36)
12.		Information Measure of Correlation I	$\frac{HXY - HXY1}{\max\{HX, HY\}}$ <p>Where HX and HY are the entropies of p_x and p_y such that:</p> $HXY = -\sum_i \sum_j p(i, j) \log(p(i, j))$ $HXY1 = -\sum_i \sum_j p(i, j) \log\{p_x(i) p_y(j)\}$ $HXY2 = -\sum_i \sum_j p_x(i) p_y(j) \log\{p_x(i) p_y(j)\}$	(3.37)
13.		Information Measure of Correlation II	$(1 - \exp[-2(HXY2 - HXY)]^{1/2}$	(3.38)
14.		Maximum Correlation Coefficient	$\sqrt{\text{Second Largest eigenvalue of } Q}$ <p>Where</p> $Q(i, j) = \sum_k \frac{p(i, k) p(j, k)}{p_x(i) p_y(k)}$	(3.39)
15.	(Soh & Tsatsoulis, 1999)	Autocorrelation	$\sum_i \sum_j (ij) p(i, j)$ <p>Where $p(i, j)$ represents the number of occurrences of gray levels (i and j)</p>	(3.40)
16.		Cluster Prominence	$\sum_i \sum_j (i + j - \mu_x - \mu_y)^4 p(i, j)$ <p>Where $p(i, j)$ is the $(i, j)^{th}$ entry in a normalized GLCM. The mean for rows and columns of the matrix are:</p> $\mu_x = \sum_i \sum_j i \cdot p(i, j)$ $\mu_y = \sum_i \sum_j j \cdot p(i, j)$	(3.41)
17.		Cluster Shade	$\sum_i \sum_j (i + j - \mu_x - \mu_y)^3 p(i, j)$	(3.42)
18.		Dissimilarity	$\sum_i \sum_j i - j \cdot p(i, j)$	(3.43)
19.		Homogeneity	$\sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j)$	(3.44)
20.		Maximum Probability	$\underset{i, j}{MAX} p(i, j)$	(3.45)

21.	(Clausi, 2002)	Inverse Difference Normalized	$\sum_i \sum_j \frac{p(i,j)}{1 + i - j ^2 / N_g}$	(3.46)
22.		Inverse Difference Moment Normalized	$\sum_i \sum_j \frac{p(i,j)}{1 + (i - j)^2 / N_g}$	(3.47)
23.	Matlab	Homogeneity	$\sum_i \sum_j \frac{1}{1 + i - j } p(i,j)$	(3.48)
24.		Correlation	$\sum_i \sum_j \frac{(i - \mu_x)(j - \mu_y) p(i,j)}{\sigma_x \sigma_y}$	(3.49)

Countless number of works in the literature used the GLCM for the purpose of texture feature extraction. Computer-based cancer diagnosis is one of the active areas that exploit GLCM features. An example of this is the work carried out by (Yang et al., 2012) where GLCM was used to extract sonographic texture features from ultrasound images for parotid gland injury detection which is one of the most common side effects of head-and-neck cancer radiotherapy. The author observed that GLCM features can significantly differentiate between normal parotid gland and postradiotherapy parotid glands. GLCM was also successfully used for colon cancer detection by classifying cancer and non-cancer Colonic histopathology images (Jiao et al., 2013). The work by (Nithya & Santhi, 2011) highlighted the effectiveness of the GLCM features in distinguishing between normal mammogram images and malignant ones for breast cancer detection. Hematological malignancy detection is another domain where GLCM was successfully implemented, such as, the work by (Madhloom et al., 2012b) where GLCM features were used along with shape features to differentiate between *ALL* and normal lymphocytes,

3.3.3 Color-Based Features

Color features describe the color distribution of the images, which are the most discriminative features of blood and bone marrow cells (Díaz, & Manzanera, 2011).

Since any pixel in a colored image can be described by three components in a certain color space such as the Red, the Green, and the Blue channels of the RGB color space. Hence, a histogram of each color channel can be defined, such that, three different histograms can be produced, each of which belongs to a different color channel. Hence, the same features extracted from gray level histogram (Please Refer to Section 3.3.2.1) can be used to extract color features (Reta et al, 2010; Nandagopalan et al., 2010).

3.4 Feature Selection

A general practice in pattern classification is to extract a feature set as large as possible in order to cover all aspects of the phenomenon under analysis. This practice is widely acceptable. However, large number of feature can lead to problems (Kuncheva, 2004). This is where feature selection comes into the picture. Feature selection (also known as variable selection, attribute selection or subset selection) is the process of choosing a powerful subset of features that can efficiently classify the target classes (Li et al., 2007; Guyon & Elissee, 2003; Liu & Yu, 2005).

In most of the cases in pattern recognition, many of the features can be irrelevant or redundant; however, if no feature selection is performed on the dataset, these features would be included in the dataset yielding a low classifier performance. Besides that, a high dimensionality of features makes the classification task a difficult problem, because of the effect known as, the curse of dimensionality (Bellman et al., 1961). There are several advantages of feature selection: (1) dimensionality reduction in order to reduce the computational cost; (2) reduction of noise in order to improve the classification accuracy;

(3) more interpretable features or characteristics that can help identify and monitor the phenomenon under study (Ding & Peng, 2005).

Feature selection techniques are grouped into three different categories, namely, filter methods, wrapper methods and embedded methods. This categorization is formed based on how the feature selection search is combined with the construction of the classification engine (Jain et al, 2000; Guyon & Elissee, 2003; Saeys et al., 2007).

Filter methods do not take into account the properties of the classifier, as it performs statistical tests to rank the features. Feature set is arranged based on the score, and then low-scoring features are removed (Saeys et al., 2007, Guyon & Elissee, 2003).

Filter methods ignore feature dependencies; this may lead to worse classification performance when compared to other types of feature selection techniques such as wrapper or embedded methods (Saeys et al., 2007).

The second feature selection approach is wrapper, in this method; the feature selection process is wrapped around the learning algorithm. It searches the usefulness of different feature sets by judging the estimated accuracy of the learning algorithm. Various search methods can be used in this process, such as exhaustive search which is generally searched over all feature sets. However, this method is intractable due to the exponentially large number of possible sets. Alternatively, some other search methods employ a variety of heuristics such as forward search and backward search.

Advantages of wrapper approaches include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. On the other hand, these techniques are classifier dependent and more computationally intensive than filter methods (Jain et al, 2000; Guyon & Elissee, 2003; Saeys et al., 2007).

The third class of feature selection techniques is embedded methods. Embedded methods are carried out by combining the learning part and the feature selection part. It directly uses the parameters of a classifier rather than using the classifier as a black box to estimate the classification accuracy. Just like wrapper methods, embedded methods are also classifier dependent (Guyon & Elissee, 2003; Saeys et al., 2007).

3.5 Pattern Classification

Pattern classification is the process of classifying input patterns (e.g. blast cells) to one of a predefined set of classes (e.g. *ALL*, *AML*) based on the features which have been extracted in the feature extraction stage (e.g. Shape, Texture, Color) (Patel & Marwala, 2006).

There are several different methods to create a pattern classifier; the two most popular classifiers are, namely, Artificial Neural Network and Support Vector Machine (Cristianini & Shawe-Taylor, 2000; Yao et al., 2001) which are described in the next subsections.

3.5.1 Artificial Neural Network

Artificial Neural Network (ANN) is an information-processing system that attempts to imitate biological neural networks (Sivanandam & Deepa, 2006). The neural network was invented in order to overcome the technical limitation of the computer's ability to perform certain tasks. These tasks, such as, reading a handwritten document or recognizing a face, may seem simple for human beings, but are difficult for even the most advanced computers (Abdul-Kareem et al., 2000)

The applications of ANN in biomedicine have gained a tremendous interest from many researchers because of its ability to perform non-linear data processing with relatively simple algorithm (Cohen & Hudson, 1999).

ANN is composed of large number of highly interconnected processing elements namely neurons, working together to solve a specific problem (Deepa, 2006). Figure 3.15 diagrammatically depicts a common ANN hierarchal architecture composed of several layers. The layers are connected and the neurons are organized along these layers. The network is linked to the outside environment through the neurons of the input and output layers.

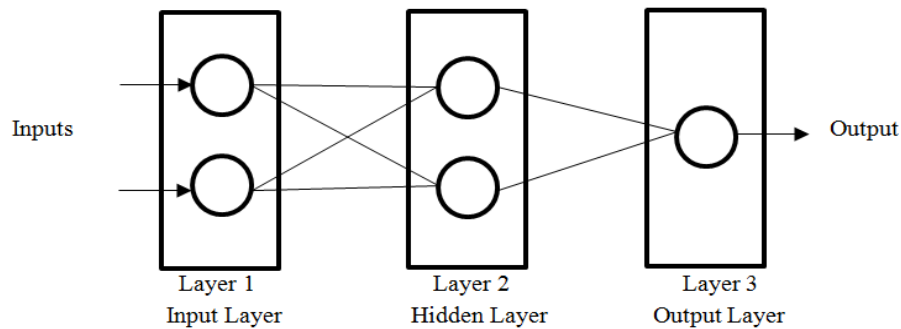


Figure 3.15: Neural Network (Verma and Blumenstein, 2008)

An artificial neuron is the basic component and fundamental unit that performs a simple mathematical operation on its inputs and imitates the functions of a biological neuron and its unique process of learning (Hayati & Shirvany, 2007). An artificial neuron receives multiple inputs and calculates its output which corresponds to the impulse frequency of a real neuron.

ANN performs its processing by accepting inputs, x , which are then multiplied by a set of weights, w . The neurons then, nonlinearly transform the sum of the weighted inputs, by means of an activation function into an output value y , as illustrated in Equation 3.50. Figure 3.16 shows the basic architecture of an artificial neuron.

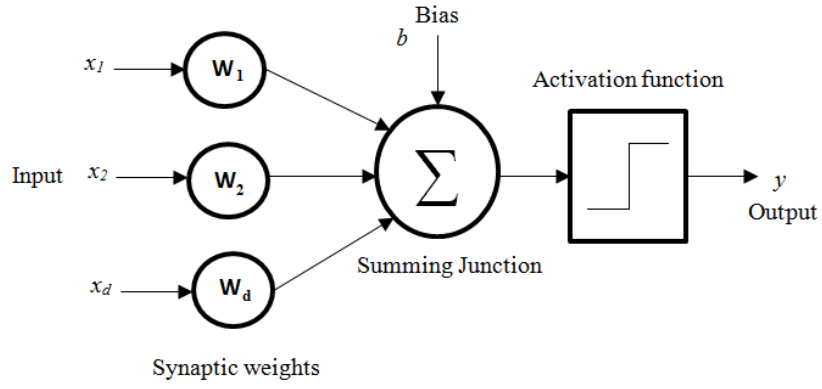


Figure 3.16: Basic Architecture of Artificial Neuron (Negnevitsky 2005)

$$y = \sum_{i=1}^d w_i x_i + b \quad (3.50)$$

The output of a neuron y , thus, depends on the neuron's input and on its activation function. Sometimes a bias b is also added to the network. The bias is then regarded as a weight, with a constant input of 1 (Fausett, 1994; Negnevitsky, 2005). There are many kinds of neuron activation functions, such as the logistic function, the hyperbolic-tangent function the sigmoid function, etc. of which the sigmoid function is the most widely used (Zhen-Zhen & Su-Yu, 2012).

3.5.1.1 Multilayer Perceptron Feed-Forward Neural Network

Various ANN architectures are used for classification or prediction purpose. However one of the most common is the multilayer perceptron feed-forward neural network (MLP-NN). In a MLP-NN, the connections between neurons in each layer are unidirectional where the information being processed pass through the input layer, to the hidden layer(s), and then to the output layer. An example of a MLP-NN with two hidden layers is shown in Figure 3.17.

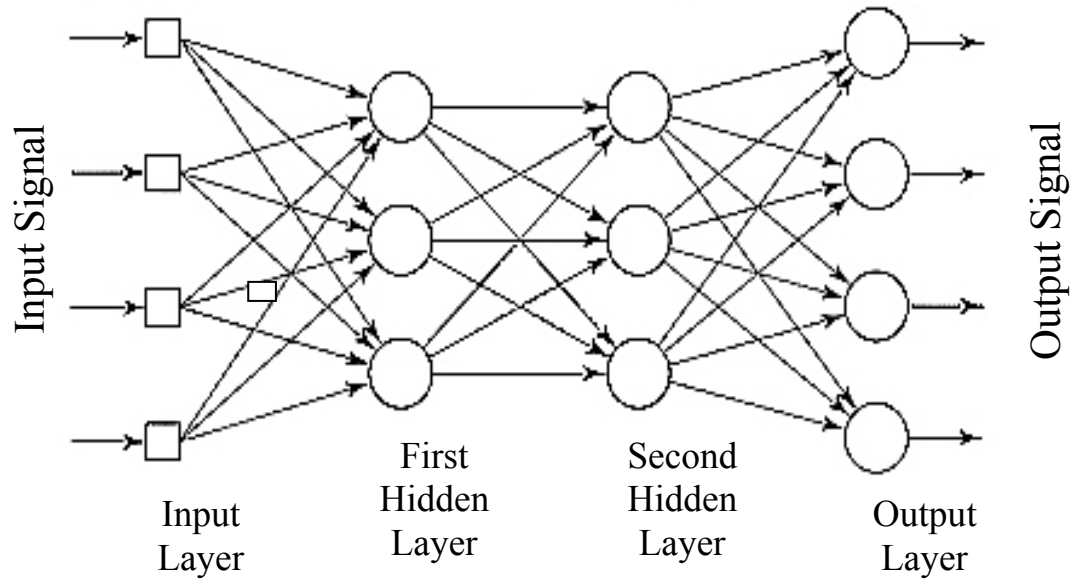


Figure 3.17: MLP-NN with Two Hidden Layers (Negnevitsky 2005)

There are several works implemented the MLP-NN for the purpose of blast cells classification, such as the works by (Scotti, 2005, Mohapatra, et al., 2013).

The design of the typical MLP-NN consists of, an input, at least one hidden layer, and an output. Theoretically, there is no specific limit on the number of hidden layers. However, in most cases, one or two hidden layers are adequate (Bishop, 1995).

MLP-NN commonly uses the Back-Propagation (BP) (Jiang et al., 2010) supervised learning rule to dynamically alter the weights and bias values for each neuron in the network. Back-propagation learning method is implemented through the delta rule, which is a gradient descent learning regulation for updating the weights of the artificial neurons in a single-layer perceptron. For a neuron j with activation function $g(x)$ the delta rule for weight w_{ji} is given by Equation 3.51 (Sivanandam & Deepa, 2006).

$$\Delta w_{ji} = \alpha (t_j - y_j) g(h_j) x_i \quad (3.51)$$

Where, α is a small constant, called, the learning rate, $g(x)$ is neurons starting or activation function, t_i is the target output, h_i is the sum of the product of the weight w_{ji} and x_i , y_j is the actual output, x_i is the i^{th} inputs. It holds that $h_j = \sum x_i w_{ji}$ and $y_j = g(h_i)$.

The training of a MLP-NN with back-propagation is an iterative process carried out in two phases. In the first phase, the input is propagated forward to the output unit where the error of the network is measured. The error e is usually defined as the square difference between the output and the target as illustrated in Equation 3.52 (Henseler, 1995, (Sivanandam & Deepa, 2006).

$$e = \sum_j (t_j - y_j)^2, \quad \text{for each of the output pattern} \quad (3.52)$$

In the second phase, the error is propagated backward through the network, and used for adapting the connection. On the other hand, the testing process is carried out by giving new unseen input features to a trained network and eventually obtaining the target output.

The MLP-NN architecture is particularly suitable for applications in medical imaging where the inputs and outputs are numerical and pairs of input/output vectors provide a clear basis for training in a supervised manner (Jiang et al., 2010). It is claimed to be the most common, most competent, and the most efficient model (Fausett, 1994).

There are several MLP-NN parameters that need to be optimized, in order to obtain the most suitable network structure, which can give the best testing performance of the phenomenon under study, such as, the number of hidden layers, the number of neurons in each hidden layer, the number of training cycles (epochs) and the learning rate. These parameters are discussed in more details in Chapter 4 Section 4.5. The final MLP-NN architecture used in this research to classify blasts cells is presented in Chapter 7 Section 7.5.

3.5.2 Support Vector Machine

The Support Vector Machine (SVM) introduced by Vladimir Vapnik and colleagues (Vapnik, 1995) is a powerful solution to the classification problems (Osowski et al., 2009).

The main advantage of the SVM network used as a classifier is its very good generalization ability and extremely powerful learning procedure, leading to the global minimum of the defined error function (Schölkopf & Smola, 2002; Smola, & Schölkopf, 2004; Vapnik, 1998).

SVM has been in a variety of applications, including e-mail spam classification (Drucker et al, 1999), gene data expression (Brown et al., 2000), handwriting recognition (Kaensar, 2013), and blood cell recognition (Osowski et al., 2009).

The SVM data classification is done by constructing an N-Dimensional hyper plane that classifies the data points into distinct categories. The selection of the optimal hyper plane is performed based on largest separation, or margin between the two classes, i.e. the hyper plan which maximizes its distance from the nearest data point should be chosen. This type of hyper plane is known as the Maximum-Margin Hyper plane (MMH) (Vapnik 1998; Cristianini & Shawe-Taylor, 2000). Figure 3.18 illustrates this theory in classification of data in 2-D space.

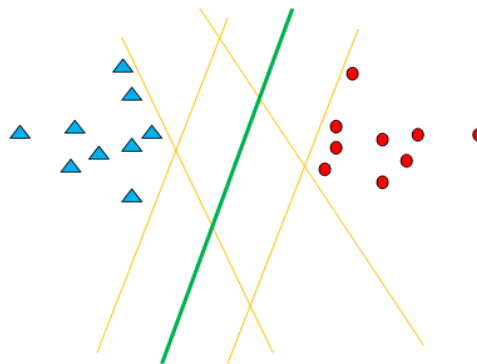


Figure 3.18: Optimal Separating Hyperplane

The red colored circles symbolize data points belonging to class 0, and the blue colored triangles symbolize samples belonging to class 1. In 2-D space, the separating margins are straight lines. In Figure 3.18, there is more than one line to separate the two classes of data, however the bold green line (hyper plane) is considered as the optimum. The optimal hyper plane is the one which leads to the maximum distance between the two set of samples. This hyper plane is considered as the MMH. The formula of a hyper plane is given by Equation 3.53 (Vapnik 1998).

$$(w^T x) + b = 0 \quad (3.53)$$

Where " x " is a point in n -dimensional space and " w " and " b " are the coefficients of a hyper plane. In a binary classification problem, the decision function is denoted by Equation 3.54 (Vapnik 1998).

$$\text{if } (w^T x_i) + b > 0 \text{ then } y_i = 1 \quad (3.54)$$

$$\text{if } (w^T x_i) + b < 0 \text{ then } y_i = -1$$

Where $i = 1, 2, \dots, l$ " l " indicates the number of classes. If $l = 2$, then, it is a binary classification problem where y_i represents the class label and x_i is the data vector.

Therefore, a SVM classifier design is equivalent to finding the MMH or finding the values of w and b . It has been shown by (Burges, 1998; Vapnik 1998; Cristianini & Shawe-Taylor, 2000), that the coefficients w and b can be obtained by minimizing the following optimization problem, which is referred to as a first SVM formulation.

$$\min_{w, b} \frac{1}{2} w^T w \quad (3.55)$$

$$\text{subject to } y_i (w^T x_i) + b \geq 1 \quad i = 1, \dots, l.$$

This type of mathematical problem is called an optimization problem, and it is effectively solved by using the method of Lagrange's multipliers (Vapnik, 2000) which provide a practical approach for finding the maxima and minima of function with subject to

constraints. A Constraint is a condition that the solution to an optimization problem must satisfy. The Lagrange multiplier method defines a function, named the Lagrange function, by taking the partial derivative of this function and equalizing it to zero, producing a solution to the optimization problem. The value of w can be obtained by following this procedure. Once w is obtained and substituting that value in the equation of a hyper plane (Equation 3.53) gives the value for b

In most of the classification cases, the data may not be separated linearly. In that case, finding the values of w and b is not straight forward (Hsu et al., 2003). For obtaining these values, we artificially add dimensions to the vectors so that they are linearly separable. This is done by mapping each point in N -dimension to higher dimensional space. Hence, a new optimization problem needs to be solved namely a standard SVM problem (Burges, 1998, Bottou & Lin, 2007).

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i \\ \text{subject to} \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, \dots, l. \end{aligned} \tag{3.56}$$

There are several differences in the second optimization problem compared to the first. Firstly x_i in the constraint term is replaced by $\phi(x_i)$ which is the mapped values of x in the higher dimension. Secondly, an error term (ξ) and penalty parameter C is introduced which corresponds to training errors. These errors are obtained when data is mapped to higher dimension, we want the error to be zero which happens only in the infinite dimension and since this is not possible, we tend to settle down by setting a large value for the parameter C and the summation of error term is set to a value which tends to zero such that the inequality in the constraint term is similar to the first SVM problem. While mapping the data into an infinite space, the dimension of w tends to infinity.

This makes the optimization problem very complicated and difficult to solve. In order to solve this problem, the coefficient w is represented as a linear combination of training vectors which is represented in Equation 3.57 (Bottou & Lin, 2007).

$$w = \sum_{i=1}^l \alpha_i y_i \phi(x_i) \quad (3.57)$$

This problem is referred to as the Dual optimization problem and α is the dual parameter which has to be found (Bottou & Lin, 2007).

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, i = 1, \dots, l \\ & y^T \alpha = 0, \\ & \text{where } Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j) \text{ and } e = [1, \dots, 1]^T \end{aligned} \quad (3.58)$$

A finite number of variables will be obtained from solving the Dual problem rather than solving the primal problem.

This is considered an advantage of the Dual problem. The primary aim in the Dual problem is to determine the value of alpha (α). Since it is derived from the primal problem, a solution to either one determines a solution to both. But the Q_{ij} term in the Dual problem has an inner product which is very difficult to obtain in higher dimensional space. So in order to make the computation easy Kernel Functions are introduced.

Kernel functions operate with the values in the regional space and obtain the results in much fewer operations than the direct inner product. This is called the kernel trick. Therefore, wherever an inner product is used in the optimization problem, it is replaced by a kernel function. Some of the well known SVM Kernel functions are Polynomial, Radial Basis Function, and Sigmoidal (Hsu et al., 2003).

3.6 Review of Computer-Based Acute Leukemia Diagnosis and Classification

Microscopic morphological examination of the PB smear is often the first step in the sequence of the leukemia diagnostic process (Mohaptra et al., 2013), despite the existence of other advanced diagnostic procedures such as flow cytometry, immunophenotyping, and cytogenetic analysis

However, this diagnostic procedure is still challenging and its accuracy is limited where the accuracy of the manual procedure is estimated to be about 60%-70% (Nasir et al., 2013). Moreover, this diagnostic procedure is inherently subjective and suffers from inter-observer variability. This issue highlights the demand for receiving an in vivo second opinion which (i) increases the diagnostic accuracy, thus saving more lives, and (ii) decreases the number of false diagnosis, hence reducing the medical and emotional cost imposed on individuals by unnecessary laboratory tests such as bone marrow biopsy.

Image processing and ML techniques have been applied by many researchers to solve this problem. A computer-based acute leukemia diagnosis provides a quantitative and objective evaluation of the blast cells versus the subjective manual procedure.

It allows for reproducible diagnosis by diminishing the inter-observer variability. It also automates the analysis, and thereby reduces the amount of repetitive and tedious tasks to be done by the hematologist or laboratory practitioner.

Due to improvements in microscopy imaging technology and image processing techniques and owing to the morphological differences between blast cells, which makes it an ingratiating problem to work on, there has been a significant increase in interest in the development of computer-based acute leukemia diagnostic system. The next sub-sections gives a critical description of the most relevant methods in the literature applied for acute leukemia diagnosis and classification based on morphological characteristics of blast cells.

3.6.1 Peripheral Blood Image Acquisition

Image acquisition is basically considered as the initial stage of any image recognition system (Mora-González et al., 2011). Acquiring image with high resolution, clarity, accuracy and proper illumination intensity is considered as a complicated task due to the complex setting of the video camera and the microscope, for medical and non-medical people (Madhlloom et al, 2012).

Up to the early nineties, most microscopic applications that needed image acquisition was with the use of an analog video camera, often simply closed circuit TV cameras with support of a frame grabber to digitize the image (Rajendran et al., 2008).

Nowadays, these analog cameras have been replaced with digital ones, called CCD (charge-coupled device) camera that is mounted on the microscope. Figure 3.19 shows an example of a modern microscope with a mounted digital camera attached to a computer.

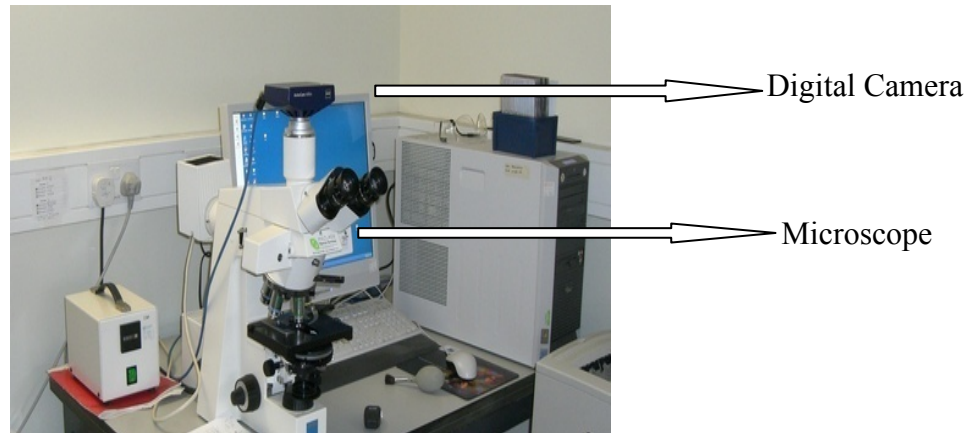


Figure 3.19: Microscope with a digital camera (Zephyris, 2007)

The basic microscopic system used in clinical laboratory is the light microscope (Estridge & Reynolds, 2011) in which visible light passes directly through the lenses and specimen. A light microscope has at least three objective lenses. In general, these lenses magnify an object 10, 40 and 100 times, respectively (Pommerville, 2009).

A proper visual analysis requires that the image of the blood smear is captured with a 100x objective lens (Díaz & Manzanera, 2011).

In order to capture high-quality blood smear images, there are several factors that need to be considered, such as well-prepared blood smear (Rodak et al., 2007), balancing illumination (Sadeghian et al., 2009), and the optimal assessment area of the blood smear where the images should be captured from (Rodak et al., 2007).

Many previous works in the literature (Osowski et al., 2004; Markiewicz et al., 2005; Markiewicz & Osowski 2006; Siroic, et al., 2007; Osowski et al., 2009; Sadeghian et al., 2009; Mohapatra et al., 2013) did not discuss the characteristics of the blood images acquisition process. This could be due to the reason that the researchers obtained the data sets of blood images from the medical center as digital images without being involved in the acquisition process. Meanwhile, other researchers discussed the details regarding the equipment and the setting used during the blood smear image acquisition process.

Table 3.3 summarizes the characteristics of blood smear acquisition process used by a number of researchers in the literature. Six different points are highlighted in the table, namely, the camera type, the microscope type, the lens objective, the image format, the color depth and the resolution.

Table 3.3: Summary of the acquisition process characteristics reported in the literature

Author	Camera Type	Microscope Type	Lens Objective	Image Format	Color Depth	Resolution
Sabino, et al., 2003,2004	Kodak DC290	Zeiss	100x	-----	24	720x480
Putzu & Ruberto, 2013, 2013a, 2013b, Labati, et al., 2011, Scotti, 2005, 2006	Cannon PowerShot G5	-----	30x-50x	JPG	24	2592 × 1944
Mohaptra & Patra, 2010, Mohaptra, et al., 2011, 2011b, 2013	-----	Carl Zeiss	100x	-----	24	1024x1024
Nee et al., 2012; Huey Nee et al., 2012	Luminera Infinity 2	Leica	100x	-----	24	-----
Madhlloom, et al., 2012, 2012b	Olympus UC30	Olympus	40x	JPG	24	2080×1544
Nasir et al., 2013	Luminera Infinity 2	Leica	40x	BMP	24	800x600

As seen in Table 3.3, most of these previous works shared a number of common acquisition characteristics such as using a light microscope and CCD camera which produce 24-bits color images. However, other characteristics such as camera type, lens objective and image resolution were different. The difference in image acquisition setup could lead to some variability in the appearances of blood images; therefore understanding acquisition conditions is tremendously important to diagnose acute leukemia. As a conclusion, better understanding of the blood image acquisition process contribute strongly to the success of the overall diagnostic system as the images will go through several processing steps in order to attain the intended results (Díaz & Manzanera, 2011).

However, if the images have not been acquired with a proper level of clarity (clarity here means, the absence of blurriness, balanced illumination, suitable microscopic magnification), then the intended results may not be achievable, even with the help of some form of image enhancement.

In this research, we have applied high-resolution PB images obtained from University Malaya Medical Center, Kuala Lumpur, Malaysia captured using Olympus UC30 camera. At a later stage of the work, we have also used images from other resources (Labati et al., 2011), whose contributions are gratefully acknowledged.

3.6.2 Blast Cells Segmentation

Image segmentation is used to detect the blast cells region, in order to separate them from the background components such as plasma and RBCs. It is considered as the most difficult stage in the acute leukemia diagnostic system (Mao-jun, et al, 2008; Rezaatofghi & Soltanian-Zadeh, 2011; Patil et al., 2012). In this section, the most-recent techniques relevant to blast cells segmentation are discussed. There are three major blast cells segmentation problems that are not completely solved by the algorithms presented in the literature; these problems are as follow:

1. Localization of the blast cells in the PB image and extracting each into a single sub-image.
2. Adjacency and overlapping between cells.
3. Adaptation to color, illumination, staining variation.

Finding solutions to these problems are significant yet challenging. The blood smear consists of various types of cells such as leukocyte, erythrocytes and platelets in addition to the abnormal cells (blasts). Identifying a blast cell and cropping it into a sub-image is a desirable step (Mohaptra et al., 2013) where each blast cell can be separately evaluated in order to classify it into either *ALL* or *AML (Problem1)*. Heterogeneous distribution of cells in PB smear is a common problem originating from the PB smear preparation procedure (Nee et al., 2012).

Since the quality of cells distribution cannot be guaranteed, the segmentation algorithm should remove the clustered cells adjacent to the cell of interest, otherwise the adjacent cell(s) and the cell of interest will be considered as one object. Hence, wrong features could be extracted (**Problem2**) (He & Liao, 2008). Blast cells segmentation algorithm should be impervious to color variation in the blood smear; the segmentation algorithm should be able to extract a complete blast cell with its internal components (nucleus and cytoplasm) from blood smears stained with different staining procedures such as Wright-Giemsa, May-Grünwald, Leishman's stain, etc. The segmentation algorithm should also be impervious to blood images acquired using different types of acquisition equipment (**Problem3**) (Markiewicz et al., 2005).

Based on the literature review, existing blast cells segmentation algorithms can be generally categorized into six different approaches according to their underlying techniques. (1) Pixel-based Threshold approaches, (2) Edge-based approaches, (3) Region-based approaches, (4) Color-Clustering approaches, (5) Mathematical Morphology approach, (6) Active-contour based approach. However, most of the proposed algorithms in the literature combined several techniques in order to extract blast cells from the image.

A two-step semi-automatic segmentation algorithm was proposed by (Sabino et al., 2003, 2004, Ushizima, et al., 2005) to segment leukocytes and one type of leukemic cell namely **CLL** (Chronic Lymphocytic Leukemia). The Green channel threshold was used to localize the nucleus and to crop a sub-image, and then the Bayesian Supervised Learning Algorithm based on the RGB color pixels was used to group the pixels into four distinct regions, namely the nucleus, the cytoplasm, the erythrocytes, and the plasma.

The segmentation was sensitive to light variations and it was often limited to a subset of images with similar staining properties. Besides that, using the nucleus bounding box to crop a single cell sub-image could produce an incomplete cell, where some part of the cytoplasm could be missed.

(Osowski et al., 2004) used the watershed algorithm to segment blast cells in bone marrow aspirate images. The watershed was applied on the original image after it was converted to gray level then to binary image and then processed with morphological closing and erosion. As stated by the (Osowski et al., 2004), this method was inaccurate in separating a complete blast cell from the background, especially at the border of the cytoplasm. This could be due to the similarities in gray level values between the erythrocytes adjacent to the cytoplasm of the blast cell. Later, the same method presented by (Osowski et al., 2004) was adopted in other works, such as the works by (Markiewicz et al., 2005; Markiewicz & Osowski 2006; Siroic, et al., 2007; Osowski,et al., 2009). Canny edge detection combined with mathematical morphology operator (i.e. dilation and erosion) was used in the work by (Scotti, 2005) to localize a single lymphoblast cell from a gray scale blood smear image. Edge detection needs some subsequent edge linking in order to close the contour of regions; however, this is a very difficult process as it is not clear which edges are of interest and which are not (Zhang et al., 2004). It has been assumed by (Scotti, 2005) that the edge size can fall in a specific range so morphological dilation can be used to link the edges. However, if the edge value is not within the assumed range, wrong cell detection can occur. Later, (Scotti, 2005) proposed to use the Otsu Threshold (Ostu, 1979) to separate the nucleus region from the cytoplasm region.

Some other researchers implemented their segmentation algorithms directly on a manually cropped sub-image that contained only a single blast cell such as the work by (Khashman & Al-Zgou 2009; Sadeghian, et al., 2009) . The purpose of this was mainly to separate the nucleus region from the cytoplasm region. In the work presented by (Khashman & Al-Zgou 2009), a pixel-based threshold with a fix range of threshold was used to separate the nucleus from the cytoplasm, while in the work by (Sadeghian, et al., 2009), the sub-image was firstly processed with canny edge detection, followed by a gradient vector flow (GVF) active contour to detect the nucleus, and finally, the Zack threshold was employed to define the cytoplasm component.

(Reta et al., 2010) introduced a segmentation algorithm to segment five types of acute leukemia blast cells, including *ALL* (L1-L2) and *AML* (M2-M3-M5). The Lab color space and the 2-D wold decomposition texture model were used. The color and texture information were modeled by using the Markov Random Field in order to obtain the regions of cell elements. A rule based classifier with respect to color and shape properties were used to separate the nucleus from the cytoplasm. A cell overlapping separation algorithm was also proposed using a linear interpolation in the polar space to provide a conical shape. However, this overlapping separation algorithm produces some edges discontinuities where edge linking is a very difficult process as it is not clear which edges are of interest and which are not (Zhang et al., 2004).

The blast cell segmentation algorithms used a fixed threshold value to extract the blast cell from the blood smear image, such as the works by (Patil, et al., 2012; Aimi Salihah, et al., 2009; Aimi Salihah, et al., 2010; Harun et al., 2010; Halim, et al, 2011; Halim, et al., 2011b). Unfortunately, the segmentation performance as a result of manually selected threshold deteriorates, since the blood smear image could be acquired from different

sources; therefore, a fixed threshold value could not suitably segment every image. All of these approaches that used a fixed threshold value were not able to extract a complete blast cell as the cytoplasm region was always missing.

In our early work (Madhloom et al., 2012), we proposed a new method that integrated color features with the morphological reconstruction to localize and isolate lymphoblast cells. Based on the results obtained from our work, this method was able to localize the complete lymphoblast cells; however, the adjacency problem was not addressed. Besides that, this approach did not further segment the nucleus region from the cytoplasm region.

(Nee et al., 2012; Huey Nee et al., 2012) adopted the same morphological reconstruction process presented in the work by (Madhloom et al., 2012). (Nee et al., 2012; Huey Nee et al., 2012) proposed a color segmentation methodology for three type of *AML* (M2-M5-M6). The saturation band of the HSV color space was used to construct a mask and marker from the original image, and then the morphological reconstruction was used to retrieve the cell. A binary image was produced from the morphological reconstruction using Otsu's global threshold (Otsu, 1979), then watershed transform was applied to the resultant binary image construction supported by a copy of the saturation gradient magnitude image to generate the ridge line. This approach failed to localize a complete blast cell, especially, blast cells with indistinct boundaries between the nucleus and the cytoplasm.

Some other researchers used an unsupervised clustering segmentation approach, such as K-mean clustering and fuzzy c-mean. Within this context, (Mohapatra & Patra, 2010; Mohapatra, et al., 2011) presented a segmentation algorithm to segment two types of blood cells, namely Lymphocyte and Lymphoblast.

The blood smear image are clustered based on Lab color space into four distinctive regions, namely erythrocyte, nucleus, cytoplasm and plasma.

Only the nucleus cluster was retained in order to extract a sub-image of the nucleus for further processing. The nuclei sub-images were cropped based on the smallest bounding box. Later, (Mohapatra et al., 2010; Mohapatra, et al., 2011b) repeated the same clustering approach; however, this time the fuzzy c-mean clustering approach was used instead of the K-mean clustering. (Mohapatra et al., 2011c) used a rough K-mean clustering to divide the image of blood smear into four different regions, where the rough K-mean clustering showed better performance than both K-mean and fuzzy c-mean clustering.

A two step clustering approach was proposed by (Mohapatra et al., 2013). In the first step, the blood smear image segmented using K-mean clustering based on RGB color space to identify the nucleus region, then a sub-image of the whole cell is cropped based on the nucleus bounding box. Later, the cropped sub-image is segmented into three different regions namely nucleus, cytoplasm and background.

Unsupervised clustering segmentation approaches separate distinct regions in the image based on the color or intensity. This method works fine when a prominent difference in the color or intensity between regions exists. However, it performs poorly when similar color is presented between different regions in the image, for instance the similarity in brightness between erythrocytes and cytoplasm (Won et al., 2005). For that reason, it has been observed by (Mohapatra & Patra, 2010; Mohapatra, et al., 2011) that the cytoplasm and the erythrocytes are classified into same cluster. This situation can also occur in clumped (overlapping) cells, where clustering approaches are not able to separate them.

A WBC identification algorithm was proposed in the work by (Putzu & Ruberto, 2013; Putzu & Ruberto, 2013a; Putzu & Ruberto, 2013b).

The algorithm included three stages, namely, leukocytes Identification, separation of grouped leukocytes, and nucleus/cytoplasm selection.

The leukocytes identification was performed using the Zack threshold applied on the Y component of the CMYK color model. This was followed by the watershed algorithm in order to separate the leukocytes clusters. Threshold-based operation using Otsu method was applied on each leukocyte sub-image. The threshold-based operation was performed on an intensity image which resulted from the combination of the green component of the RGB color space and the a component of the Lab color space. The algorithm was tested on the ALL-IDB1 database (Labati et al., 2011), which is the same as *Dataset-B* used here in this research (Please Refer to Section 4.2.2). The test was carried out with a sample of 33 PB images contained 267 cells. The algorithm was able to identify 245 cells out 267 with an average accuracy of 92%.

A new algorithm for segmentation of both healthy WBC and acute lymphoblastic leukemia blast cell using both peripheral blood and bone marrow images was proposed in the work by (Arslan et al., 2014). This algorithm modeled the WBCs based on color and shape characteristics by defining two transformations and introduced an efficient use of these transformations in a marker-controlled watershed algorithm. This was followed by a post-processing step to eliminate false white blood cells. The algorithm was tested on 650 WBCs. Although, the algorithm was able to segment a total number of 637 cells out of 650, however, in some cases, the algorithm was unable to remove some of the false positives objects that show similar coloration and shape with lymphoblast, such as, RBCs or dead cells.

Table 3.4 summarizes the characteristics of each methods presented in the literature and how they deal with the three problems highlighted earlier in this section.

From Table 3.4 it can be noted that there is no blood segmentation algorithm that completely addresses the three issues.

Table 3.4 Review of Previous Segmentation Algorithms

Author	Segmentation Method	Problem 1	Problem 2	Problem 3	Quantitative Accuracy Rate	Remarks
Sabino, et al., 2003	Pixel-Based Threshold, Bayesian Supervised Learning Algorithm	Yes	No	No	No Quantitative Evaluation	Sensible to light variations, limited to a subset of images with similar staining properties. The training set size should be increased to produce better segmentation result.
Sabino, et al., 2004						
Ushizima, et al., 2005						
Osowski, et al., 2004	Watershed	No	Yes	No	No Quantitative Evaluation	Inaccuracies were visible, especially at the border of cytoplasm.
Markiewicz, et al., 2005						
Markiewicz & Osowski 2006						
Siroic, et al., 2007						
Osowski, et al., 2009						
Scotti, 2005	Edge Detection, Mathematical Morphology, Pixel-Base Threshold	Yes	No	No	No Quantitative Evaluation	In some images the nucleus area was overestimated
Khashman & Al-Zgou, 2009	Bimodal Threshold Mathematical Morphology	No	No	No	Average accuracy was 98.33%	Sub-image cropped manually and threshold range value was fixed
Sadeghian, et al., 2009	Edge Detection active contour Zack threshold	No	No	No	92% Nucleus and 78% for Cytoplasm	Sub-image cropped manually
Aimi Salihah, et al., 2009	Pixel-Based Threshold	No	No	No	No Quantitative Evaluation	The threshold value was selected manually
Aimi Salihah, et al., 2010	Pixel-Based Threshold	No	No	No	No Quantitative Evaluation	The threshold value was selected manually
Harun, et al., 2010	Pixel-Based Threshold	No	No	No	No Quantitative Evaluation	The threshold value was selected manually
Halim, et al., 2011	Pixel-Based Threshold	No	No	No	No Quantitative Evaluation	Threshold value has a fix range Inaccuracies in cytoplasm extraction
Halim, et al., 2011b						

Reta, et al., 2010	Markov Random Field Rule Based classifier	No	Yes	No	95.87% Nucleus 95.75% Whole Cell	Fix Range thresholds for the rule-based classification. Edges discontinuities produced by overlapping segmentation algorithm.
Mohapatra & Patra, 2010	K-mean Clustering	Yes	No	Yes	No Quantitative Evaluation	Only Nucleus Extraction
Mohapatra et al., 2011						
Mohapatra et al., 2010	Fuzzy C-Mean Clustering	Yes	No	Yes	No Quantitative Evaluation	Only Nucleus Extraction
Mohapatra, et al., 2011b						
Mohapatra et al., 2011c	Rough K-mean Clustering	Yes	No	Yes	No Quantitative Evaluation	Completed cell sub-image was estimated based on the nucleus only
(Nee et al., 2012, Huey Nee et al., 2012)	Mathematical Morphology and Watershed	No	Yes	No	94.5% in the whole cell extraction.	Uncompleted localization Cytoplasm missing
Patil, et al., 2012	Pixel-Based Threshold	No	No	No	No Quantitative Evaluation	Uncompleted localization Cytoplasm missing
Madhloom, et al., 2012	Threshold and Mathematical Morphologically	Yes	No	No	90-95% in the whole cell extraction.	Fix threshold used to produce the Mask image
Mohapatra, et al., 2013	K-Mean Clustering Shadowed C-Mean Clustering	Yes	No	Yes	No Quantitative Evaluation	Only Lymphocyte and Lymphoblast are considered in the experiment
Khot, et al., 2013	Mathematical Morphologically	No	No	No	No Quantitative Evaluation	Uncompleted localization Cytoplasm missing
Putzu & Ruberto, 2013	Color Transformation Threshold-Based operation Watershed Transform	Yes	Yes	No	245 cells out of 267 correctly identified with an average accuracy of 92%.	Only Leukocyte and Lymphoblast are considered in the experiment. Generates incorrect results with the Presence of holes in the whole leukocyte.
Putzu & Ruberto, 2013a						
Putzu & Ruberto, 2013b						

Arslan et al., 2014	Color Transformation and marker- controlled watershed	Yes	Yes	No	637 cells out of 650 detected correctly	Only Leukocyte and Lymphoblast are considered in the experiment. Unable to remove some of the false positives objects that show similar coloration and shape with lymphoblast
------------------------	---	-----	-----	----	--	--

Many of the previous researches did not focus on finding a proper algorithm to localize the blast cells for the purpose of sub-imaging. Besides that, the problem of overlapping cells did not get enough attention, where most of the methods were based on the assumption that the cells are well spread and are well distributed in the blood smear. Also, the problems of color, light and appearance variations of the blood smear image were not fully discussed. Furthermore, most of the previous blast cells segmentation algorithms were tested on a single dataset acquired using a particular type of camera and microscope without validating the algorithm on a different dataset of images acquired from different sources to test the robustness of the algorithm.

3.6.3 Feature Extraction, Selection and Classification

Feature extraction is the process of extracting certain characteristic attributes and generating a set of meaningful descriptors from an image. The purpose of the feature extraction stage is to extract various features from a given blood image which best characterizes a given blood cell.

This section discusses the most related publications in the literature from the perspective of the three different processes, namely feature extraction, feature selection and classification. The reason why all this processes are discussed together is because the performance of feature extraction and selection cannot be tested alone without considering the classification accuracy.

We only included the previous researches that reported these three stages together. As discussed earlier, the research aims to classify acute leukemia blasts into either *ALL* or *AML* using peripheral blood smear images. To our knowledge, this is the first attempt to classify acute leukemia blast cells based on peripheral blood smear images. Most of the works in the literature are not very relevant to the work introduced in this thesis. Some of the researchers tried to distinguish between acute leukemia blast cells and healthy WBCs such as the works by (Mohaptra et al., 2010, 2011, 2013; Madhukar et al, 2012; Madhlloom et al., 2012; Kizrak & Ozena, 2012; Halim, et al., 2011; Nasir et al, 2011). While other established an approach to classify one of the acute leukemia types (i.e. *AML*) into their sub-types such as the work by (Osowski et al., 2004; Markiewicz et al., 2005; Scotti, 2005; Markiewicz & Osowski, 2006; Siroic et al., 2007; Osowski et al., 2009; Ismail et al., 2010). There are very few attempts reported in the literature on acute leukemia classification. However, in all these attempts, BM smear images were used such as the work by (Nasir et al., 2013; Supardi et al., 2012; Harun et al., 2011; Reta et al., 2010). Although, all of the proposed approaches have been carried out using images acquired from bone marrow. However, in a study carried out by (Weinkanff et al., 1999) it was found that there are no differences in morphological features, cytochemistry or immunophenotyping between blasts cells in PB and BM sample. Moreover, examination of stained PB smears by microscopy remains the main way of leukemia diagnosis (Angulo., et al, 2006). Using PB smear in leukemia diagnosis has clinical, cost and emotional benefits where it is much easier to draw blood from the vein than from the bone marrow, particularly, when dealing with a very ill or very young patient. In some clinical situation, the laboratory technologist is often ask to render a complete diagnostic and prognostic work up of leukemia on a PB sample due to poor specimen quality or blast yield in bone marrow sample.

Clinically, the reason behind using PB smears instead of BM is not only the desire to spare the patient an invasive and painful procedure, but also an intuitive assumption that the PB and BM blast cells in the same patient at a given point of time are identical (Almarzooqi et al., 2011).

Table 3.5 summarizes the extracted features used before for the purpose of blast cell recognition.

(Sabino et al., 2004) established an approach to classify normal leucocytes and *CLL*. A combination of 62 features was extracted from each cell nucleus and cytoplasm including simple shape features such as area, perimeter, etc., texture (5 GLCM) and color (first-order histogram features based on RGB color space). The 12 best features were selected using wrapper techniques, namely, the Sequential Forward Selection (SFS). The method was tested on 718 sample using Naïve Bayes classifiers and the accuracy reported was 89.07%. Later, (Ushizima et al., 2005) repeated the same procedure as in the work by (Sabino et al., 2004). However, the data was classified with SVM. The obtained result showed that SVM outperformed Naïve Bayes at 95.14% average accuracy.

A classification approach based on SVM used for the distinguishing between different myelogenous blast cells in BM images, taking into consideration different maturation levels of the myelogenous blast cells was proposed in the work by (Osowski et al., 2004; Markiewicz et al., 2005; Markiewicz & Osowski 2006; Siroic et al., 2007; Osowski et al., 2009). Various features were extracted from each blast cell, such as geometrical, texture and color statistics based on the RGB color histogram. The main focus was on comparing several feature selection approaches. (Osowski et al., 2004), evaluated two filter approaches (correlation analysis, mean and variance measures), to select the best 70 features from a set of 99 features in order to distinguish 12 classes of myelogenous blast cells.

(Markiewicz et al., 2005) used a wrapper of linear support vector machine (SVM) for selecting features from a vector of 87 dimensions, in order to distinguish 10 classes of myelogenous blast. Later, Markiewicz et al., (2006) compared two filter approaches (correlation analysis, mean and variance measures) and one wrapper of linear support vector machine (SVM) for selecting features from a vector of 164 dimensions. The best approach was the correlation between the feature and the class as well as the linear SVM ranking. A feature selection approach based on genetic algorithm was proposed in the works by (Siroic et al., 2007; Osowski et al., 2009). In this approach, each set of possible features was represented as chromosomes. Genetic operators such as mutation and crossover were applied in order to find the best solution(s) according to a fitness function which was defined as the classification error on the validation data set. The reported results showed that, features selected using this approach, obtained a better classification performance compared to wrapper feature selection based on the linear SVM.

Scotti (2005) evaluated three different classifiers, namely the K-nearest neighbors (KNN), the feed-forward neural network (FFNN) and the Naïve Bayes for the classification of **ALL**. 23 simple shape features were extracted from each cell such as, area, perimeter, circularity, etc., and the mean and the standard deviation of the gray level image as a texture features. The best three features were selected using a wrapper technique, namely the sequential forward selection. The reported results showed that, the best classification performance was obtained using the FFNN.

Reta et al., (2010) presented a methodology to classify acute leukemia blast cells into **ALL** and **AML** based on blast cell morphology. Five different classifiers (KNN, Random Forest, Simple Logistic, Sequential Minimal Optimization, and Random Committee) available in Weka data mining software (Hall et al., 2009) were evaluated based on 27 features using

simple shape feature, first-order statistics based on gray level, and RGB color histogram. Among all the five classifiers, the best classification accuracy of 92.20% was obtained using the Sequential Minimal Optimization.

The SVM was applied to distinguish between *ALL* and healthy Lymphocyte using PB images in the works by (Mohapatra & Patra, 2010; Mohapatra et al., 2010; Mohapatra et al 2011b). Various features were extracted from the cell nucleus such as fractal dimension, contour signature, simple shape features, first-order statistics, GLCM and mean color value based on RGB and HSV color spaces. The SVM showed a high classification performance with over 90 % true positive accuracy. Later, (Mohapatra et al., 2013) applied an ensemble classification approach evaluated with k-fold validation which gave 94.73% accuracy.

Table 3.5: Review of previous feature extraction, selection and classification methods with their reported results

- NS=Not Specified

Author(s)	Blood Cells Types	No. of Features	Features Extraction Techniques			Feature Selection	Classification Technique	Reported Result
			Shape	Texture	Color			
Sabino et al 2004	PB (5 normal Cell and <i>CLL</i>)	62	√	√	√	Sequential Forward Selection (12)	Naïve Bayes	89.07%
Ushizima, et al., 2005			(12)	(15)	(36)		SVM	95.14%
Osowski., S et al ,2004	BM <i>AML</i> (12)	99	√ (11)	√ (52)	√ (36)	Correlation analysis Mean-variance analysis (70)	SVM	88.4% on testing data
Markiewicz et al., 2005	BM <i>AML</i> 17 Types	87	√ (19)	√ (44)	√ (24)	30 of the best features ranked by the SVM	SVM	81.29%
Scotti F, 2005	PB <i>ALL</i> vs Lymphocyte (260)	23	√ (21)	√ (2)	× (0)	Sequential Forward Selection (3)	Linear Bayes KNN FF-NN	10-fold CV 99.96% 99.97% 99.98%
Markiewicz and Osowski 2006	BM <i>AML</i> 10 Types (1850)	164	√ (35)	√ (105)	√ (24)	analysis of variance and means Correlation Analysis, Linear SVM	SVM	>80%
Siroic et al 2007	BM <i>AML</i> 9 Types	158	√ (NS)	√ (NS)	√ (NS)	Genetic Algorithm SVM Ranking	SVM	GA 86.84% SVM ranking 80.9%
Osowski., S et al ,2009	BM <i>AML</i> 11 Types (1717)	164	√ (31)	√ (106)	√ (27)	Genetic Algorithm SVM Ranking	SVM	83.2% GA 77.5% SVM ranking 72.5% with no ranking
Reta et al 2010	BM <i>ALL</i> (L1,L2) <i>AML</i> (M2,M3,M5) 633 Sample	27	√ (17)	√ (5)	√ (5)	No Feature Selection	SMO SL K-NN RF RC	92.20% with SMO
Mohapatra and Patra 2010	PB <i>ALL</i> vs Lymphocyte 108	17	√ (8)	√ (7)	√ (2)	No Feature Selection	SVM	95%
Mohapatra et al., 2010	PB <i>ALL</i> vs Lymphocyte 108	14	√ (8)	√ (4)	√ (2)	No Feature Selection	SVM	95%

Mohapatra et al 2011b	PB <i>ALL</i> vs Lymphocyte 108	15	√ (9)	√ (4)	√ (2)	No Feature Selection	SVM	93%
Harun, N. H et al 2011	BM <i>ALL</i> <i>AML</i>	6	√ (6)	× (0)	× (0)	No Feature Selection	Hybrid Multi- Layer Perceptron Trained with modifier RBF	97.4%
Nasir et al, 2011	BM acute leukemia (<i>ALL</i> vs <i>AML</i>)	32	√ (24)	× (0)	√ (8)	No Feature Selection	Neural Network Trained with (LM and BR)	LM 94.39 BR 94.51
Halim, N. H. A, et al, 2011	BM WBC vs acute leukemia	32	√ (24)	× (0)	√ (8)	No Feature Selection	Neural Network Trained with (SCG and FAM)	SCG 94.5 FAM 90.27
Madhukar et al, 2012b	PB <i>AML</i> vs WBCs (50)	Not Specified	√ (NS)	√ (NS)	× (0)	No Feature Selection	SVM	93.5%
Kizrak M.A and Ozena F., 2012	PB <i>ALL</i> vs WBCs	7	× (0)	√ (NS)	× (0)	No Feature Selection	Kernel Ridge Regression	96.43%
Supardi et al, 2012	<i>ALL</i> vs <i>AML</i> 1500 samples 750 <i>ALL</i> and 750 <i>AML</i>	12	√ (NS)	× (0)	√ (NS)	No Feature Selection	KNN, k=4 Cosine distance	86%
Madhloom et al., 2012	PB <i>ALL</i> vs Lymphocyte (260) cells	30	√ (15)	√ (15)	× (0)	Fisher's Discriminati on Ratio (7) Exhaustive search (3)	K-NN	92.5%
Nasir et al., 2013	BM <i>ALL</i> vs <i>AML</i> 500 200 <i>ALL</i> and 300 <i>AML</i> 40x 1683 cells	42	√ (30)	× (0)	√ (12)	No Feature Selection	MLP_BR MLP_LM SFAM	95.70% 95.55% 92.43%
Mohapatra, et al., 2013	PB <i>ALL</i> vs Normal Lymphocyte s	44	√ (34)	√ (15)	√ (12)	T-Test	NB KNN MLP RBFN SVM Ensemble	k-fold cross- validation 94.73

Summary

This chapter discussed the background literature on the digital Image processing, including, color spaces, and image segmentation. An extensive discussion regarding feature extraction was provided, given that the work presented here exploits feature extraction for the identification of acute leukemia blast cells. This chapter provided an introduction to a number of other concepts relevant to this thesis. This includes discussion of feature selection and classification techniques including, Artificial Neural Network and Support Vector Machine. We also looked into existing solutions of acute leukemia diagnostic systems which have been previously developed in order to assist and facilitate hematologists for the accurate diagnosis of the disease.

CHAPTER 4

RESEARCH METHODOLOGY

4.1 Introduction

The proposed acute leukemia diagnostic methodology contains several phases with primary emphasis on detecting, segmenting, and then classifying acute leukemia blast cells. The PB images are first obtained as inputs to the diagnostic process and then analyzed through the proposed acute leukemia diagnostic phases. The phases that form the acute leukemia diagnostic methodology are shown in Figure 4.1 (a-c). The first step is image acquisition, followed by the image processing phase. Within the image processing phase, the ROI is first segmented, this is followed by analysis and classification of the blast cells. Lastly, the performance of the whole process is evaluated. In order to model the implementation and evaluation of the proposed methodology, a computer-aided diagnosis for acute leukemia system (CAD-AL) is developed. The processes embedded within the research are discussed in the following sections.

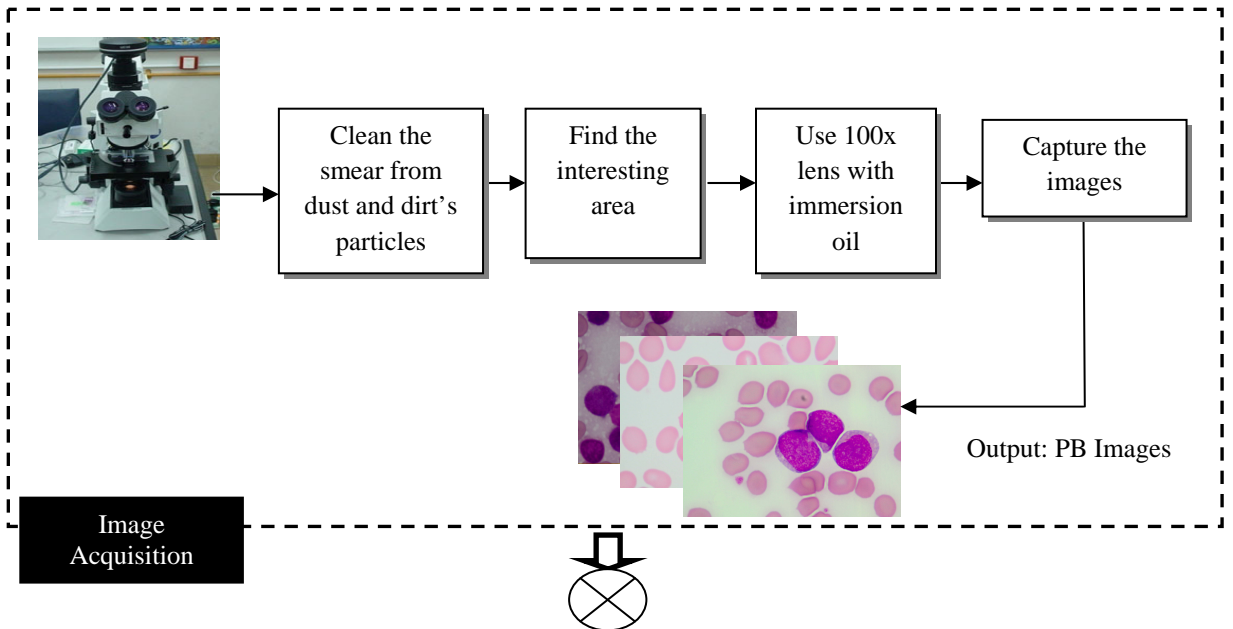


Figure 4.1 (a): The Proposed Acute Leukemia Diagnostic Methodology Phases (Image Acquisition)

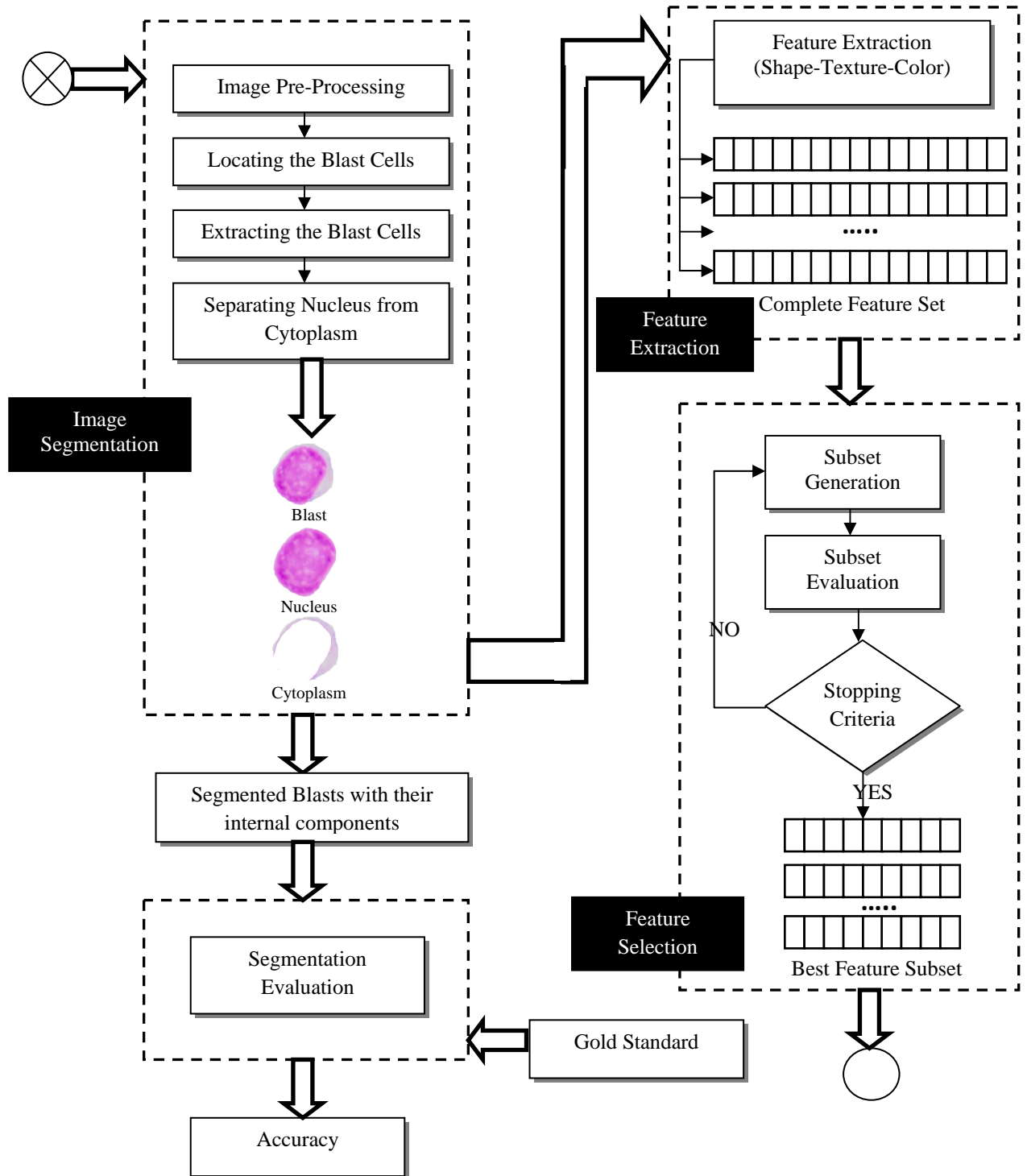


Figure 4.1(b): The Proposed Acute Leukemia Diagnostic Methodology Phases (Segmentation, Feature Extraction, and Selection)

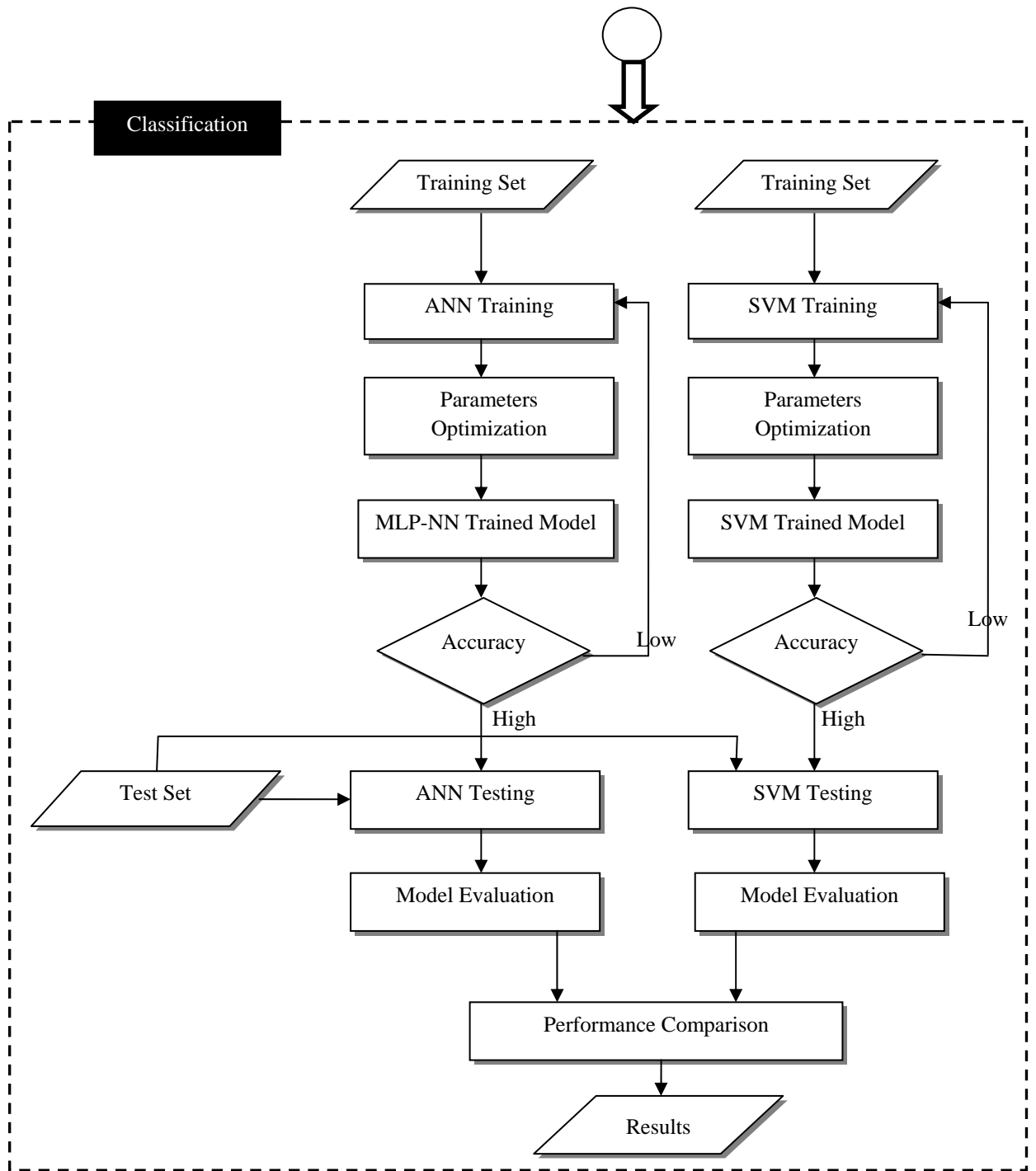


Figure 4.1(c): The Proposed Acute Leukemia Diagnostic Methodology Phases (Blast Cells Classification)

4.2 Data Acquisition

The initial stage of the proposed CAD-AL system is the PB images acquisition. The main focus of this stage is to acquire a dataset of images with high resolution, clarity, accuracy and proper brightness. This process is considered a complicated task for both medical and non-medical people due to the complex setting of the video camera and the microscope. In this research, two datasets were used, namely, *Dataset-A* and *Dataset-B*.

Dataset-A was obtained from local Malaysian patients where the image acquisition process took some time depending on the availability of the PB smears. However, in the later stages of the work, and for the purpose of validating the proposed segmentation algorithm on foreign data, we obtained images from other sources (*Dataset-B*) (Labati, et al.,2011) , whose contributions are gratefully appreciated. It is worth mentioning that, the collection of images from distinctive sources has been advantageous, as testing the performance of blast cells extraction with images collected from various sources and obtaining highly accurate results, is regarded as an indication of the robustness and reliability of the proposed approach in the real-world application (Díaz & Manzanera, 2011). We are, for this reason thankful for the contribution of Labati, et al., for the provision of *Dataset-B*

The most common type of microscope used in laboratory hematology is the light microscope (Stevens, 1997). A light microscope usually has at least three objective lenses namely the low power, the high power and the oil immersion lenses. In general, these lenses magnify object 10, 40 and 100 times, respectively (Pommerville, 2009). A proper visual analysis of blood smear requires the images to be captured with a 100x objective lens (Díaz & Manzanera, 2011). Besides that, there are other factors that can affect the acquisition of clear blood smear images such as poorly prepared stained blood smear, imbalance illumination (Sadegian et al.,2009) and choosing the optimal assessment area (Angulo & Flandrin, 2003). These factors, affect the performance of the subsequent stages.

4.2.1 Dataset-A

Dataset-A was acquired from the University of Malaya Medical Center (UMMC) located in Kuala Lumpur, Malaysia. It comprises of PB images taken from local patients treated at the Department of Pediatric Oncology at the (UMMC) Faculty of Medicine, University of Malaya. Before capturing the images, each smear was obtained from patients who have been diagnosed with acute leukemia (lymphoid or myeloid). In each case we had the result of the morphological examination available together with the acute leukemia sub-types. The process was performed with the help of the domain expert, who assisted us in choosing samples to be digitized, based on her experience. The acquisition characteristics of **Dataset-A** are summarized in Table 4.1.

Table 4.1: Acquisition Characteristics of **Dataset-A**

Image Acquisition Characteristics	
Camera:	Olympus UC30
Microscope :	Olympus CX31
Magnification of the microscope:	1000x
Image format:	JPG
Color:	24-bits RGB color
Resolution:	2080×1544
Staining Method:	MGG

The PB smears were then digitized using the Olympus UC 30 digital camera mounted on the Olympus CX31 light microscope, where the camera was connected to a PC using a Firewire cable.

The Olympus UC 30 camera is a 3.2 mega pixels digital color camera with a charge coupled device (CCD) chip. The live frame rate is 7.0 frames per second at 2080×1544, which is the resolution used for the images in this research. All the PB smears had been stained with May– Grünwald–Giemsa (MGG) stains. This staining method contains EOSIN-METHYLENE blue (May-Grunewald) & AZURE-EOSIN-METHYLENE blue (Giemsa) giving a dark blue-purple leukocyte nuclei, light cytoplasm and blue or light-orange cytoplasmic granules appearances.

Figure 4.2 shows the equipment used to capture the images for this research, where the camera and the microscope were assembled together and attached to an LCD monitor. Figure 4.3 and 4.4 show the Olympus UC 30 camera and the Olympus CX31 light microscope, respectively.



Figure 4.2: Equipment Used for *Dataset-A* Image Acquisition



Figure 4.3: Olympus UC30 Digital Camera



Figure 4.4: Olympus CX31 Optical Microscope

The process that was used to obtain an image of PB smear was as follows. First, the smears were cleaned from any dust or dirt particles with a wiping tissue and then placed on the microscope specimen stage. The smear was then observed using a 10x or 40x lens to find the working area, i.e. the well spread part of the smear by navigating through the smear using the mechanical stage knobs while looking through the ocular.

Later, the 100x lens with immersion oil was used to look for cells of interest within the selected area. When the cells of interest were located, a digital image was captured. This process was repeated as necessary depending on the number of cells of interest found in each area. This process was also repeated to collect images from all different sub-types of acute leukemia (Lymphoid and Myeloid). Table 4.2 demonstrates the number of images captured and the number of blasts cells corresponding to each acute leukemia type.

Table 4.2: Number of Images and Blast Cells (*Dataset-A*)

Acute Leukemia	<i>ALL</i>		<i>AML</i>					
Sub-type	L1-L2	L3	M1	M2	M3	M4	M5	M7
Number of Images	207	62	17	170	149	86	125	175
Number of Blast Cells	260	65	17	194	187	191	190	199
Total Number of Blast Cells	325		978					

The distribution presented in Table 4.2 shows the number of images captured during the duration of the research. The number of captured images was solely dependent on the availability of the PB smears provided by UMMC. Due to time limitation and the lack of sufficient data, two sub-types of *AML* namely M0 and M6 were not included in the research. Some sample images from *Dataset-A* are shown in Figure 4.5.

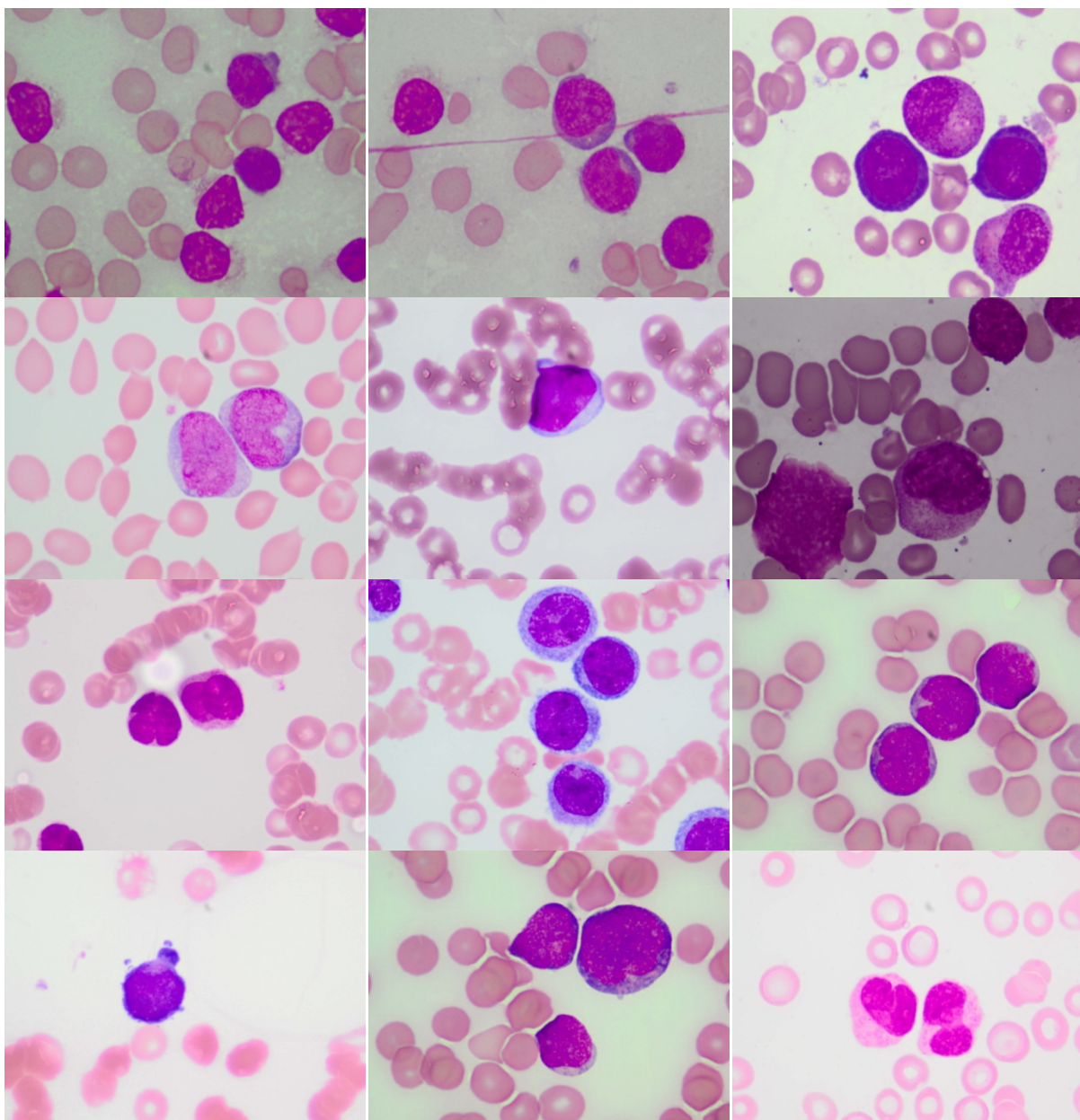


Figure 4.5: Sample Images from *Dataset-A*

Dataset-A is considered the main dataset where the segmentation, feature extraction and classification steps were performed. However, the collection of the *Dataset-A* PB samples was started after the acquisition of the *Initial-Dataset*. The *Initial-Dataset* consisted of 100 *ALL* (L1-L2) PB samples, which had similar specifications to *Dataset-A* except the magnification of the microscope used here was 400x. The *Initial-Dataset* was used in the development of the blast cell localization algorithm (*BCL*) (Please Refer Section 5.2).

After obtaining the initial results produced by the *BCL*, it was eventually discovered that for better PB smear image analysis, PB images should be captured at 1000x microscopic magnification, as the 1000x gives purer PB images and better outlines the boundaries of the blood components as well as the boundaries between the blast cell's nucleus and cytoplasm, particularly when compared to the 400x magnification.

4.2.2 Dataset-B

Dataset-B was obtained from an accredited image repository. It is composed of 108 images collected during September, 2005. All the samples of the dataset were collected by experts of the M. Tettamanti Research Center for childhood leukemia and hematological diseases, Monza, Italy (Labati et al., 2011). The acquisition characteristics of *Dataset-B* are summarized in Table 4.3.

Table 4.3: Acquisition Characteristics of *Dataset-B*

Image Acquisition Characteristics	
Camera:	Canon PowerShot G5
Microscope :	-----
Magnification of the microscope:	300x-500x
Image format:	JPG
Color:	24-bits RGB color
Resolution:	2592 x 1944
Staining Method:	-----

The images of this dataset were captured with an optical laboratory microscope coupled with a Canon PowerShot G5 camera. All images were in JPG format with a 24 bit color depth, and a resolution of 2592 x 1944. The images were taken with different microscopic magnification ranging from 300x to 500x. The dataset was used for the purpose of cell segmentation and classification by many researchers such as (Scotti, 2005; Scotti, 2006, Madhukar, 2012; Putzu & Ruberto, 2013; 2013a, 2013b) and is available online at (<http://www.dti.unimi.it/fscotti/all/>). Figure 4.6 shows sample images taken from *Dataset-B*.

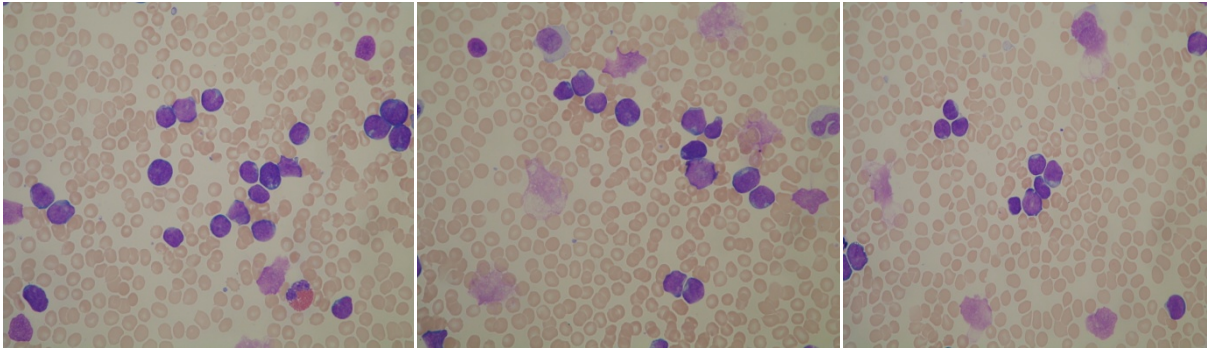


Figure 4.6: Sample Images from *Dataset-B*

4.2.3 Gold Standard

For the purpose of validating and testing the performance of the proposed segmentation algorithm, a gold standard dataset was obtained. This dataset comprised of manually segmented cells where the border of the cells of interest and their internal parts were manually drawn for all the images in *Dataset-A* and *Dataset-B* using a photo editing software.

By considering the nature of the blast cells as well as the inter-observer variability (differences in diagnosis results among different observers), each cell of interest was extracted manually from the PB smear image. This was further segmented in order to extract its internal parts i.e the nucleus and cytoplasm. This process was carried out under direct supervision of a domain expert from the UMMC. Figure 4.7 (a) shows an original sample image, while Figure 4.7 (b) shows the same image with a manual highlight of the whole cell and Figure 4.7 (c) the manual highlights of the nucleus.

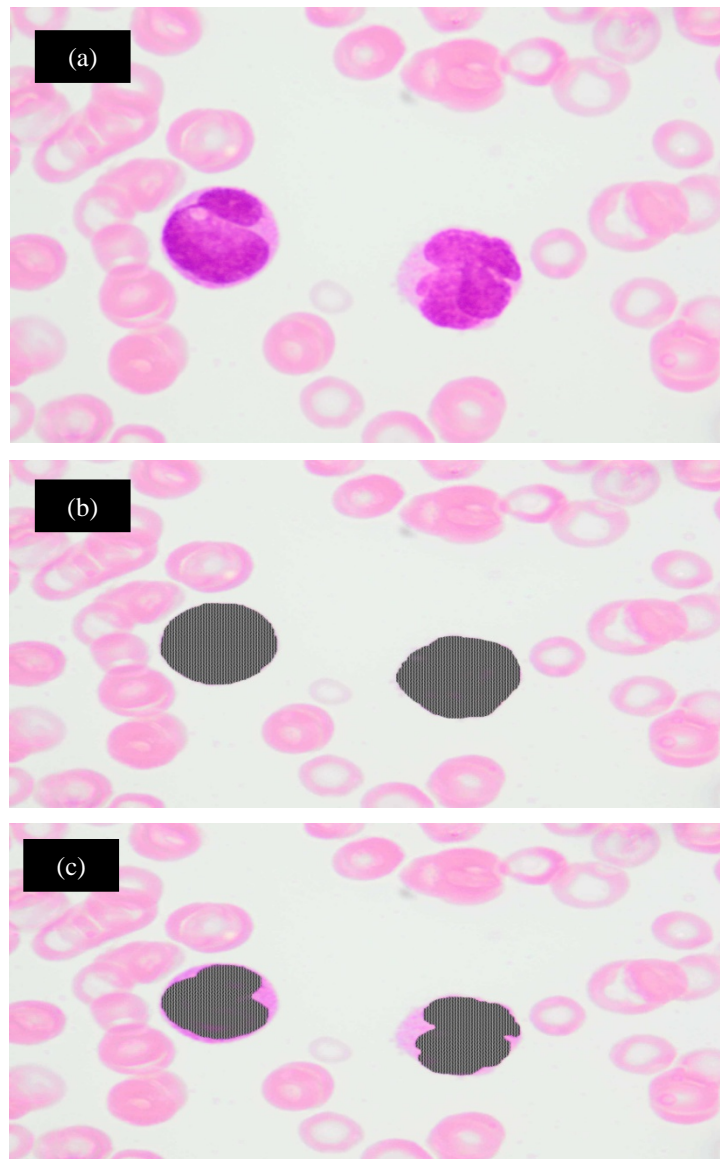


Figure 4.7: Sample of Gold Standard Images (a) original image (b) manual highlights of the blast (c) manual highlights of the nucleus

4.3 Image Segmentation

PB images are the output of the image acquisition stage; this stage is followed by the image segmentation stage. Referring to Figure 4.1 (b), the PB smear image segmentation comprises of several stages. Firstly, the images were subjected to some pre-processing steps in order to overcome the defects that may arise during the image acquisition stage such as noise reduction, stabilizing the color variation, illumination imbalance and lower uniform luminous intensity.

Therefore, pre-processing steps are needed to address the image flaws and to prepare the data for further analysis. Contrast enhancement can be used to reduce the effects of color and illumination variability and make the color of different components of the PB image more prominent. Color model transformation is also another way of highlighting the ROI while suppressing the other parts of the image such as the transformation of the RGB color into HSV or Lab. As a prerequisite step for extracting the blast cells from the obtained images, the location of each blast cell in the image should be identified, and made more noticeable than the other parts of the PB image, such as the Erythrocytes and plasma. Furthermore, each located blast cell should be separated from the other components that are in direct contact with it, such as Erythrocytes, Leucocytes or other blast cells. Through this step, the border of the blast cells can be identified, and each cell can then be cropped based on its bounding box into a sub-image that contains only a single blast cell on a white background. Sub-imaging is a desirable step since each blast cell has to be individually evaluated in order to differentiate Lymphoblast from Myeloblast and vice versa. Moreover, sub-imaging makes the process of separating the nucleus from the cytoplasm easier. Several image processing techniques namely contrast enhancement, color transformation, mathematical morphology, watershed segmentation and seeded region growing were combined together to form an automated algorithm that reduce the effects of images acquisition and result in the extraction of the cells of interest from their internal components. Section 5.2 and 5.3 discuss these proposed segmentation algorithms in details.

4.4 Feature Extraction and Selection

Given that the cells of interest have been appropriately segmented, a set of features i.e. blast cell measurements, are extracted. Acute leukemia blasts will be classified based on these features. It is widely agreed that there is no general powerful set of features that is appropriate for many different applications (Esposito & Malerba, 2001). Effective features, are solely dependent on the application and the purpose of the study. To have an optimal set of features for the problem of acute leukemia blasts classification, we need to carefully choose the features as proper feature extraction is essential for the later recognition stage (Lai et al., 2008). This is because such features should represent the unique morphological characteristics of blast cells as much as possible.

The features that domain experts use to manually recognize various types of cells are rather obvious and even straightforward for themselves. However, it is very difficult to devise equivalent measures for a computer vision application. Nevertheless, there are some apparent differences between Lymphoblast and Myeloblast. Examples of such differences include having a relatively larger nucleus and less cytoplasm in Lymphoblasts than in Myeloblasts and having a greater proportion of the area of the nucleus occupied by heterochromatin in Lymphoblasts than in Myeloblasts (Ochiai & Eguchi, 1986). However, in some cases, blast cells which belong to the same class could exhibit significant morphologic variation (Estey et al., 2007). To overcome this issue, the morphological characteristics in terms of similarities and differences between the acute leukemia types, based on the FAB classification (Table 2.5 and Table 2.6), have been studied intensively by arranging a number of meetings with a domain expert from the UMMC. Hence, a wide range of features were extracted from each blast cell including shape, texture and color features.

Shape features were used to study the geometrical differences between the blast cells in each acute leukemia type and their nuclei characteristics such as, the ratio of the blast cell area to the area of its nucleus, and other features such as perimeter, blast cell circularity, etc. Texture and color feature extraction techniques were used to address the characteristics of the nuclear chromatin texture and color characteristics. Two types of texture analysis techniques were used namely first-order statistics, represented by histogram statistics and, second-order statistics represented by GLCM.

Color features are also a very useful and distinguishing characteristic of an object. There are several color spaces that are appropriate for different applications. For this work, two different color spaces were considered. These are the RGB and the HSV.

After the full feature set had been constructed, the optimal subset of features was extracted. The philosophy behind feature selection is that not all features are useful for learning. Hence, the aim was to select a subset of the most informative or the discriminative features from the original feature set (Gu et al., 2011). Although, some researchers argue that the feature selection step is optional, this step is significant as it helps in obtaining optimal accuracy (Chu et al., 2012). On the other hand, keeping redundant or insignificant features could potentially cause confusion during the recognition process (Lai et al., 2008).

In this research, a wrapper feature selection technique namely sequential feature selection (SFS) (Kohavi & John, 1997) was used. For that reason, filter methods and embedded methods are outside the scope of this thesis.

The SFS works in an iterative manner starting with an empty feature subset, which then evaluates all possible single-feature expansions of the current subset. The feature that leads to the best score of the criterion function is added permanently. The search terminates when there is no single-feature expansion that improves the score.

Similar to Kohavi and John (1997), the improvement is defined as an accuracy enhancement of at least ε when compared to the current score.

SFS remains as one of the widely adopted supervised feature selection technique. This is because SFS is easy to implement and usually produce quite effective results (Ren et al., 2008; Marcano et al., 2010).

The pseudo code for SFS is shown below.

Purpose	1:	Selecting the best discriminative subset of features from the set of extracted features
Input	2:	The set of extracted features $F = \{f_1, f_2, f_3, \dots, f_n\}$
	3:	The used criteria function $J(\cdot)$
Output	4:	The selected subset
Procedure	5:	<i>Current</i> = $\{\phi\}$
	6:	<i>Selected</i> = <i>Null</i>
	7:	<i>For all feature</i> $f \notin \text{Current}$
	8:	<i>if</i> $J(\text{Current} \cup f) > J(\text{Selected})$
	9:	<i>Selected</i> = $(\text{Current} \cup f)$
	10:	<i>Current</i> = <i>Selected</i>
	11:	<i>else</i>
	12:	<i>break</i>
	13:	<i>return Selected</i>

SFS is still popular due to its superior results, and researchers use it in many real-world problems such as blood cells recognition (Rezatofighi et al., 2011, Piuri and Scotti, 2004) and leukemia classification (Sabino , et al., 2004, Scotti, 2005), solar power prediction (Hossain, et al., 2013), breast cancer diagnosis using digital mammography (Luo & Cheng, 2012), classification of liver tissue (Gletsos et al., 2003), prostate cancer classification and diagnosis using multispectral imagery (Bouatmane, et al., 2011) among others.

Our goal for using the feature selection process is not limited to maximize the classification accuracy, but is also to identify the most relevant features that carry a clinical significance. Such features can help in improving the subjective manual diagnosis as the specialist can objectively focus more on the most relevant features.

4.5 Classification

The next stage of the diagnostic process is passing the best subset of features onto a classifier that analyzes the quantified characteristics and determines the class of each blast cell. In this research and as shown in Figure 4.1 (c), two of the most popular classifiers are used namely the Multi-layer perceptron (MLP) and the SVM. These have proven outstanding classification performance in practice (Cristianini & Shawe-Taylor, 2000, Yao et al., 2001). Moreover, based on the literature (Ushizima, et al., 2005, Osowski., S et al ,2009, Nasir et al., 2013, Mohapatra, et al., 2013) these two classifiers are the most usable types of classification engines applied to the problem of blood cell classification in general. There are two processes involved in defining the method of classification; these include the training and the testing of the proposed classifier. During the training process, the algorithm, the parameter and feature boundaries of the classifier are optimized. The training process is very important in the classification stage; its purpose is to describe the characteristics of the classification categories. On the other hand, testing process is used to classify unseen cases based on the previously determined training parameters and feature boundaries. It usually requires the use of other independent data for testing; this is to avoid bias in the classification accuracy (Demir & Yener, 2005).

The Rapidminer application was used for the blast cells classification process. Rapidminer is a data analysis software tool which implements a set of machine learning algorithms for data mining tasks (Shafait et al., 2010).

4.6 Parameters Selection

The accuracy achieved by most learning methods strongly depends on the choice of parameters that enable the learners to adapt to the problem-specific characteristics of the training data. Thereby, parameters often experience high interdependencies (e.g. kernel parameters and regularization constant for SVMs). Hence, finding the optimal parameter configurations is critical for the results of the induction process (Cherkassky & Ma, 2004, Rocha, 2008).

In order to identify the most suitable parameter configurations, algorithms should be applied in a way that allows for an unattended optimization to the problem at hand. These algorithms must train several hypotheses with various parameter configurations by choosing the value of each parameter within a user-defined range and evaluating each configuration on a separate validation set or by a cross-validation procedure (Rocha, 2008). John (1994) investigated the possibility of applying k-fold cross-validation for the optimization of an inducer's parameters. Parameter configurations should be selected in order to optimize the adopted generalization performance measure. The parameter selection process is often called “model selection”, as each parameter configuration results in another model or hypothesis.

4.6.1 MLP-NN Parameters Optimization

In order to get the best fit MLP architecture, several parameters, that are required to be tuned or optimized, exist. These include, the number of hidden layers, the number of neurons, the activation function, the learning rate, and the training cycle (epoch) (Dheeba & Selvi, 2011). As depicted in Figure 3.17, a multi-layer feed-forward neural network consists of one input layer, one output layer and two hidden layers.

The number of input and output layers is fixed to one, and the hidden layer is of a variable size. However, expanding the number of hidden layers may result in over-fitting the data. This explains why one hidden layer is usually adequate (Günther & Fritsch, 2010). The number of neurons in the input and output layers are easily identified based on the number of input patterns and the number of output classes, respectively. A different number of hidden neurons are required to be tested for getting the best architecture. The quantity of hidden neurons can greatly affect the learning ability of the neural network (Theodoridis et al., 2010).

The activation function is used to calculate the output response of a neuron (Jayaraman et al., 2011). There are several activation functions. However, in a study conducted by (Shenouda, 2006) to compare the performance of various activation functions namely Sigmoid, Tangent, Linear and Gaussian Radial Basis Function (RBF) on 10 different public datasets (medical and non-medical), the results showed that the sigmoid activation function outperformed other activation functions.

Regarding the learning rate, it has been stated by (Abdul Kareem, 2002) that the learning rate should be small enough to guarantee smooth convergence of the model. Nonetheless, very small learning rate may lead to very slow convergence rate.

Unfortunately, there is no precise rule or mathematical definition for selecting a particular value of these parameters. Normally, the setting of the parameters is done empirically (Sun, 2011).

There are various heuristic rules and common practices for selecting the parameters (Walczak & Cerpa, 1999), but the selection process remains an art rather than a science, and varies from problem to problem (Shamsuddin et al., 2008).

4.6.2 SVM Parameters Optimization

In this work, SVM with RBF kernel was used. This kernel nonlinearly maps samples into a higher dimensional space so unlike the linear kernel, it can handle the case when the relationship between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF (Keerthi & Lin., 2003), since the linear kernel with a penalty parameter C has the same performance as the RBF kernel with some parameters (C, γ) .

In addition, the sigmoid kernel behaves like RBF for certain parameters (Lin & Lin, 2003). Another reason for choosing the RBF kernel is, the number of hyperparameters which influences the complexity of the model selection. The polynomial kernel has more hyperparameters than the RBF kernel.

Discovering appropriate values for SVM hyperparameters (C, γ) can dramatically improve the classification accuracy (Deepajothi & Selvarajan, S. 2013).

Intuitively the γ parameter defines the distance that a single training example can reach, with low values meaning ‘far’ and high values meaning ‘close’. The C parameter trades off training examples misclassification against decision surface simplicity. A low C ensures a smooth decision surface whereas a high C attempts to classify training examples correctly (Deepajothi & Selvarajan, 2013). In this research, experiments were undertaken to evaluate SVM performance through variations of the pair (C, γ)

4.7 Imbalance Data

Imbalance data means that classes are not equally represented in the training data. This issue is an intrinsic property of medical data as disease cases are rare (Mennicke et al., 2009). In our situation, the reason for the imbalance data is the lack of occurrences of **ALL** cases during the life time of the research.

Imbalance data has a serious impact on the classifiers performance (Chawla et al., 2011). Learning algorithms that do not consider class-imbalance tend to be overwhelmed by the majority class and tend to ignore the minority class. In a real world application; the ratio of minority to majority sample can be significantly low as 1:100, 1:1000 or 1:10000 (Chawla et al., 2004, Chawla et al., 2011). Misclassifying a minority class is usually more serious than misclassifying a majority class. For example, approving a fraudulent credit card application is more costly than declining a credible one (Liu et al., 2009). In our situation, as both of the samples are leukemia cases, the degree of accuracy is very important as each acute leukemia type (*ALL*, *AML*) requires different treatment pathway.

One of the solutions to tackle the problem of imbalance data is the sampling technique. This approach alters the distribution of the classes to ensure that more balanced data are obtained. This can be done either by over sampling or by down sampling and sometimes both (Batuwita & Palade, 2013).

Down sampling removes a number of observations from the majority class. It aims to attain the equal number of samples from the majority and the minority classes. This method has some drawbacks because it can eliminate some useful information. On the other hand, over sampling increases the number of samples in the minority class by replicating them and hence, reaching an equivalent number to the majority class samples.

Both down sampling and oversampling approaches have been shown to improve classifier performance over imbalanced data sets. However, it was shown by (Japkowicz & Stephen., 2002) that, oversampling is lot more useful than down sampling. The performance of oversampling algorithms was shown to improve classifier performance dramatically even for complex data sets

The Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2011) is an approach used to form new minority class examples by interpolating between several minority classes examples that lie together. It has been shown that SMOTE is very successful in dealing with imbalance data sets (Barua et al., 2011). This method used the k -nearest neighbors and it is based on the distance measure. Consequently, it may not work well on a very high dimensional datasets like gene expression datasets due to the curse of dimensionality.

In this research, the oversampling technique was adopted in order to increase the number of samples in the minority class and reduce the effect of imbalanced classes.

4.8 Evaluation Measures

4.8.1 Blast Cells Segmentation Evaluation

To achieve the fourth objective of this research, the proposed approach was tested and assessed against some metrics. First, the proposed PB image segmentation algorithm was empirically tested. In order to assess the quality of the computer generated segmentation, one manually generated segmentation, namely, the gold standard or ground truth is required in order to compare it with the computer generated one. One of the key challenges for the evaluation of automated image segmentation methods is the lack of a gold standard against which segmentation methods could be compared (Babalola et al., 2008). For that reason, most researches followed the most common method for segmentation quality assessment which is the visual inspection made by the domain expert. The disadvantage of such method is that visual or qualitative evaluation is inherently subjective (hence their namesake). Subjective evaluation scores may vary significantly from one human evaluator to another, because each evaluator has his/her own distinct standards for assessing the quality of a segmented image.

Furthermore, the results of the evaluation can depend upon the order in which evaluators observe them. Subjective evaluation is a very tedious and time-consuming process, and intrinsically, such methods cannot be used in a real-time system to rank the performance of segmentation algorithms or even different parameterizations of a single segmentation algorithm (Zhang et al., 2008).

In this research, well-known objective image segmentation evaluation protocols, that ranked high in Zhang's survey (Zhang, 1996), were employed namely the Relative Ultimate Measurement Accuracy (*RUMA*) and the Misclassification Error (*ME*).

The *RUMA* protocol uses the spatial feature disparity between the gold standard and the computer generated segmented image, where the feature for example can be the size of the area, the length of perimeter, etc. This protocol is very effective, since the main intention is to extract a completed blast cell from a PB smear image. This metric can be used to accurately compare the extracted cell to the gold standard image.

The *RUMA* protocol is defined as follows. Let GS_f represent the feature value extracted from the gold standard image and S_f represent the feature value measured from the Computer-based segmented image, then the *RUMA* protocol is defined as (Zhang, 1996):

$$RUMA_f = \frac{|GS_f - S_f|}{GS_f} \times 100 \quad (4.1)$$

On the other hand, the *ME* gives the percentage of background pixels wrongly assigned to foreground and conversly, the number of foreground pixels wrongly assigned to the background and which has often been used as a performance metric for image segmentation (Zhang et al., 2008). For the two class segmentation as in our case which consist of the foreground and the background, *ME* is defined as in Equation 4.2 (Shaikh et al., 2011):

$$ME = 1 - \frac{|B_g \cap B_s| + |F_g \cap F_s|}{|B_g| + |F_g|} \quad (4.2)$$

Where B_g is the background of the ground truth image, B_s is the background of the segmented image, F_g is the foreground of the ground truth image, F_s is the foreground of the segmented image, while the denominator $|B_g| + |F_g|$ represents the number of pixels in the ground truth image.

4.8.2 Classification Performance Measures

Generally speaking, classification problems are evaluated using a matrix known as the confusion matrix. The confusion matrix contains the numbers of correctly and incorrectly classified samples for each class. Table 4.4 shows a confusion matrix containing two classes (Positive and Negative) which is suitable for a binary classification problem (Costa et al., 2007).

Table 4.4: Confusion Matrix

True Class	Predicted Class	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

In general, True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) are computed for all outputs in the “classifier testing set” through testing the classifier. FP is the proportion of non-targeting output classified incorrectly as targeting output. On the other hand, TP is the proportion of targeting output classified correctly as targeting output. TN is the proportion of non-targeting output classified correctly as non-targeting output while FN is the proportion of targeting output classified incorrectly as non-targeting output. TN and FN are the complements of FP and TP, respectively.

It should be noted that when the results of TP and TN are 100%, an ideal ratio is obtained, while the ideal ratio of FN and FP will be 0%. Five metrics are used to evaluate the performance of the targeting output classifier. The first metric is called sensitivity which describes the rate of the TP (Sovierzosk et al., 2010), it is the proportion of examples belonging to the positive class which were correctly predicted as positive (Costa et al., 2007). Sensitivity is defined as in Equation 4.3

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (4.3)$$

The second metric is called specificity, this refers to the rate of TN (Sovierzosk et al., 2010). A clear figure of the percentage of negative examples correctly predicted as negative can be obtained using this metric (Costa et al., 2007). Specificity is defined as in Equation 4.4

$$Specificity = \frac{TN}{(TN + FP)} \quad (4.4)$$

The third metric, known as the geometric mean (G-mean), was proposed by (Kubat & Matwin, 1997) and has been used by several researchers for evaluating the classifiers on imbalance data sets (Bekkar et al., 2013). G-mean indicates the balance between the classification performances of the majority and the minority classes. This metric takes into account both specificity and sensitivity, as shown in Equation 4.5

$$G - mean = \sqrt{\text{specificity} \times \text{sensitivity}} \quad (4.5)$$

The specificity, sensitivity and G-mean are used when the performance of both classes are of concern and are expected to be high concurrently (Nguyen et al., 2009).

The last two metrics are precision and accuracy. Precision shows how consistent the results can be reproduced, while accuracy reflects the overall correctness of the classifier. Precision and accuracy are defined using Equation 4.6 and 4.7, respectively.

$$Precision = \frac{TP}{(TP + FP)} \quad (4.6)$$

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)} \quad (4.7)$$

Another very useful evaluation measure used for making a decision about the optimum models is the ROC (Receiver Operating Characteristics), which relates sensitivity and specificity (Ballabio & Todeschini., 2009). The area under the ROC curve (AUC) is often taken as a comprehensive scalar indicator of the model performance. An AUC value of 0.5 suggests a poorly performing model with random sample assignment, while an AUC value of 1 indicates a model of maximum efficiency (Truchon & Bayly, 2007).

4.9 Summary

This chapter fully describes the design process and the reasoning behind the requirements of each design step. In Section 4.2, details of the acquisition process were illustrated, with full description of the acquisition equipment and the number of acquired samples. Later, in Section 4.3 the requirements of the segmentation phase were identified and justified. This was followed by Section 4.4, which explained the features that should be extracted from each blast cell and the process of selecting the optimal set of features. The requirements analysis for obtaining the best classification model was discussed in details in Section 4.5 and 4.6, respectively. Section 4.7 elaborated on the problem of imbalance data and provided solutions for this problem. Finally, the evaluation criteria for both segmentation and classification were discussed in Section 4.8.

CHAPTER 5

PERIPHERAL BLOOD SMEAR IMAGE SEGMENTATION

5.1 Introduction

Blast cell segmentation is a fundamental step in the development of the CAD-AL. It involves separating the blast cells from the background components and further segmenting the blast into the nucleus and cytoplasm. The accuracy of detecting the blast cell is essential for accurate implementation of the subsequent stages of the system, i.e., feature extraction and classification, since diagnostic features such as area, perimeter, major axis length, etc., are strongly dependent on the segmentation result. In addition, the nucleus, which is part of the detected blast cell, reveals information about chromatin density patterns and color. Therefore, excluding any portion of the blast cell may result in significant information being overlooked, and increase the possibility of a false diagnosis. Extracting blast cells from PB smear image is challenging (Mao-jun, et al., 2008; Rezaatofghi & Soltanian-Zadeh, 2011; Patil, et al., 2012) due to a number of reasons: (1) The blood smear image contains a mixture of many blood components. (2) The adjacency of cells, e.g. Erythrocytes adjacent to blast cells, blast cells adjacent to each other. (3) The color variation which can result in increased fuzziness around the borders of the blast cells and the surrounding objects and between the nucleus and the cytoplasm. In Section 5.2 of this chapter, a novel blast cell extraction method, based on color space transformation and mathematical morphology analysis, is discussed. The method first determines the most effective and discriminative color channels in PB images that can highlight the cells of interest, followed by morphological operations to extract the blast cells. In Section 5.3, the proposed blast cell extraction method is improved by proposing new modules, namely, erythrocytes removal, touching cells segregation and nucleus/cytoplasm separation.

5.2 Blast Cells Localization

This section proposes an automatic segmentation algorithm namely; Blast Cell Localization (*BCL*), based on color space transformation and mathematical morphology operations. The ultimate aim of this algorithm is to localize the blast cells and extract them into sub-images, where each sub-image contains a single blast cell placed on a white background. The localization and sub-imaging process makes the process of nucleus and cytoplasm separation simpler and more efficient, since each sub-image will contain only two regions (nucleus and cytoplasm). This step can be done through two stages (1) Remove most of the blood components such as Erythrocytes, Platelets etc. and retain only the blast cells on a white background. (2) Extract and place each blast cell into a small sub-image to be used for the subsequent steps, i.e., nucleus and cytoplasm separation. The two stages are shown in Figure 5.1.

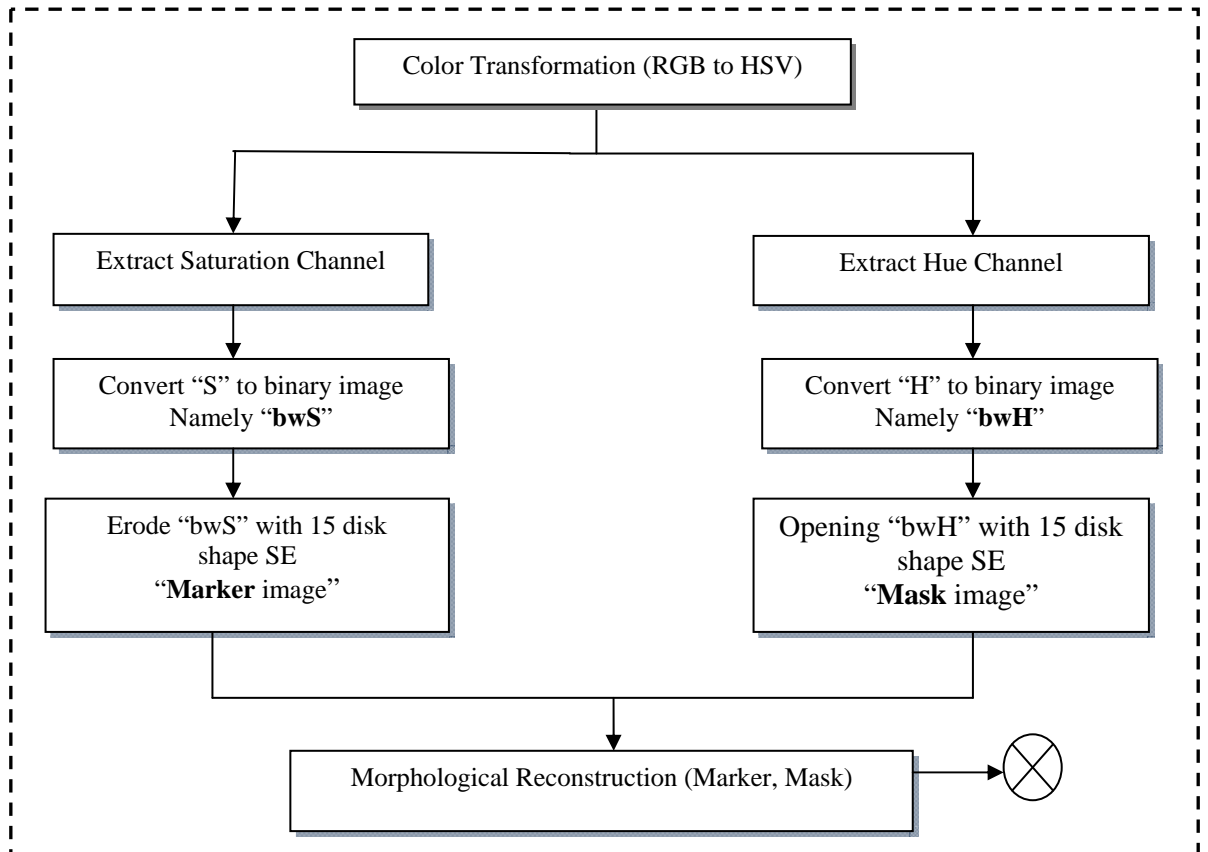


Figure 5.1: Blast Cell Extraction Flowchart (Part A)

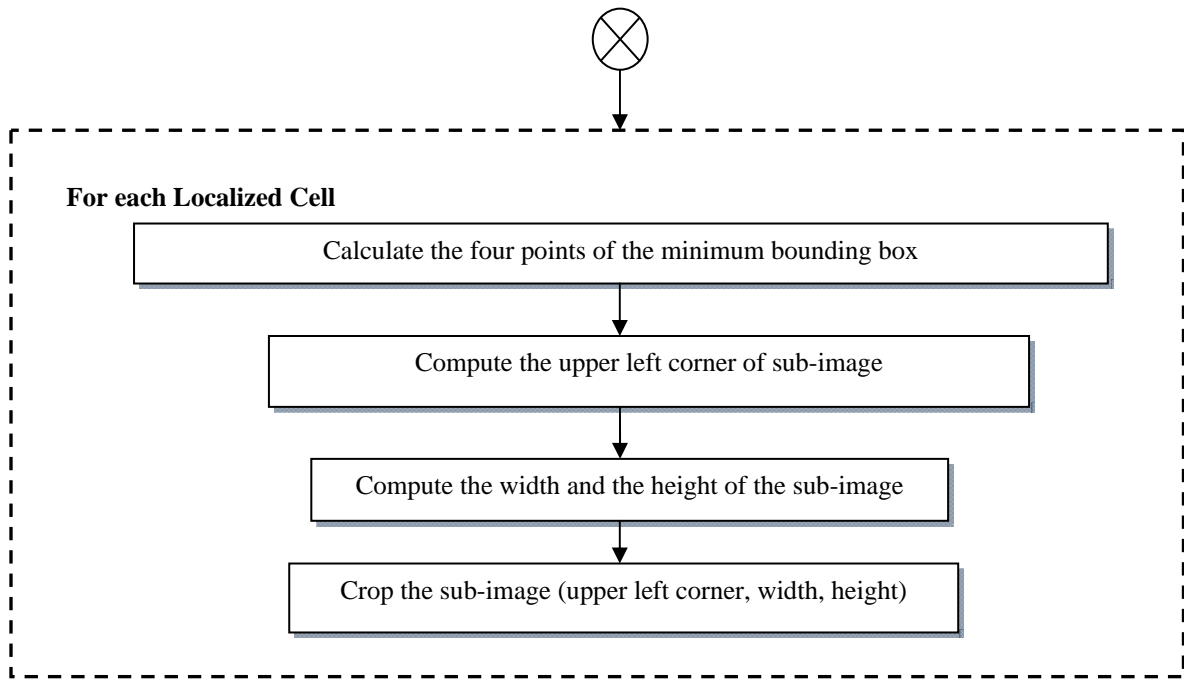


Figure 5.1: Blast Cell Extraction Flowchart (Part B)

5.2.1 Color Transformation

All the digital PB smear images acquired for this research are in the RGB color space. As a preprocessing step, the RGB images were converted to another color space namely the HSV color space. Although the HSV color space strongly represents the colors in a similar way as how the human eye senses color, this was not the major reason behind selecting it here. The main reason was that the HSV color space highlights the cell of interest and makes it more prominent than the other components; hence, this makes the localization process more efficient. Figure 5.2 (a) shows an original sample of a PB image in RGB color space while Figure 5.2 (b-d) shows its HSV counterpart, Saturation band and Hue band respectively.

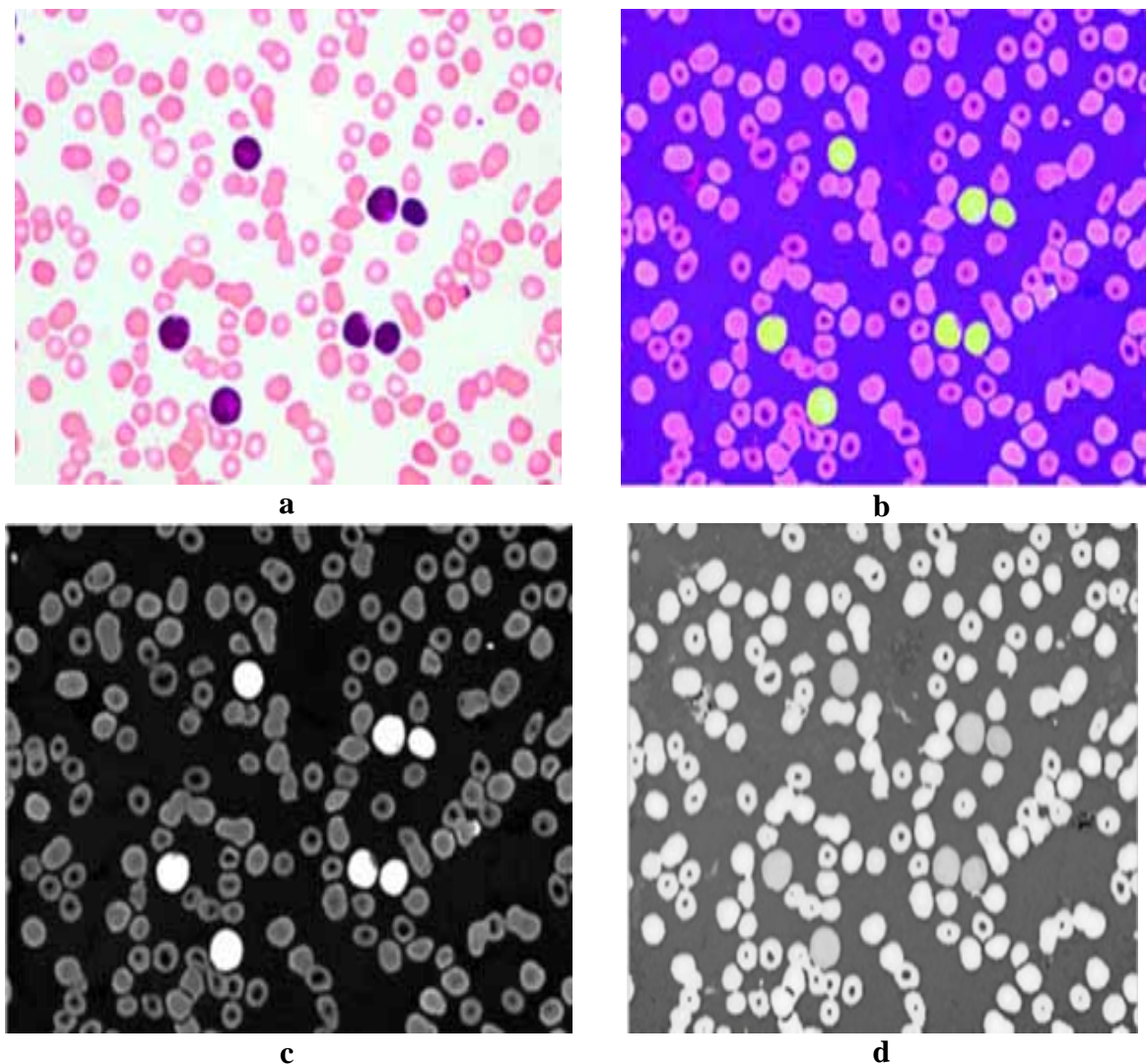


Figure 5.2: PB Image Color Transformation (a) Original RGB Image, (b) Original HSV Image, (c) Saturation Band, (d) Hue Band

The Hue and the Saturation channels were extracted from the HSV image as shown in Figure 5.2 (c) and (d), respectively. These two channels were pre-processed to be used as a Mask and Marker for the morphological reconstruction process. The hue band is considered as the shade of the color while the saturation band can be described as the color purity or the degree of white in the color tone. More hue means more saturation and a deeper or more intense version of the color.

The Saturation image shows the nucleus of the cell of interest as the brightest object in the image due to the blood staining which highlights the nucleus with a dark purple color.

This feature can be used as a starting point to extract the cell of interest. In the Hue band image, all the cells are uniformly highlighted and the difference in intensity is prominent between the foreground and the background as shown in Figure 5.2 (d). This attribute acquired from the Hue band image is tremendously useful for the subsequent step, namely blast cells localization.

5.2.2 Mask and Marker Preparation

Two grayscale images were extracted from the original HSV image. The Saturation image was represented as “**S**” and the Hue image was represented as “**H**”. To prepare the images for binary morphological operations such as erosion, dilation, etc., the images had to be converted into binary format. For this purpose, the two images “**S**” and “**H**” were converted into binary format using a simple global threshold operation with threshold value of 0.5.

The reason behind choosing **0.5** as a threshold value is that it will not cause any effect on the **Hue** image “**H**”, this means that the full size of the cells will be retained. This is because the **Hue** channel image produces a bimodal histogram. Figure 5.3 illustrates the histogram of the “**H**” image. In Figure 5.2 (d), it can be seen that the image pixels are distributed in a bimodal form. The histogram area surrounded by the green color frame in Figure 5.3, represents the foreground i.e. all the blood components, whereas the area surrounded by the yellow box represents the background.

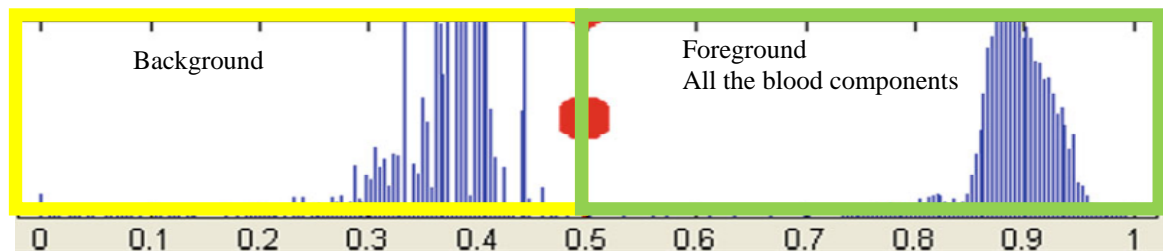


Figure 5.3: Histogram of the Hue channel image in Figure 5.2 (d)

The same threshold value was used for the “S” image, although its histogram was not bimodal. However, the binary version of the “S” image will be used as a **Marker**, so it would not matter if part of the cell of interest is lost after the threshold process. The most important point here is that this operation will keep at least one pixel from the cell of interest to be used later as a seed point for the morphological reconstruction. The use of this threshold value will guarantee that at least one pixel of the nucleus will be retained. This is because the pixels of the blast cells nuclei appear as the brightest component in the image i.e, the nuclei pixels value are higher than 0.5 where all these pixels take the value of 1 after the threshold process. Figure 5.4 (a-b) shows the binary images of the “S” and “H” images, respectively.

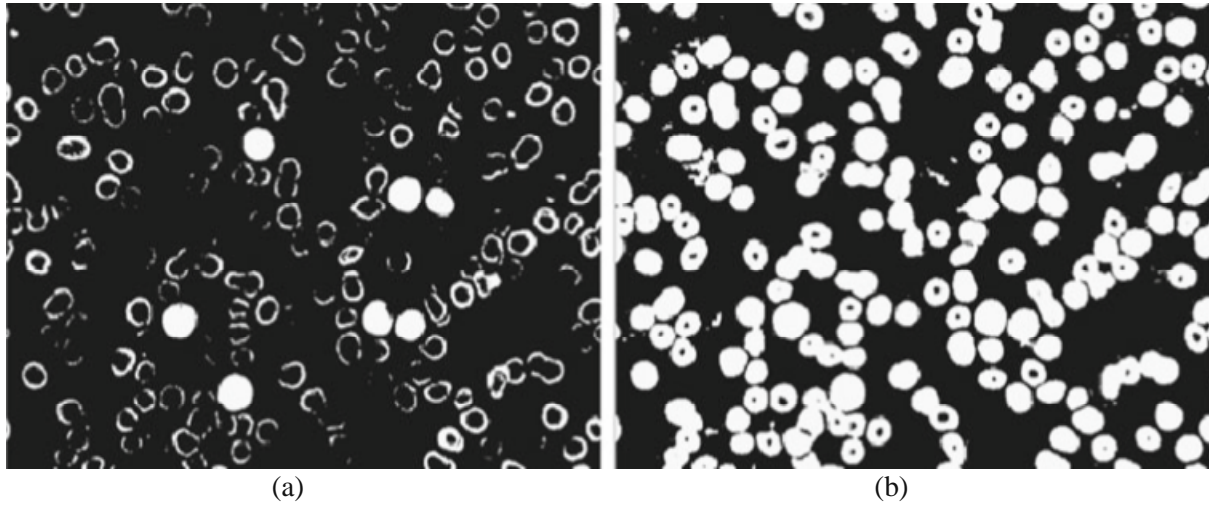


Figure 5.4: Mask and Marker Preparation (a) Binary version of “S” image presented in Figure 5.2(c), (b) Binary version of “H” image presented in Figure 5.2(d)

Now the two new binary images in Figure 5.4 (a-b) represented by **bwS** and **bwH** need to be prepared for the morphological reconstruction process. A morphological opening was performed on the **bwH** image in Figure 5.4 (b) by using a disk shape structuring element (**SE**) with the size of 15 pixels. Morphological opening (Equation 3.18) was achieved by first eroding (Equation 3.17) the image **bwH** by **SE** where the image pixels that were smaller than the **SE** were corroded, then dilating (Equation 3.16) the result by the same **SE**.

Morphological opening was performed in order to smoothen the contour of the cells as well as to segregate the cells that were attached to each other by a thin connection.

On the other hand, the **bwS** image was eroded with the same size of *SE* used earlier for the morphological opening. The erosion process applied to **bwS** removes all objects in the image that are smaller than the *SE* such as the remaining parts of erythrocytes and platelets and retains only a portion of the cells of interest to be used as a Marker. The size of the *SE* used for the opening and erosion operation is a crucial parameter as a big size *SE* may remove even the cell of interest from the **bwS** image. It has been proven experimentally here after trying different *SE* sizes, ranging from 3 to 20, the most suitable size of *SE* for both morphological operations (opening and erosion) was disk shaped *SE* with a 15 pixels radius. Figure 5.5 and Figure 5.6 shows the **bwH** and **bwS** after morphological opening (**Mask**) and erosion (**Marker**), respectively.

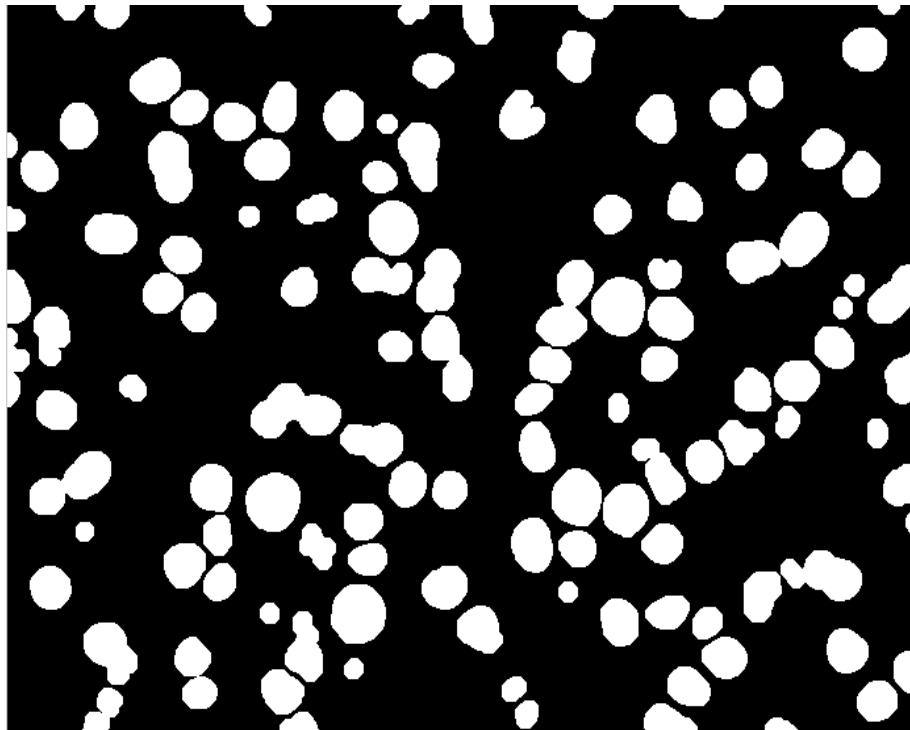


Figure 5.5: The **bwH** Image after Morphological Opening (**Mask**)

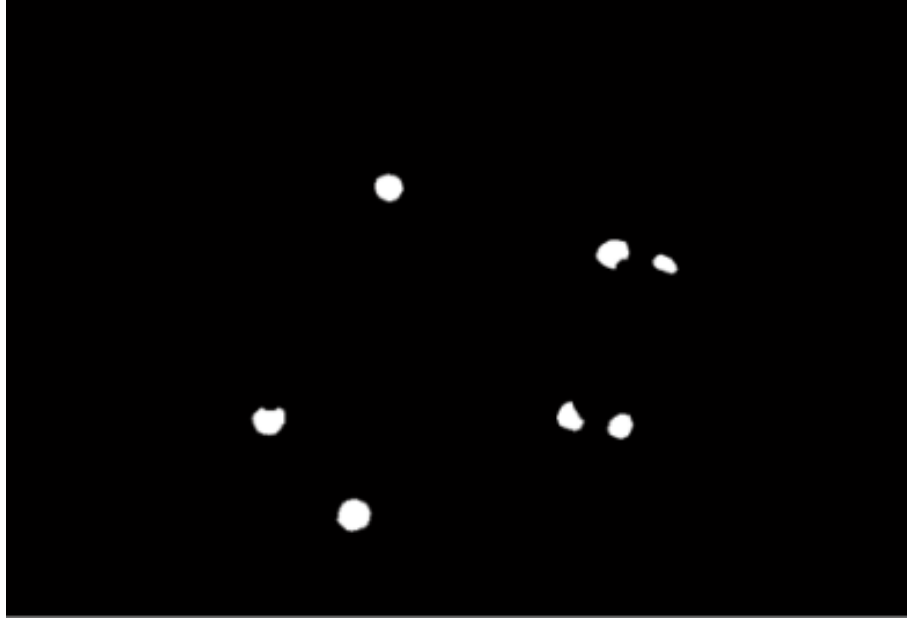


Figure 5.6: The **bwS** Image after Morphological Erosion (**Marker**)

5.2.3 Blast Cells Reconstruction

After the **Mask** and the **Marker** had been prepared, morphological reconstruction was used to reconstruct the cell of interest from the **Mask** image based on the **Marker** image. The cornerstone concept behind the morphological reconstruction is to make repeated dilations of an image, called the **Marker** image, until the contour of the **Marker** image fits under a second image, namely, the **Mask**. The following morphological reconstruction algorithm was applied to retrieve all the blast cells.

<p><i>Repeat</i></p> $Marker_{K+1} = (Marker_K \oplus SE \cap Mask)$ <p><i>Until</i> $Marker_{K+1} = Marker_K$</p>

The process of reconstructing the blast cells using the **Mask** constrained by the **Marker** is illustrated diagrammatically in Figure 5.7.

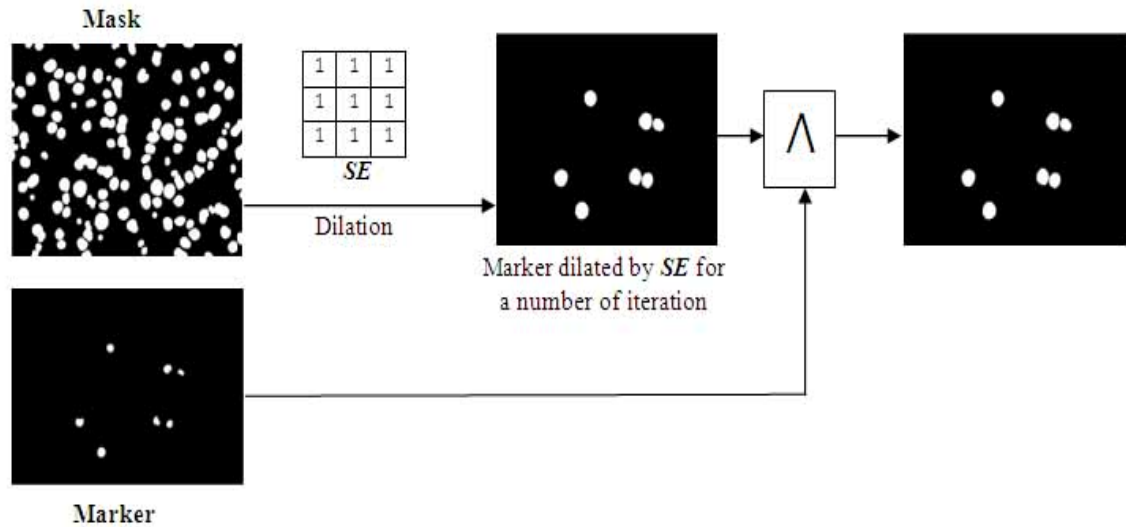


Figure 5.7: Illustration of the blast cells reconstruction from **Marker** and **Mask**

The shape of the **SE** that was used for the purpose of reconstruction was a square shape **SE** with 3×3 pixel size. The selection of the **SE** size and shape in the reconstruction process was not as crucial as with the morphological opening and erosion that were used for the perpetration of the **Marker** and **Mask**. This is because the reconstruction process would not affect the final result as all the pixels outside the contour of the **Mask** would be removed by the intersection operation. Figure 5.8 shows the result of the reconstructed blast cells

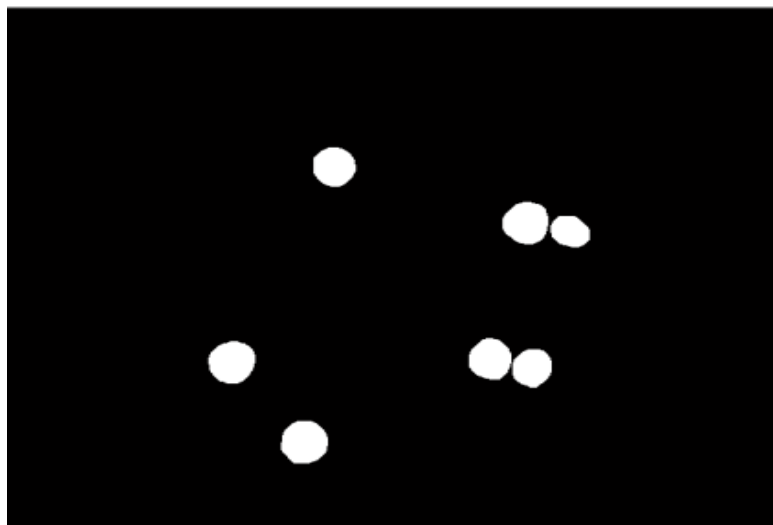


Figure 5.8: Reconstructed blast cells

The result of the reconstructed blast cells image can be used as a mask to restore all the blast cells from the original image with all the other components cleared. Figure 5.9 shows the localized cells of interest retrieved on a white background image.

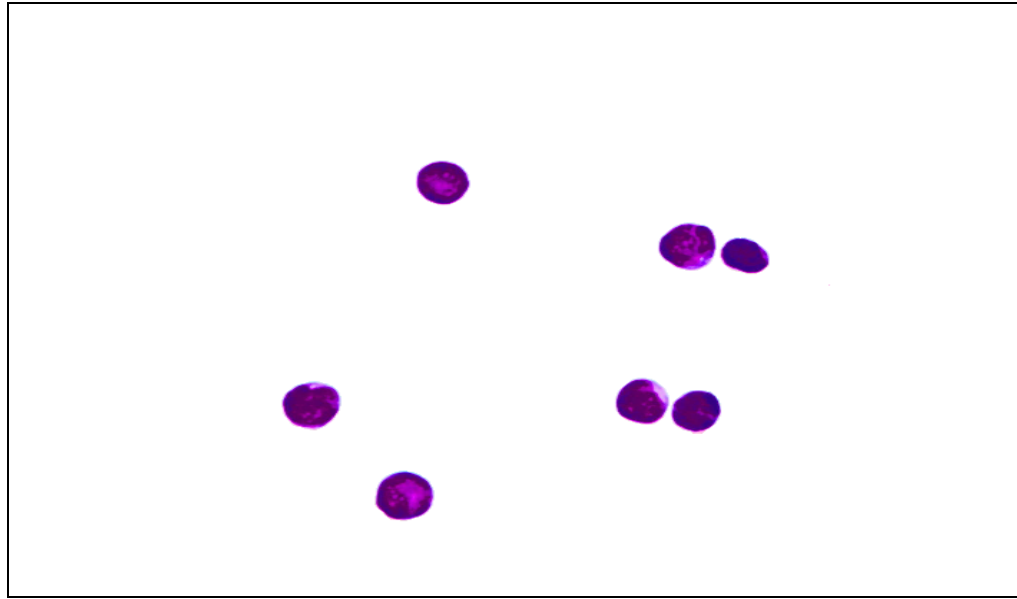


Figure 5.9: The Localized blast cells with the original RGB pixels

5.2.4 Sub-Imaging

Most of the PB images contain more than one blast cell per image. Nonetheless, each blast cell must be extracted into a single sub-image. This is desirable, since each blast cell has to be evaluated separately for the purpose of differentiating Lymphoblast from Myeloblast and vice versa. For this reason, a procedure to crop a sub-image of each blast cell was prepared.

Now, every blast cell in Figure 5.9 could be extracted separately using shape features such as area size, axis length and centroid.

Every blast cell in the resulting binary image (Figure 5.8) was labeled with a different number. For instance, Figure 5.8 shows 7 blast cells; each one was labeled with a different number such that all the pixels for the first blast cell were labeled with 1's, the pixels of the second blast cell were labeled with 2's and the third with 3's and so on as shown in Figure

5.10. An important step in this operation was to create a label matrix that could be used for distinguishing every blast cell in the image.

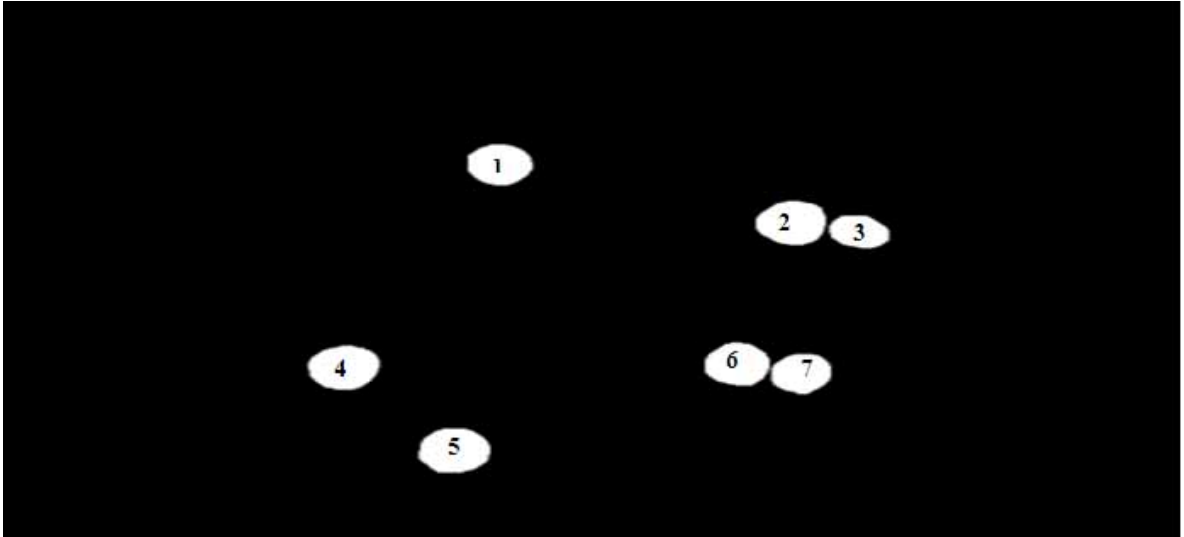


Figure 5.10: Label Matrix where each blast cell are labeled with different number

The first step in the sub-imaging process was to find the center point of each blast cell in the binary image. The center point of the blast cell would be the center point of the sub-image that would contain a single blast cell and the width of the sub-image would be double the axis length (AL) of the blast cell. The reason behind doubling the width of the sub-image was to ensure that enough room was kept for further processing of each sub-image. For example, if the fitting bounding box was used, in some cases a portion of other blast cell which is very near to the cell of interest would be cropped inside the bounding box, this portion needed to be removed for instance, by using morphological clear border, otherwise it would be considered in the feature extraction stage. When using the fitting bounding box, the cell of interest would still be touching the border, and when clearing the objects that were in touch with the sub-image border, the cell of interest could be removed from the sub-image. That is why enough room around the four directions of the blast cells was retained.

The cropping of blast cells sub-images was carried out based on the following algorithm.

Purpose	1:	Cropping sub-images for each blast cell.
Input	2:	Function $f_n(i,j)$ represent a label matrix of size $R \times C$ image
	3:	(Original_ localized) Image of the localized blast Cells with original RGB Pixels
Output	4:	Sub-images of each blast cell
Procedure	5:	For each object in the label matrix image.
	6:	Find the <i>center point</i> of each WBC (x^c, y^c) using Equation 5.1 and 5.2
	7:	For counter=1 \rightarrow n^{th} object number
	8:	$x_n^c = \frac{1}{A_n} \sum_{i=0}^{R-1} \sum_{j=0}^{C-1} x_n(i,j) \dots\dots(5.1)$
	9:	$y_n^c = \frac{1}{A_n} \sum_{i=0}^{R-1} \sum_{j=0}^{C-1} y_n(i,j) \dots\dots(5.2)$
	10:	End For
		Where n is the n^{th} object, A_n is the area of the n^{th} object and it can be found using Equation (5.3)
		$A_n = \sum_{i=0}^{R-1} \sum_{j=0}^{C-1} f_n(i,j) \dots\dots\dots(5.3)$
	11:	Find the horizontal and vertical Axis Length (<i>VAL and HAL</i>) of the WBC
	12:	If $VAL_n > HAL_n$ then
	13:	(Axis length) $AL_n = VAL_n$
	14:	Else
	15:	(Axis length) $AL_n = HAL_n$
	16:	Endif
	17:	(The <i>upper Left Corner</i> point) $ul_n = ((x_n^c - (AL * 2)), (y_n^c - (AL * 2)))$
	18:	(The <i>Width</i> of the sub image) $w_n = (2 * AL)$
	19:	End For
	20:	For each object in the image
	21:	From the image (Original_ localized), Crop the (sub-image) $_n$ (ul_n, w_n)
	22:	End For

Figure 5.11 illustrates the sub-imaging procedure

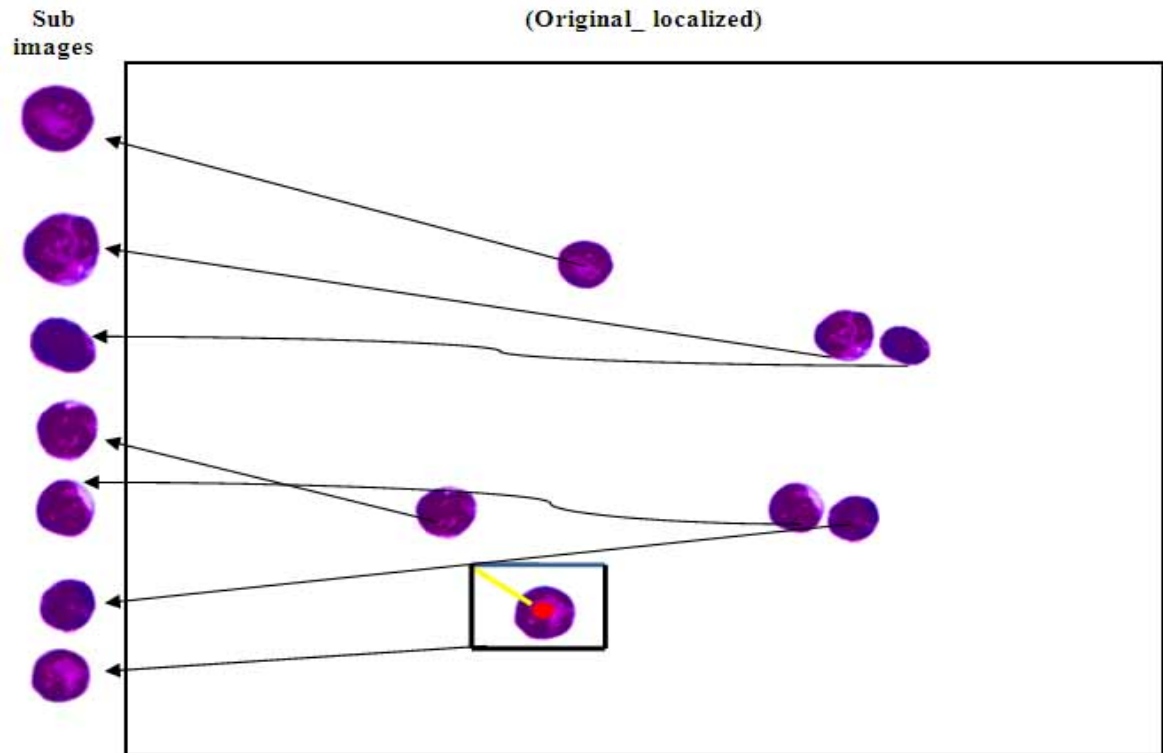


Figure 5.11: Illustration of sub-imaging procedure

In Figure 5.11, the red colored point in the center of the last blast cell is considered the center point (X_n, Y_n) , and the yellow line represents the distance from the center of the blast cell to the top left corner which was found based on the Axis Length (AL) of the blast cell itself. The AL is of variable distance and based on the size of the blast cell. The blue line represents the width of the sub-image, and finally the black lines represent the lines that complete the sub-image rectangle. If another blast cell falls in the same sub-image, part of it will touch the border of the sub-image. In this case all the object's pixels that are touching the border of the image can be removed. This is achieved by clearing the border object, by using an application of a mathematical morphology that removes all connected components of a binary image that touch any image border

The algorithm shown in Section 5.2 was implemented mainly for blast cells' localization and sub-imaging. This algorithm was produced during the early stage of this research when the dataset of images was not yet complete. However, this algorithm did not address a number of important issues in blast cell segmentation such as segregating touching cells and nucleus/cytoplasm separation. The next section discusses improvements to this algorithm where issues such as color variation, touching cells, nucleus/cytoplasm separation are considered.

5.3 A Completed Blast Cell Segmentation Algorithm (CBCSA)

The segmentation method discussed in this section is an advanced alternative to the *BCL* method that was presented in Section 5.2. This method consists of three main modules namely the Erythrocytes Removal and Background Suppression, Blast Cell Reconstruction and the Nucleus/ Cytoplasm Separation module. Figure 5.12 shows the sequence of the three main modules.

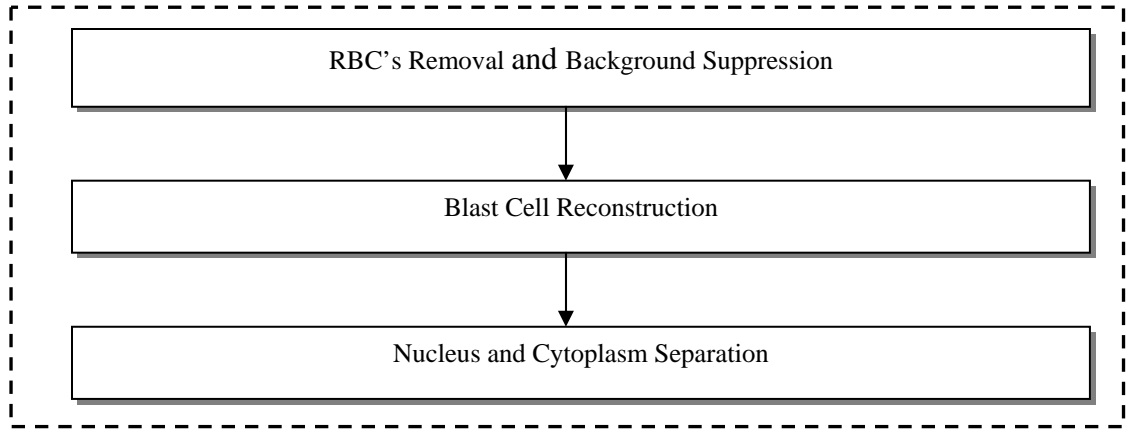


Figure 5.12: Stages of the Completed Blast Cells Segmentation Algorithm (*CBCSA*)

As the image acquisition process continued, more complex PB images were acquired. The complexity here comes in terms of the color uniformity and the homogeneity of the cells distribution in the blood smear. These two problems actually result from the camera and microscope settings that was used as well as from the blood smear preparation procedure. This type of complexity appears with some of the new images acquired with a 1000x magnification in *Dataset-A* (Figure 4.5) and also in *Dataset-B* (Figure 4.6), which is considered a foreign Dataset.

5.3.1 Erythrocytes Removal

In the localization algorithm presented in Section 5.2, the binary version of the **Hue** channel image, namely the **bwH**, was used as a **Mask** (Figure 5.5) in the blast cell reconstruction process. In the **bwH** image, all the erythrocytes and blast cells were presented. Using the **bwH** as a **Mask** works very well when the cells distribution is uniform, in other words, the cells are far away from each other. However, this could cause some problem in cases where the Erythrocytes are touching the cells of interest since the Erythrocytes may be reconstructed together with the cell of interest. In order to tackle this problem, a new module namely the “Erythrocytes Removal” was introduced to prepare the **Mask** image.

The main target of this module is to remove most of the erythrocytes from the peripheral blood image, while highlighting the blast cells and making it more prominent. This module also focuses on enhancing the color distribution of all the blood components presented in the image so that the borders between all the cells become clearer and sharper. This module involves several image processing steps combined in an appropriate sequence to fulfill the required target. Figure 5.13 demonstrates the key steps involved in this module.

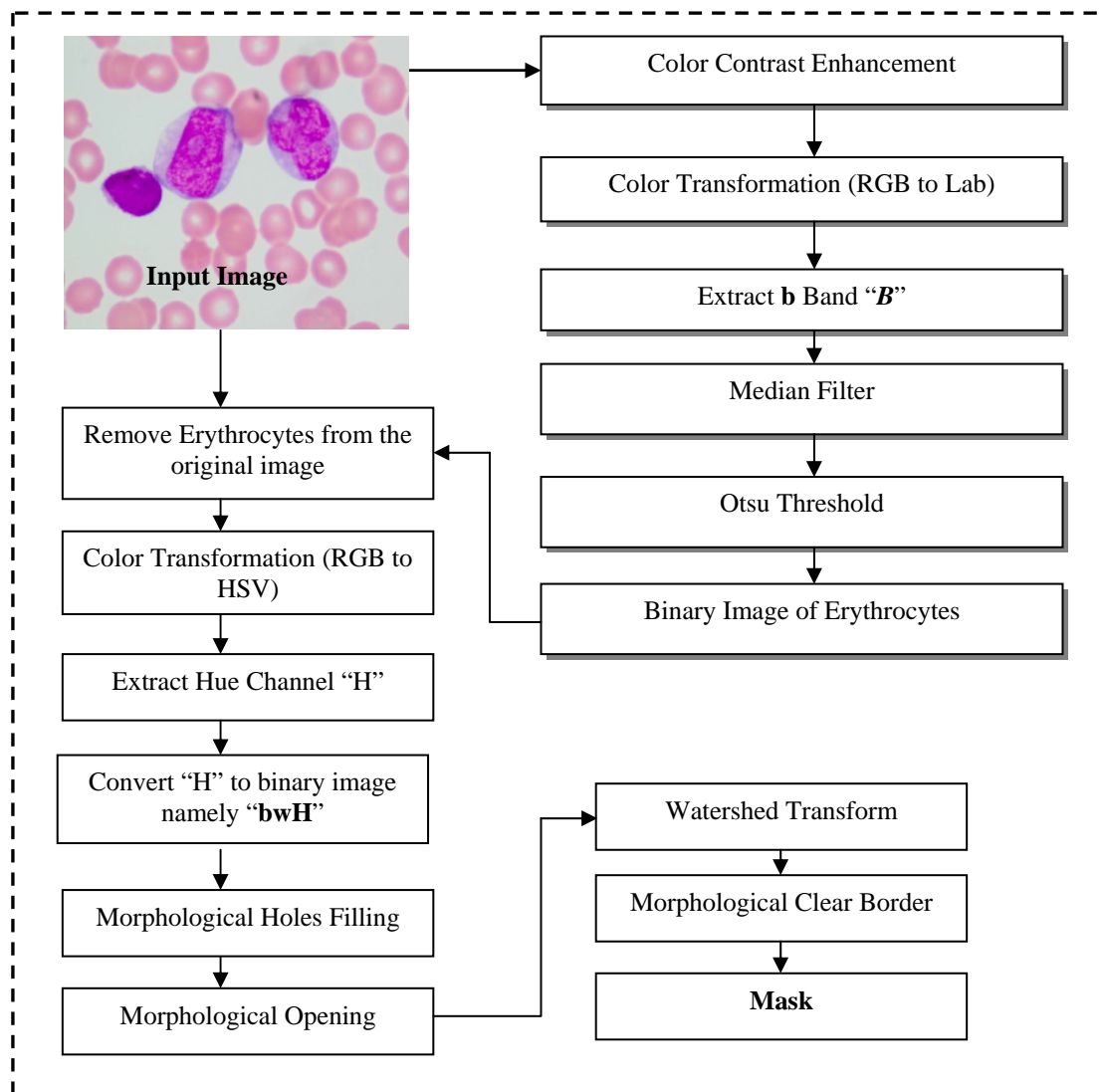


Figure 5.13: Flowchart of Erythrocytes Removal Process

First, the original image color was enhanced using color contrast stretching. This step was applied to reduce the noise generated during the acquisition process such as illumination variation and to make the color of the erythrocytes highly different from the blast cell cytoplasm.

The original RGB image was decomposed into its original three bands namely **Red**, **Green** and **Blue**. Then, the Intensity of each band was adjusted by mapping the image intensity value to a new range such that 1% of the image data was saturated at low and high intensities. After that, the enhanced images were combined to form a new RGB color image with color contrast enhanced, as shown in Figure 5.14.

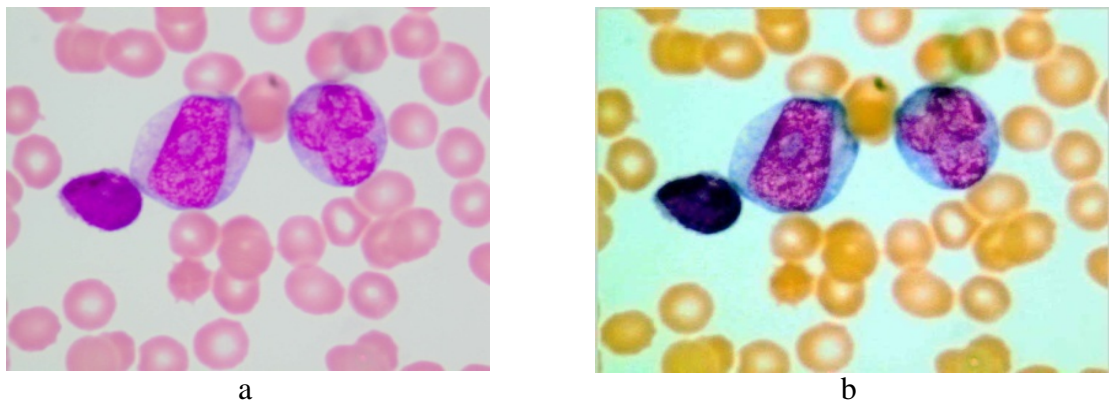


Figure 5.14: Color Contrast Enhancement (a) Original image, (b) Enhanced image

The enhanced image was then transformed to the Lab color space. In the Lab color system, the **L** represented lightness and it can take a value from 0 (black) to 100 (white). The **a** value has a variation from $-a$ (greenness) to $+a$ (redness) while the **b** value has the variation from $-b$ (blueness) to $+b$ (yellowness) (Sahin & Sumnu, 2006). The **b** channel from the Lab color space was then extracted. In fact, it has been observed that the **b** channel makes the Erythrocytes appear more vivid as compared to the nucleus and cytoplasm of the blast cell. This is because yellow is present in the Erythrocytes, where it is almost absent in the blast cells.

The median filter was applied to the **b** channel image to enhance the pixels, which are likely to be affected by noise and also to reduce the pixels discontinuities in the blast cell region especially in the nucleus region (the region surrounded with red color borders in Figure 5.15 (a-b)) due to the non-homogeneous nuclear chromatic pattern. The non-homogeneous nuclear chromatic pattern makes some of the pixels in the nuclear region have a high intensity value similar to the Erythrocytes pixels, where these pixels could be retained after applying the threshold process. Figure 5.15 (a-b) shows the **b** channel image and the smoothed **b** channel image after applying the median filter, respectively.

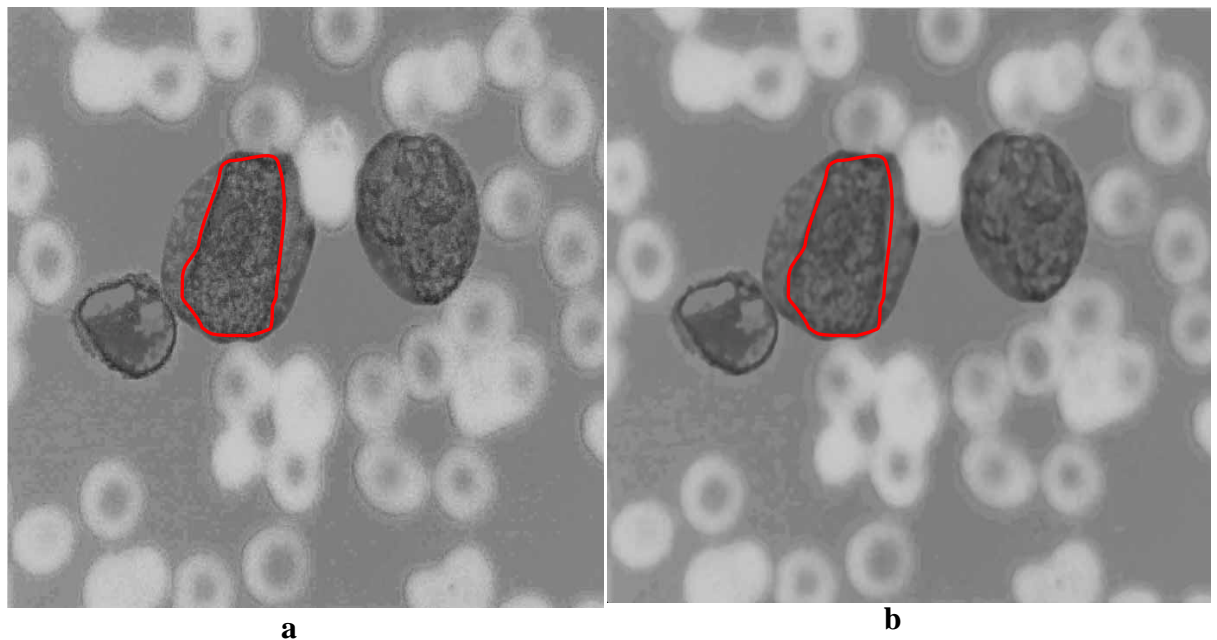


Figure 5.15: Extracted **b** channel image before and after applying median filter. (a) **b** channel image, (b) Smoothed **b** channel image with median filter

To obtain the Erythrocytes binary image, Otsu threshold was used. Figure 5.16 shows the histogram of the smoothed **b** channel image depicted in Figure 5.15(b).

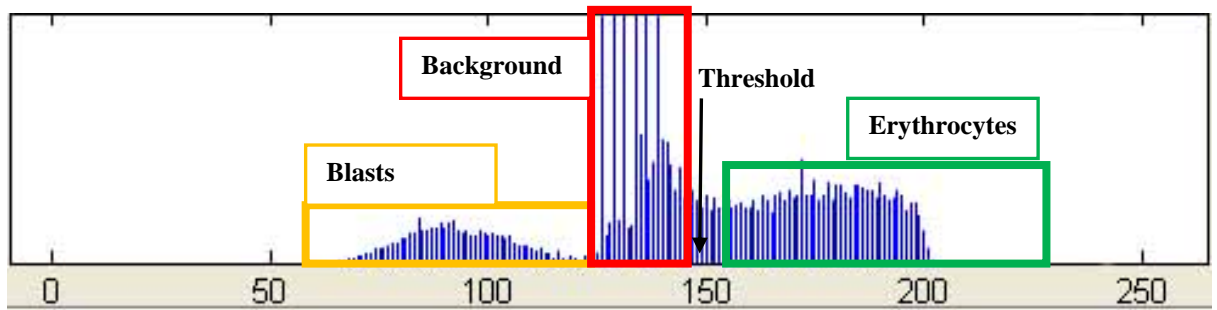


Figure 5.16: Histogram of the smoothed *b* channel image

As shown in Figure 5.16, the histogram is divided into three distinctive areas namely blasts in the yellow rectangle, background plasma in the red rectangle and Erythrocytes in the green rectangle. The Otsu method chooses the threshold value in the valley between two overlapping peaks which falls between the background and the Erythrocytes (Please Refer to Figure 5.16). In this case, the chosen threshold value is 149, this means that the pixels below the selected threshold are turned into black and the pixels value above 149 are turned into white. Figure 5.17 shows the binary image of the highlighted Erythrocytes

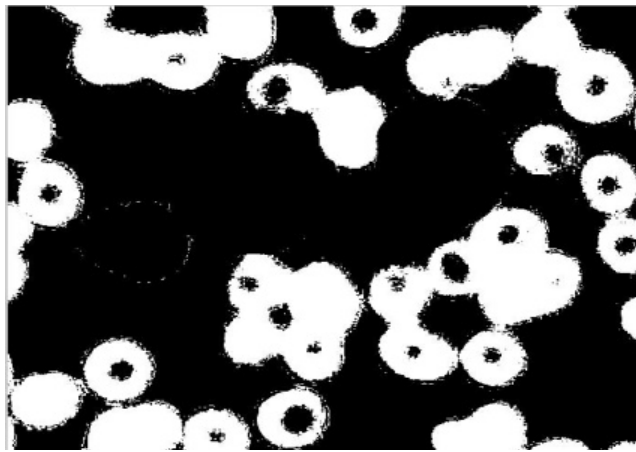


Figure 5.17: Binary image of highlighted Erythrocytes

The binary image in Figure 5.17 was used as a mask to remove all the Erythrocytes from the original image.

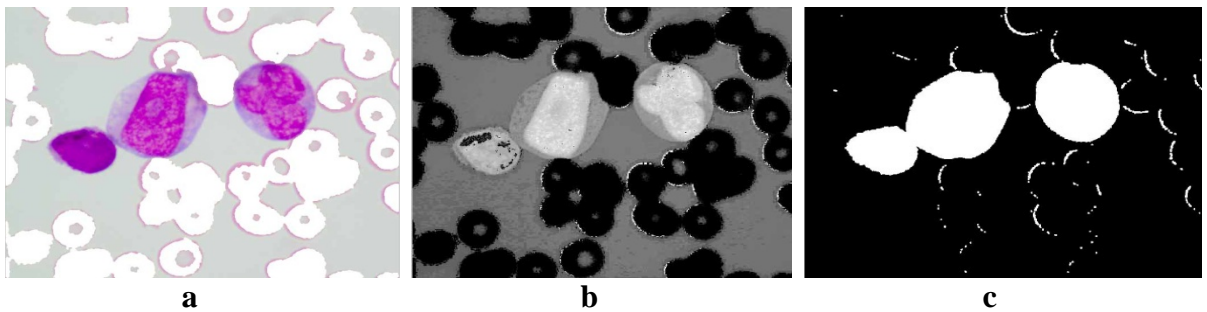


Figure 5.18: Final steps to prepare the Mask image. (a) Original image after subtracting Erythrocytes, (b) Hue channel of (a), (c) Final **Mask** image

Figure 5.18(a) shows the original image after removing almost all the Erythrocytes. Two components remain in the image, namely the blast cells and the background plasma. In order to remove the plasma part from the image, the image was transformed into a HSV color space and the Hue channel was extracted (Figure 5.18 (b)). The Hue channel was used because it can perfectly discriminate between the blast cells and the plasma. Here, all the cellular pixels were highlighted as a foreground and the plasma pixels as a background. To produce the final **Mask** image, a threshold was applied on the **Hue** channel image, and then the result was processed with morphological holes filling and morphological opening. This was done in order to close any holes caused by the threshold process and to clear the image from small particles as well as to smoothen the blast cells contours. The final **Mask** image is shown in Figure 5.18 (c).

5.3.2 Segregating Touching Cells

The problem of adjacent blast cells is actually present in many of the blood smear images of leukemia patients. For better leukemia cell classification, each blast cell should be separated individually in order to extract a group of features from each cell for subsequent classification. To solve this problem, the watershed algorithm implemented in the public domain software for image analysis, ImageJ (NIH, Bethesda, Maryland, USA; <http://rsb.info.nih.gov/ij/>) was used.

The watershed was applied on the binary **Mask** image presented in Figure 5.18 (c). In the **Mask** image, two gray levels were included (black and white). To make the **Mask** image suitable for watershed transform, a distance transform image was needed. For this purpose, the Euclidean distance transform was used. The aim of this was to process the binary image where the distance from every pixel of the object component to the nearest background pixel was computed in order to identify the Ultimate Eroded Points (UEPs). The UEPs was considered as a marker of each blast cell in the image (catchment basins maxima).

At a later stage of this research (Please Refer to Section 7.3), the marker-controlled watershed was used. We propose to use the identified nuclear region obtained from the binary version of the saturation channel (Please Refer to Figure 5.6) as a marker (catchment basin maxima) to force the watershed algorithm to use these markers in order to reduce the over-segmentation (under-segmentation) effects. Figure 5.19 shows the distance map of the **Mask** image while Figure 5.20 shows a 3-D representation of the distance map.

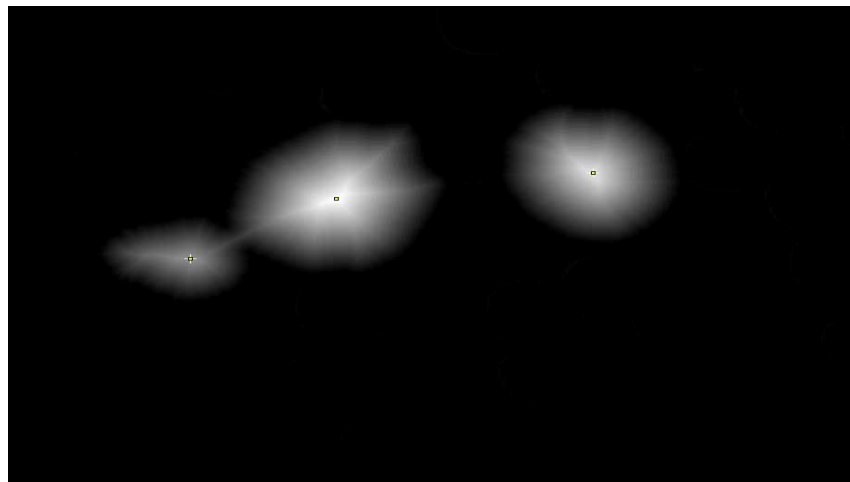


Figure 5.19: Distance map of the Mask image

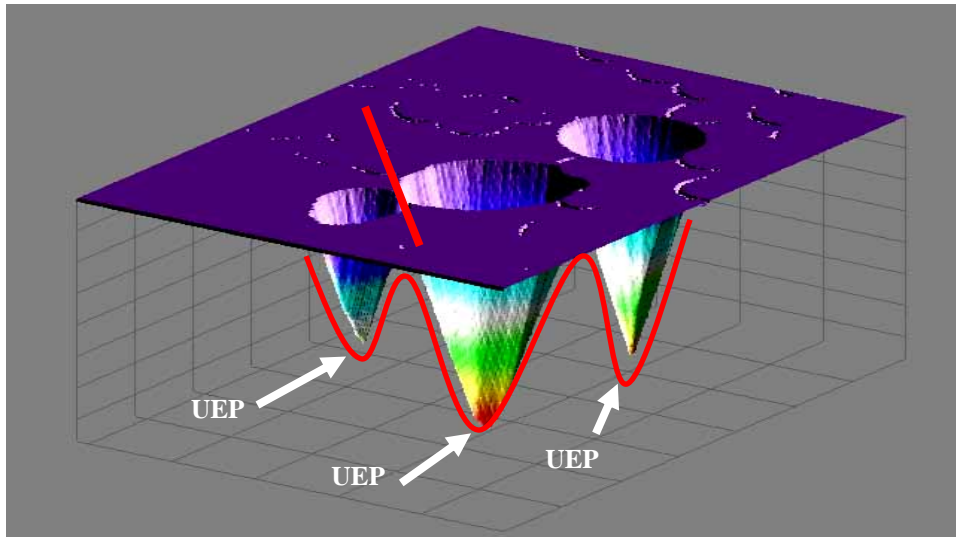


Figure 5.20: 3-D representation of the distance map

By analogy, the water is immersed in the catchment basin in a bottom-up approach, i.e. the water flow into the valley from the local maxima points filling each catchment basin. When the basin is about to overflow, a dam is built on its neighboring ridge line such as the red line in Figure 5.20. The final result of the segregated touching blasts is shown in Figure 5.21.

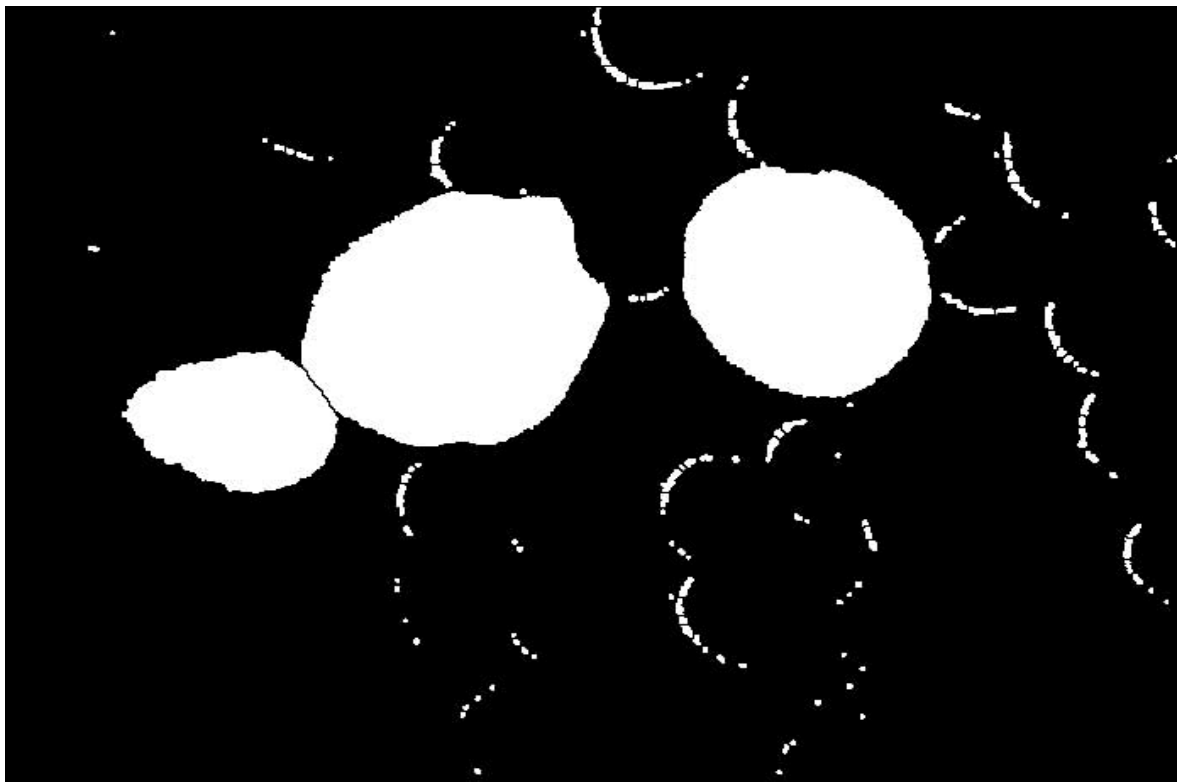


Figure 5.21: Segregated touching blast cells

During the acquisition process, the camera was unable to cover the whole blood smear area and some blast cells were connected to the border of the acquired image and hence, might not be completely visible. These cells had to be removed from the image since they were incomplete and could negatively impact the feature extraction phase. To maintain only the full sized cells, the morphological border clearing was employed.

5.3.3 Marker Image Preparation

As mentioned earlier, the morphological reconstruction process requires two main input images namely, the **Mask** and **Marker**. The preparation of the **Mask** image was described in detail in Section (5.3.1) of this chapter. The **Marker** image was prepared in a similar way to the proposed *BCL* method presented in Section (5.2.2). Here, the original image was transformed into a HSV color space in order to extract the Saturation channel. The Saturation channel image was then converted into a binary image. After that, morphological erosion was applied in order to remove small particles and retain only marker pixels which belong to the blast cell's nucleus.

After the preparation of the **Mask** and **Marker** images, the morphological reconstruction was applied to localize the completed blast cells as shown in Figure 5.22.



Figure 5.22: The Localized Blast cells

To extract each blast cell into a single sub-image, the sub-imaging process was carried out in the same manner as discussed in Section (5.2.4).

5.3.4 Nucleus/Cytoplasm Separation

After extracting each blast cell into a single sub-image, the nucleus was separated from the cytoplasm for further analysis. The blast cells are characterized by immature nuclear chromatin pattern (Rubin et al., 2008). It has been observed, from all the extracted blast cell's sub-images, that the presence of condensed and smudged nuclear chromatin pattern varies from one blast cell to another. The variability in nuclear chromatin pattern makes the separation process much more challenging. Figure 5.23 shows a blast cell sub-image extracted using the proposed extraction algorithm presented in Section 5.3 with less condensed nuclear chromatin and blur nuclear-cytoplasmic margin in some areas.

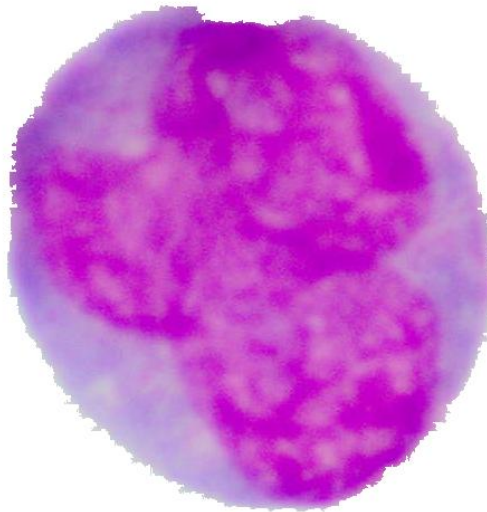


Figure 5.23: Single blast cell sub-image

To address the problem of nucleus/ cytoplasm separation, the approach presented in Figure 5.24 was proposed.

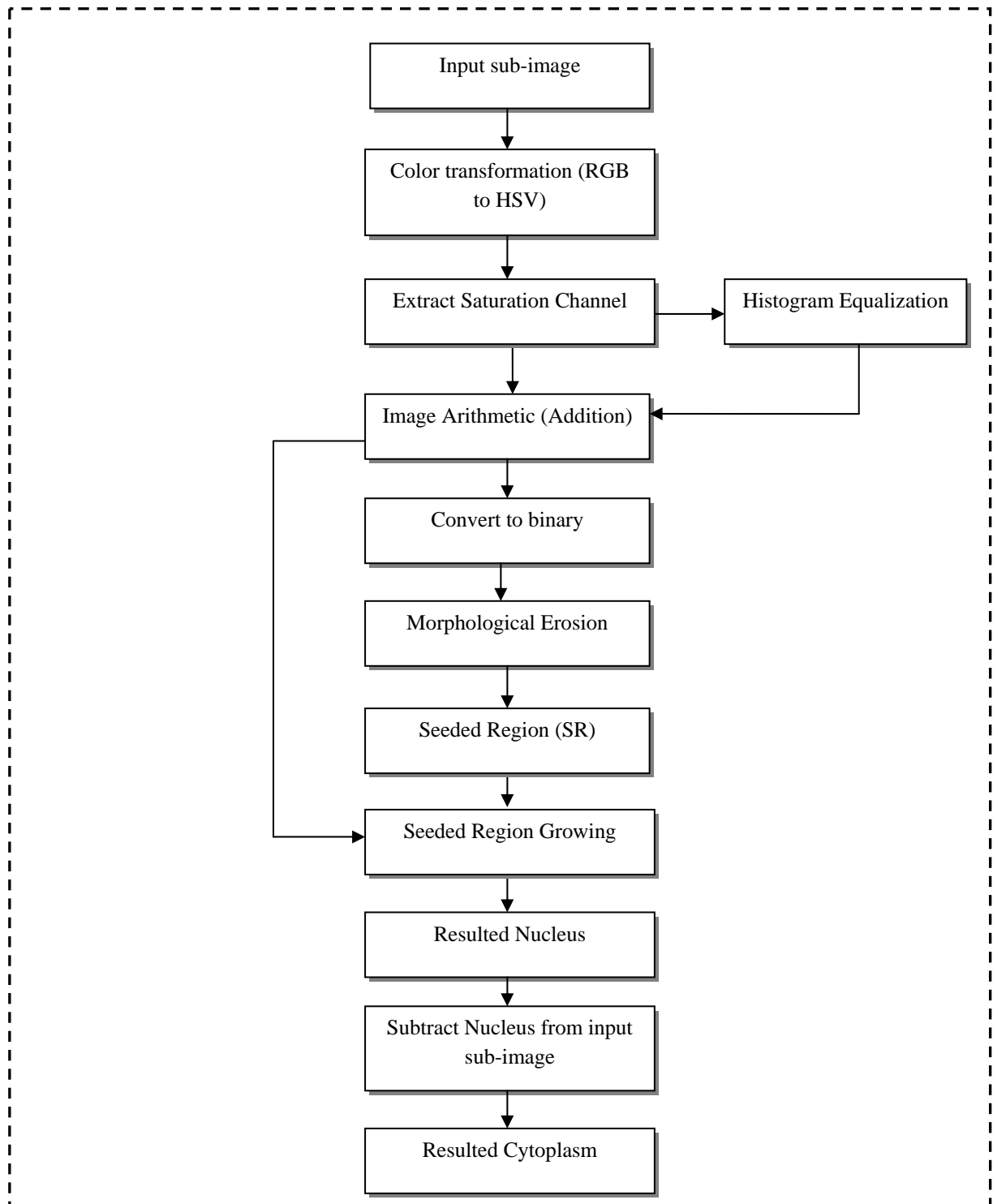


Figure 5.24: Nucleus/Cytoplasm Separation Steps.

The proposed Nucleus/Cytoplasm separation process was mainly based on the Seeded Region Growing (SRG) algorithm (Adams & Bischof, 1994). The input sub-image consists of two main regions; these include the nucleus and cytoplasm. The proposed approach involves, first segmenting the nuclear region, and then subtracting it from the original image in order to only retain the cytoplasm.

As mentioned earlier, the saturation channel of the HSV color space causes the nuclear region to appear as the brightest region in the image. Hence, the input sub-image was converted into HSV color space in order to extract the saturation channel (Figure 5.25 (a)). In order to make the nuclear region's pixels more homogenous, a copy of the saturation image was enhanced with histogram equalization (Figure 5.25 (b)). Later, both the original saturation image and the enhanced one were added together as shown in Figure 5.25 (c).

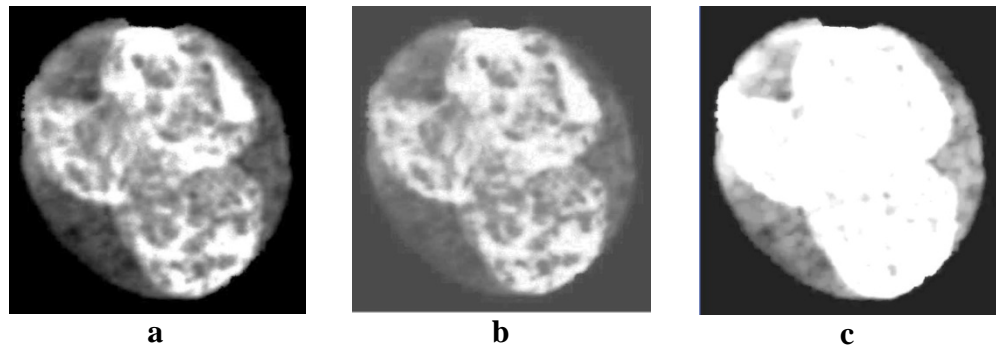


Figure 5.25: Production of homogenous nucleus. (a) Saturation channel, (b) Saturation channel after histogram equalization, (c) Resulting image after arithmetic addition of (a) and (b).

The seeded region was generated automatically by converting the image presented in Figure 5.25(c) into a binary image, as shown in Figure 5.26 (a) and then shrinking it using morphological erosion to guarantee that the seeded region is part of the nuclear region (5.26 (b)).



Figure 5.26: Generating the seeded region. (a) Binary version of image presented in Figure 5.25(c), (b) Seeded region

The seeded region was iteratively grown by comparing all the unlabeled neighboring pixels to the growing region. The region's mean intensity was used as a measure of homogeneity between the unlabeled neighboring pixels and the grown region pixels. If the difference between the unlabeled pixels intensity and the mean intensity of the grown region was below a certain threshold, the unlabeled pixels were assigned to the grown region. The final result of the grown nuclear region is shown in Figure 5.27 where the outline of the grown nuclear region is highlighted with blue color borders.

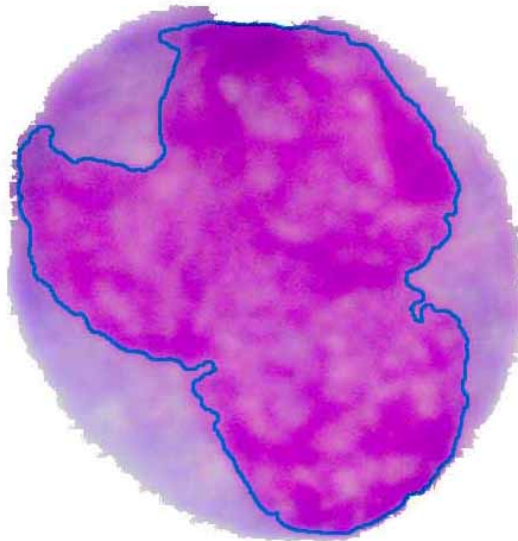


Figure 5.27: The grown nucleus region

5.4 Summary

This chapter presented two methods for localizing and segmenting acute leukemia blast cells in PB images. In Section 5.2 of this chapter, blast cells localization (**BCL**) method, based on color space analysis and mathematical morphology operations, was proposed. The method first determined the most effective and discriminative color channels that were able to discriminate blast cells from other blood components. It then applied a mathematical morphology operation to pre-process and extract the cells of interest. After that, a sub-imaging algorithm was followed in order to extract each blast cell individually.

Section 5.3 described the improvements that were carried out on the **BCL** algorithm (Please Refer to Section 5.2) by addressing a number of issues presented in PB image segmentation such as color variation, segregating touching cells and nucleus/cytoplasm separation. Various stages of the developmental process were covered and the details of the outcome of each stage were presented. The implementation process was long due to the complexity of the process, but was completed successfully, thanks to careful planning.

CHAPTER 6

FEATURE EXTRACTION, SELECTION AND BLAST CELL CLASSIFICATION

6.1 Introduction

Due to the technological advancements in microscopic imaging and image-processing techniques in the recent years, there has been a momentous increase of interest in computer-aided diagnosis of acute leukemia (Madhloom et al., 2012b). This new tool seeks to remove subjectivity and confusion from the diagnostic routine and provides a reliable second opinion to Hematologists and laboratory technologists. However, based on the literature review presented in chapter 3, it is acknowledged that a number of improvements are required for the computer-based algorithms to be routinely adopted in the diagnostic process. Figure 6.1 shows the proposed methodology stages, namely, image acquisition, segmentation (which was discussed in chapter 4 and 5 respectively), feature extraction, feature selection, and classification. This chapter presents a complete description of the experimental steps that were carried out during the implementation of the last three stages (highlighted in Figure 6.1).



Figure 6.1: The proposed methodology stages with emphasis on feature extraction, selection and classification

6.2 Feature Extraction

As discussed earlier in chapter 4, there is no general powerful set of features that is appropriate for all computer vision applications. Feature extraction was used to extract the features in a similar manner to those visually detected by a domain expert that accurately characterizes a blast cell.

The feature extraction methodology of many acute leukemia diagnostic systems has been largely based on the morphological characteristics of blast cells presented in the FAB classification scheme (Please Refer to Table 2.5 and Table 2.6).

In this research, we propose a new feature extraction method, which combines distinctive types of features including novel ones and a number of others which is adopted from existing studies. In the proposed feature extraction process, three different types of features were extracted and combined, comprising of shape, texture and color features.

6.2.1 Shape Features

In clinical diagnostic approaches, domain experts look for the visual differences within the blast cell taking into consideration the morphological and geometrical changes in the appearance of the blast cells. The shape features is a set of scalars that are produced to describe a given shape property such as size, axis length, convex area, etc. The shape descriptor attempts to quantify the geometrical appearances of blast cells in ways that agree with the domain expert's perception. A good and accurate object classification based on shape features, requires the classification process to group objects with similar shapes from a set of images. In this research, the shape features were extracted from a completed blast cell as well as from its nucleus. On the other hand, cytoplasm features were not used as stand- alone features since most of the cytoplasm features are already included in the blast cell. Moreover, stand-alone cytoplasm features are not relevant to the classification of the blast cells (Reta et al., 2010). In order to extract meaningful shape characteristics of the blast cells and its nucleus, a total of 15 simple shape descriptors as well as one ratio feature were used. Although, these features are considered primitive descriptors of an object, they are simple and generally applicable. Furthermore, these features can reflect the semantic meaning of the blast cell in a similar manner to the domain experts'.

Such features included the size of the blast cell or its nucleus, the regularity of the shape; whether it looks like a circle or look like an oval; the length and width of the cell, and other features as discussed in this chapter. Figure 6.2 depicts the graphical representation of some of the simple shape features used in this research.

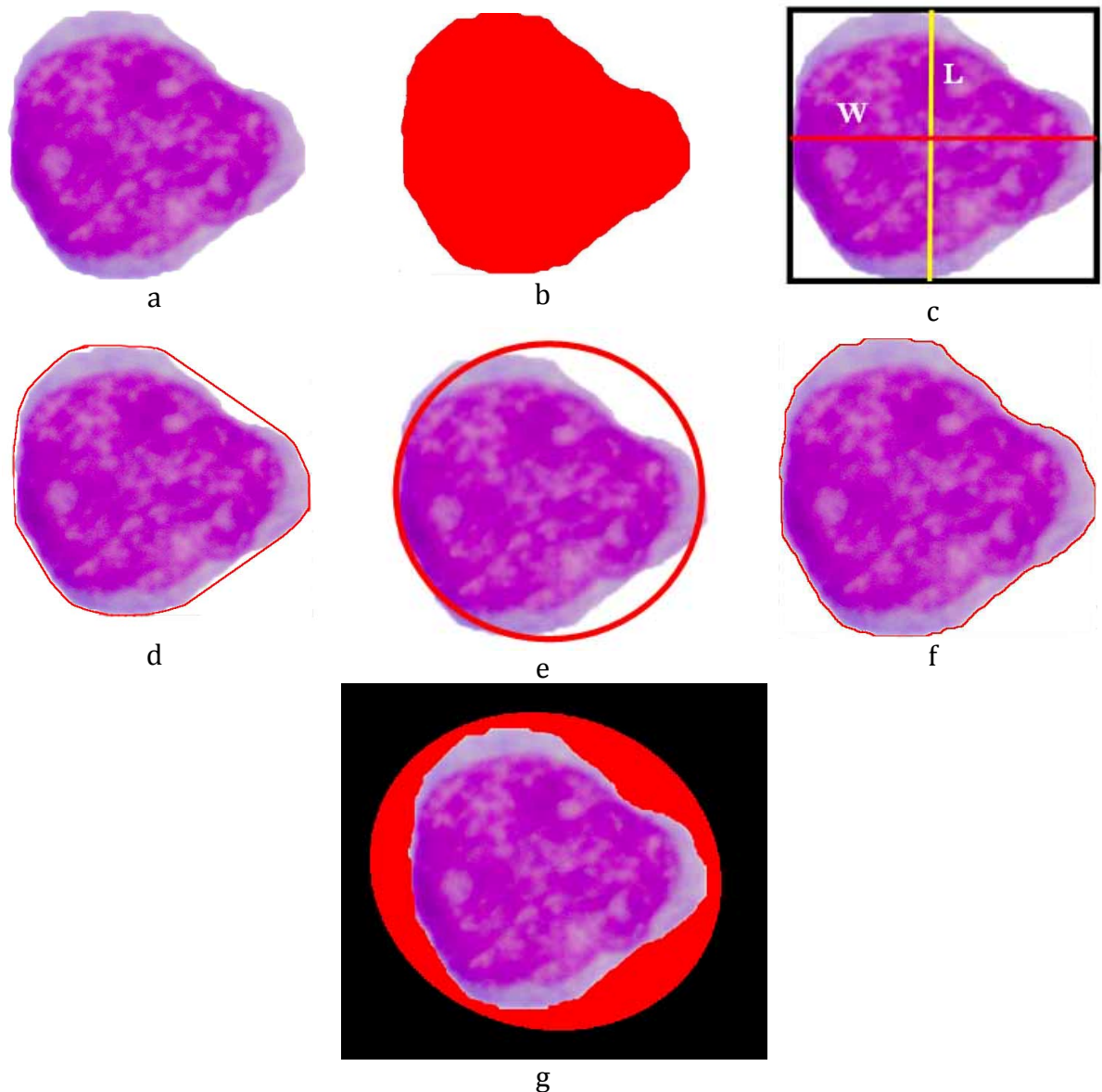


Figure 6.2: Graphical representation of simple shape features. (a) Original blast cell, (b) Area, (c) Rectangular bounding box, (d) Convex Hull (e) circularity, (f) perimeter, (g) minimum bounding ellipse

The following shape-based features were extracted from each blast cell and its nucleus:

Area (*A*): represented by the number of pixels in the ROI. This feature was used to quantify the size of the ROI. Figure 6.2 (b) shows the area of the blast cell as all the pixels are highlighted with red color.

Eccentricity (*ECC*): Figure 6.2 (c) shows the minimal bounding box of a blast cell sample. The Eccentricity represents the ratio of the length *L* and width *W* of the minimal bounding box of the ROI, that is, it measures the degree in which the blast cell and its nucleus resemble an ellipse.

Elongation (*Elo*): Another measure which can be calculated from the bounding box is elongation. It describes the extent of elongation of the blast cell and its nucleus.

Convex Area (*Cov-A*): The convex area is represented by the smallest convex polygon namely convex hull (Figure 6.2 (d)) that can contain the ROI. This feature is useful when measuring the area of the ROI without considering any irregularities in the ROI's boundary.

Solidity (*Sol*): describes the extent to which the ROI is convex or concave. It can be computed as the ratio of the *A* to *Conv-A*, where the solidity of a convex ROI is always 1.

Circularity (*Cir*): Circularity as the ratio of the perimeter squared to the area. This measure describes how much the ROI is similar to a circle and it reflects the complexity of the object boundary (Figure 6.2 (e)).

Perimeter (*P*): Number of boundary pixels that belong to an object. This measure quantifies the distance of the outside boundary of the blast cell or its nucleus (Figure 6.2 (f)).

Rectangularity (*Rec*): Rectangularity is represented as the ratio of the ROI's area to the area of its minimum bounding box. This measure describes how much the ROI is similar to a rectangle.

Equivalent-Diameter (*Equiv-D*): Diameter of the circle having the same area as the ROI.

Best-fit-Ellipse (*BFE*): the *BFE* was used to extract more irregularity features from the blast cell and its nucleus such as (1) area, (2) major and (3) minor axis length, (4) diameter, (5) eccentricity and (6) perimeter. These measures are very important in quantifying the degree of regularity since the degree of regularity of the blast cells and its nucleus varies between the two acute leukemia types especially for the nucleus part (Figure 6.2 (g)).

Ratio Cell/Nuclues ($R_{C/N}$): The ratio of the whole blast cell to its nucleus is represented by the ratio of the nucleus area size to the whole blast cell area size. In clinical diagnosis, this feature is considered a major difference between a lymphoblast and a myeloblast. The nucleus area of the lymphoblast usually occupies more space than the myeloblast's nucleus (Dunn & Pallister, 1998).

In total, 15 shape-based features were extracted from the binary image of each blast cell and its nucleus together with the ratio between the cell and its nucleus. Thus, the shape features vector consists of 31 features for each blast cell.

6.2.2 Texture Features

Another very important feature which is used to differentiate between the *ALL* and *AML* is the “nucleus chromatin structure”. In general, the nuclear chromatin in *ALL* is more condensed than in *AML* (Kumar et al., 2007). In order to capture the features of the nucleus chromatin pattern, two types of texture features were used namely the Histogram-based features and the GLCM features. Both of these texture features were extracted from the grayscale image of the nucleus.

6.2.2.1 Histogram-Based Features

In this simplest form of texture feature, the features are extracted from the gray-level histogram of an image (Please Refer to Section 3.3.2.1). Figure 6.3 (a) and Figure 6.4 (a) show two samples of an *ALL* and *AML* nuclei, respectively.

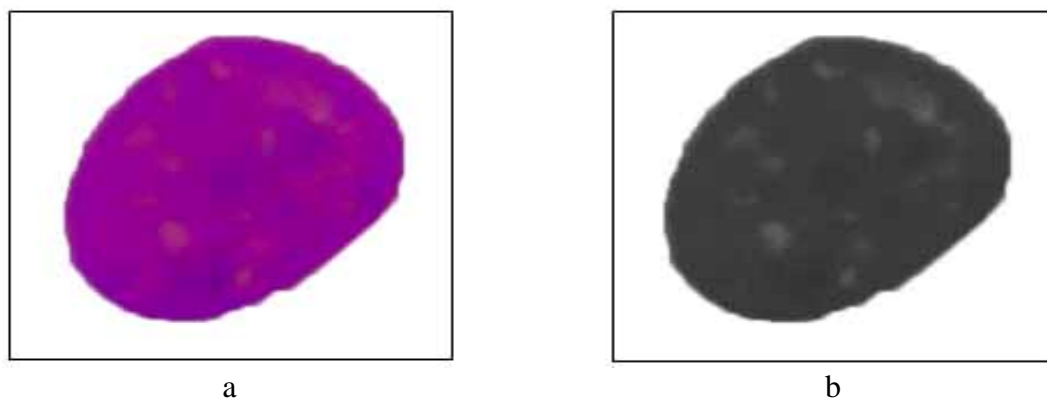


Figure 6.3: Histogram-based features (*ALL*). (a) Original *ALL* sample, (b) Grayscale image of the *ALL* sample

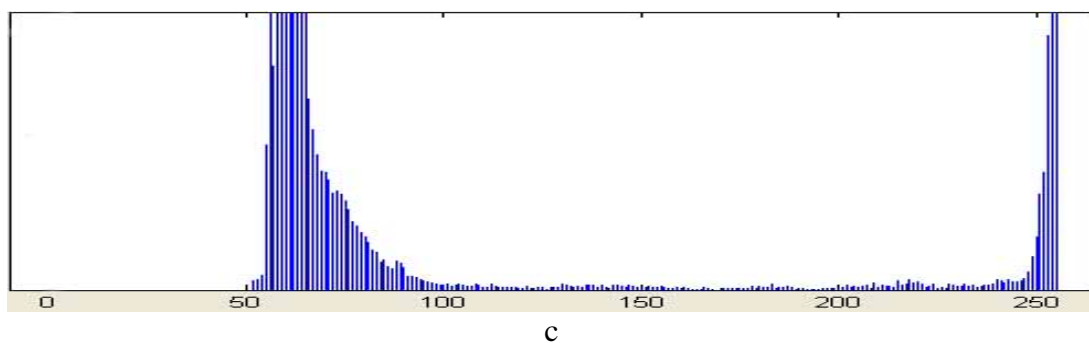


Figure 6.3: Histogram-based features (*ALL*) (c) Histogram of the grayscale *ALL* Sample

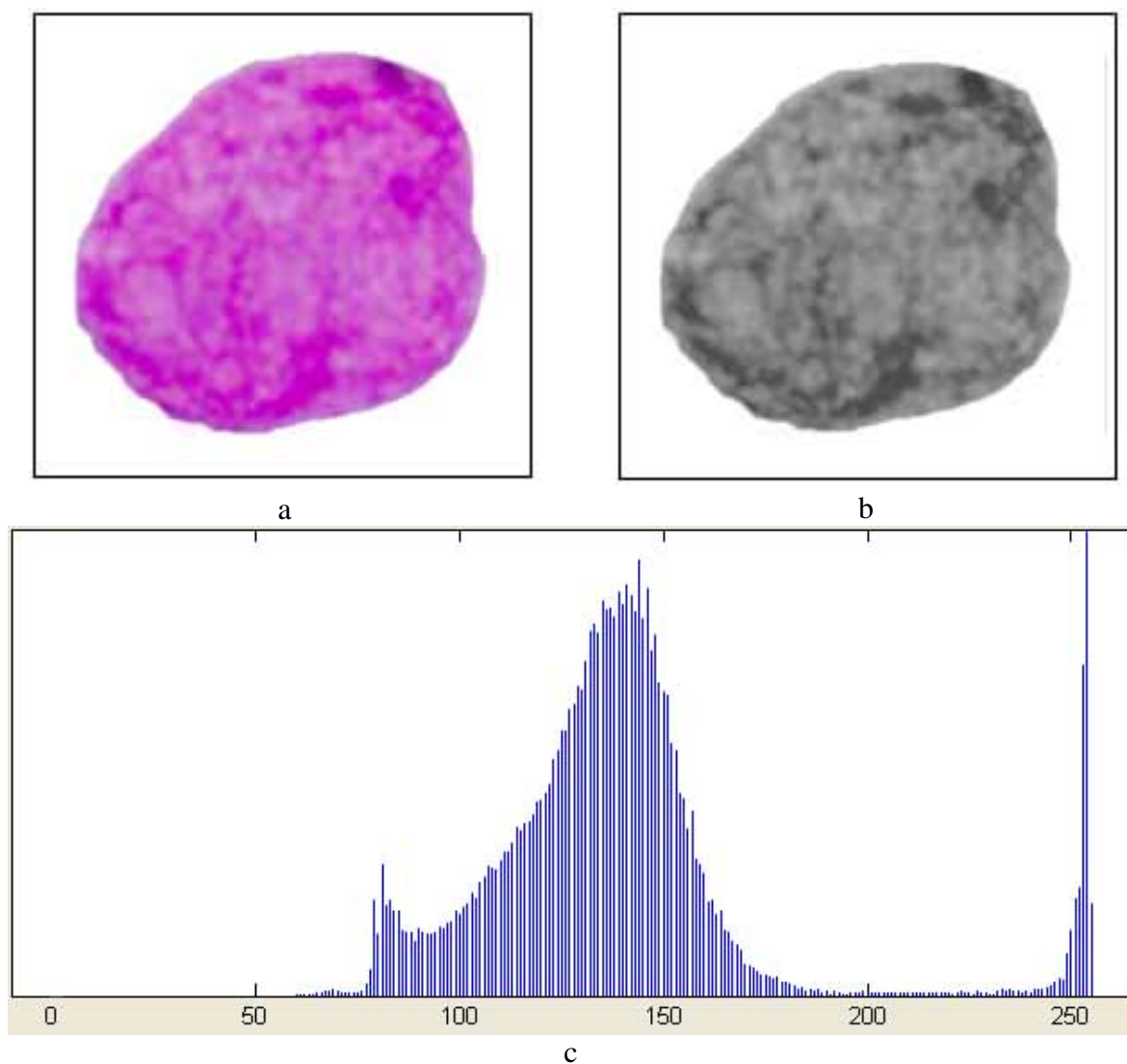


Figure 6.4 Histogram-based features (*AML*). (a) Original *AML* sample, (b) Grayscale image of the *AML* sample, (c) Histogram of the grayscale *AML* Sample

The histogram-based features were extracted from the nucleus grayscale image. A sample of the *ALL* nucleus in a grayscale format is shown in Figure 6.3 (b), while Figure 6.4 (b) shows the sample of *AML* nucleus in grayscale format. Figure 6.3 (c) and Figure 6.4(c) show the histogram of the *ALL* and *AML* nuclei, respectively. It can be observed from the histograms of both acute leukemia samples that there is a clear difference in the nuclear chromatic pattern. To capture the differences between the *ALL* and the *AML* histograms, 6 histogram-based features were extracted from the histogram of the grayscale image for each nucleus sub-image, taking into consideration to remove the white background pixels before calculating the histogram statistics. The 6 histogram-based features are tabulated in Table 6.1 below.

Table 6.1: Histogram-based features extracted from blast cell nucleus

No.	Histogram Feature	Equation No.
1.	Mean	(3.20)
2.	Standard deviation	(3.21)
3.	Energy	(3.22)
4.	Entropy	(3.23)
5.	Skewness	(3.24)
6.	Kurtosis	(3.25)

6.2.2.2 GLCM Features

GLCM features are known to be the most common and successful techniques for texture analysis. The theoretical background of GLCM is illustrated in Section 3.3.2.2, whereas Table 3.2 rows 1-14 presents the standard GLCM texture descriptors proposed by (Haralick et al. 1979).

In this research, apart from using the standard GLCM texture descriptors discussed by (Haralick, 1973), other recent GLCM texture descriptors discussed by (Clausi, 2002; Soh &

Tsatsoulis 1999) and the MATLAB Image Processing Toolbox were adopted and used for texture features computation of the nuclear chromatin structure. These features are tabulated in Tables 6.2 through 6.5, respectively.

Table 6.2: GLCM texture features from (Haralick, 1973)

No.	GLCM Feature	Equation No.
1.	Angular Second Moment	(3.26)
2.	Contrast	(3.27)
3.	Correlation	(3.28)
4.	Sum of Square Variance	(3.29)
5.	Sum Average	(3.31)
6.	Sum Entropy	(3.32)
7.	Sum Variance	(3.33)
8.	Entropy	(3.34)
9.	Difference Variance	(3.35)
10.	Difference Entropy	(3.36)
11.	Information Measure of Correlation I	(3.37)
12.	Information Measure of Correlation II	(3.38)

Table 6.3: GLCM texture features from Soh & Tsatsoulis (1999)

No.	GLCM Feature	Equation No.
1.	Autocorrelation	(3.40)
2.	Cluster Prominence	(3.41)
3.	Cluster Shade	(3.42)
4.	Dissimilarity	(3.43)
5.	Homogeneity	(3.44)
6.	Maximum Probability	(3.45)

Table 6.4: GLCM texture features from Clausi (2002)

No.	GLCM Feature	Equation No.
1.	Inverse Difference Normalized	(3.46)
2.	Inverse Difference Moment Normalized	(3.47)

Table 6.5: GLCM texture features from MATLAB Image Processing Toolbox

No.	GLCM Feature	Equation No.
1.	Homogeneity	(3.48)
2.	Correlation	(3.49)

Based on Haralick GLCM features, a total of 12 distinctive features were employed (Table 6.2). However two features were omitted from the original 14 Haralick features namely the inverse difference moment and the maximum correlation coefficient. The inverse difference moment was not considered in this research as this feature is already similar to the Homogeneity feature calculated in (Soh & Tsatsoulis 1999) method. On the other hand, it has been found in the literature (Désir et al., 2010, Markey et al., 1999) that the maximum correlation coefficient is computationally expensive. For this reason the maximum correlation coefficient was not considered.

Besides the Haralick features, a total of 6 GLCM texture feature descriptors shown in Table 6.3 (Clasi, 2002) and 2 (Soh & Tsatsoulis, 1999) shown in Table 6.4 were adopted for texture feature extraction. These texture descriptors are more recently proposed in the literature and have indicated more promising results in pattern classification problems using textures. The *graycoprops* function in the MATLAB calculates the statistics specified from GLCMs, which has four texture descriptors, namely: contrast, correlation, energy and homogeneity.

It is noticed that the calculation formulae of the *Contrast* and *Energy* texture descriptors in the MATLAB Image Processing Toolbox are similar to the ones proposed by Haralick (1973). Thus, only the two texture descriptors implemented in the MATLAB Image Processing Toolbox, namely correlation and homogeneity, were applied in this research as shown in Table 6.5.

In total, 22 GLCM texture descriptors were identified in this research, which are shown in Tables 6.2, 6.3, 6.4 and 6.5. The GLCM computational parameters: *Number of grey levels*, *Distance between pixels* and *Angle* (Please Refer Section 3.3.2.2) used for texture feature extraction are discussed as follows:

(a) *Number of gray Levels*

In this research, we considered the ability of the GLCM features to classify nucleus chromatin structure using six different gray-levels quantization. This is important since different quantization levels might produce different results. Hence, the objective of using different quantization levels is to investigate the behavior of 22 co-occurrence features, as a function of six gray-levels quantization that are powers-of-two (8, 16, 32, 64, 128, and 256 levels), to classify acute leukemia blast cells based on the nucleus chromatic structures. Figure 6.5 shows samples of a nucleus sub-image quantized to 8, 16, 32, 64, 128, and 256 levels respectively.

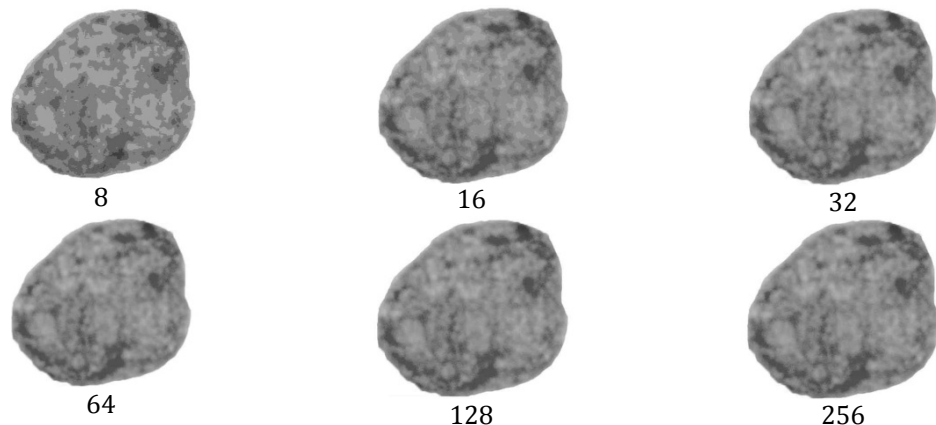


Figure 6.5: The effect of gray level quantization on nucleus chromatic pattern

To our knowledge, this approach of co-occurrence statistical features applied to the blast cell's nucleus images is original.

(b) Distance Between Pixels

GLCMs are constructed by identifying neighboring pairs of image cells with a distance d from each other and incrementing the matrix position corresponding to the grey level intensity of both cells, as shown in Figure 3.13. In this research, the value of d is chosen as $d = 1$, in order to represent the distance between the pixel of interest and the neighboring pixels in each ROI.

(c) Angle

As indicated in Figure 3.13, the most common GLCM directions for a given distance d are: $(0^\circ, d)$, $P(45^\circ, d)$, $P(90^\circ, d)$, $P(135^\circ, d)$. In this research, all four directions were used.

The total number of GLCM texture features computed using the GLCM parameters mentioned above, are listed in Table 6.6. For each nucleus sub-image, 528 feature values were computed using the 22 GLCM texture features as listed in Table 6.2 to Table 6.5. This means that, for each GLCM matrix calculated for the four directions based on specific quantization, 88 feature values were calculated.

Table 6.6: GLCM texture features calculated for each nucleus sub-image

GLCM Parameters	Parameters Value
GLCM Texture Features (Table 6.2, 6.3, 6.4 and 6.5)	22
GLCM directions (0° , 45° , 90° , 135°)	4
Distance ($d = 1$)	1
Gray-Levels Quantization (8,16,32,64,128,256)	6
Total number of GLCM features calculated for each nucleus sub-image	528

6.2.3 Color Features

The simplest and most frequently used representation of color content is the color histogram. It represents the joint probability of the intensities of the three color channels and captures the global color distribution of an image. Here, color features were derived from the two different color spaces namely RGB and HSV. Each color image was decomposed into its three distinctive color channels including the red, green, and blue channels in the RGB and hue, saturation, value in the HSV. The histogram was calculated from each color band image. A total of 6 features, tabulated in Table 6.1 were derived from each generated histogram in the same way as in Section 6.2.2.1. In total, 36 color features were extracted for each nucleus sub-image.

6.3 Feature Selection

Various types of features were employed to address the blast cell's morphology, including shape, texture and color. In total, 601 features were extracted from each blast cell, and its nucleus: 31 of these were shape-based features, while 534 were texture-based features and 36 were color-based features. As discussed in Section 4.4 of this thesis, for feature selection, SFS wrapper technique was used by employing a 10-fold cross-validation with a Naive Bayes learner to evaluate feature sets. Naive Bayes algorithm was adopted as a learner because of its robust performance with respect to irrelevant features (Kohavi and John., 1997).

6.4 Blast Cells Classification

6.4.1 Data Normalization

The main advantage of using scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation (Hsu et al., 2003). For that purpose and prior to the development of the classification engines, the optimal subset of features obtained were normalized in the range between 0 and 1.

6.4.2 Training and Testing Data Separation

To implement the two selected classifiers (MLP-NN and SVM), the normalized data needs to be separated into two distinct sets, i.e. the training set and the testing set. As illustrated in Chapter 4 Table 4.2, the total number of blast cells obtained from the acquired PB images was 1303, out of which 987 are *AML* samples and the remaining 325 are *ALL* samples.

It is worth mentioning that the acute leukemia classification was performed based on the two main types, the *ALL* and the *AML*, and not on the subtypes level. This is because acute leukemia subtypes are important for the disease prognosis (Cairo & Perkins, 2012), and the subtypes under the same group have similar treatment options except for the M3 subtype of *AML*. Recent studies in molecular biology have discovered that this subtype requires different treatment options than the other *AML* subtypes (Abdul-Hamid, 2011). However, differentiating between the M3 and other *AML* subtypes is beyond the scope of this research and could be addressed in future work.

In order to split the feature data into the training and the testing sets, the Holdout method was used. In this method the samples were randomly divided, two-thirds (70 percent) of the samples from both classes were allocated to the training set and the remaining one-third (30

percent) of the samples were allocated to the testing set. The specifications of the samples in the training and the testing sets are indicated in Table 6.7.

Table 6.7: Ratio of samples used for training and testing

Class	Number of Samples	Training Set	Testing Set
<i>ALL</i>	325	227	98
<i>AML</i>	978	685	293
Total	1303	912	391

The dataset shows a degree of imbalanced distribution in the number of samples from both types of acute leukemia, i.e. there were more *AML* samples than *ALLs*'. The *ALL* (minority class) to the *AML* (majority class) ratio is approximately 1:3, this ratio is not considered to be tremendously significant (Chawla et al., 2011). At a later stage of this research, an oversampling technique, (namely, SMOTE) was used to balance the ratio of the data samples.

6.4.3 MLP-NN Optimization, Training and Testing

In the MLP-NN design process, there were several critical parameters that need to be determined properly, i.e. the learning rate, the number of hidden layer, the number of neurons in the hidden layer, the activation function, and the number of epochs. In general, the performance of the ANN would be very poor, if these parameters were not chosen correctly. As discussed in Chapter 4 Section 4.6.1 of this thesis, the number of the MLP-NN input nodes depends directly on the number of the input features. While the number of the output nodes was set to two outputs (*ALL* and *AML*). Meanwhile, a MLP-NN trained with back-propagation was employed. The number of hidden layers was set to one and the sigmoid activation function was used throughout the model. On the other hand, there were still other parameters, namely, the learning rate, the number of neurons in the hidden layer and the training cycle (epoch) that was needed to be chosen properly. The validation training effect was compared on different network architecture using trial and error.

Three different learning rates were tested, these included, 0.001, 0.01 and 0.1. Moreover, a total of six different numbers of nodes in the hidden layer (H) were experimented. These included, $H = 2, 4, 6, 8, 10, 12$. A maximum training cycle (epoch) of 1200, taking into consideration the validation accuracy at every hundred training cycle, were experimented with. A total number of 18 different network architectures were tested. The results of the experiments showed that the best performance was obtained at a learning rate of 0.01 with four nodes in the hidden layer (Please Refer to Section 7.5).

To build an appropriate MLP-NN model using back-propagation algorithm, the available data was divided into three portions. One approach that can be used to avoid over-fitting, is the K-fold cross-validation technique. In order to select the best model architecture, the method of 10-fold cross validation was used.

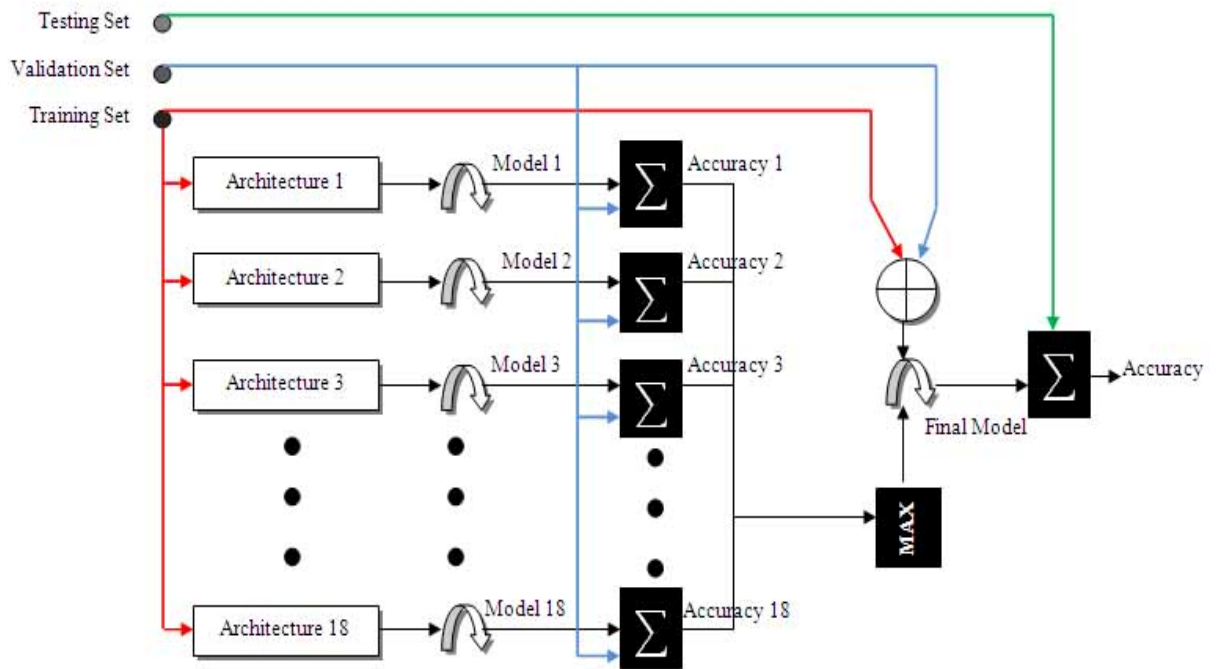


Figure 6.6: MLP-NN model optimization, Training and Testing process

As shown in Figure 6.6, we ran 18 sets of 10-fold cross validation experiments each with a different set of parameters.

The accuracy results of the 18 different models were then evaluated and compared. The model that produced the maximum accuracy was selected as the best architecture. The best model architecture was then trained with the whole training set (912 samples) and tested with an unseen data using the testing set (391 samples). For each model, the training parameters were selected, then the training set (912 samples) was iteratively divided into 10 folds; one fold was selected and removed from the dataset. For each model, the network was trained on 9 folds, and the accuracy of the model was calculated on the removed fold. This process was repeated for each fold and the calculated accuracy across the whole training set was averaged.

6.4.4 SVM Optimization, Training and Testing

Based on the SVM requirement analysis discussed in Chapter 4 Section 4.6.2, SVM classifier with the RBF kernel was used in order to classify acute leukemia blasts. Meanwhile, there were two important parameters in RBF kernel, namely the C and γ . These needed to be determined in order to construct the SVM classifier with an optimum balance between its memorization and generalization capability.

The flow chart presented in Figure 6.7 describes the whole process of acute leukemia blast cells classification using the optimal SVM classification engine. The algorithm started by splitting the dataset into training and testing sets (Please Refer to Table 6.7). The training set was used in the process of finding the best (C, γ) combination, and was also used to train the final SVM model with the best hyper-parameter found. The test set was used to evaluate the model performance. The test set was considered as an unseen data and was not used as part of the SVM training process. Nonetheless, the SVM training and testing process is more or less straightforward, however, the main issue is to find the best hyper-

parameter (C, γ) pair which can be used to minimize the classification error on the unseen testing set.

In order to optimize the SVM hyper-parameter, the Grid Search method proposed by (Hsu et al., 2003) was used. Although, the grid search is considered exhaustive and time consuming compared to the other techniques such as the heuristics method, however, the computational cost required by the grid search is not significantly higher than that by advance techniques since the search is carried out for only two parameters (Hsu et al., 2003). Moreover, the grid search is still labeled as the most reliable parameter search approach (Zhuang & Dai., 2006).

As recommended by (Hsu *et al.* 2003), an exponentially growing sequences of parameters C and γ were used to identify the optimum parameter values with respect to the best 10-fold cross validation accuracy.

In this trial and error procedure, 88 sequences of parameters in the range $C = \{2^{-5}, 2^{-4}, \dots, 2^{15}\}$ and $\gamma = \{2^{-11}, 2^{-10}, \dots, 2^3\}$ were evaluated. The grid search was evaluated on each (C, γ) pair based the SVM classification performance using the 10 fold cross-validation approach. The training data was randomly divided into 10 disjoint sets or folds. Each fold was taken from the training data and employed as a test fold for validation purposes. The model was trained with 9 folds, and then the performance measure was computed on the 10th fold. This process was repeated for each testing (validation) fold. Later, the 10 performance measures were averaged so that a single estimated performance measure could be obtained. The 10 fold cross-validation process was repeated with each (C, γ) pair for 88 times. The hyper-parameter (C, γ) corresponding to the highest average performance was adopted. Consequently, the elected hyper-parameters were used to train the SVM classification engine using the completed training set.

Nevertheless, during the training process, the values of the selected hyper-parameters were tuned manually within the range of the selected hyper-parameters bounds to investigate if the classification error can be further minimized. The SVM hyper-parameter search and the SVM classification engine were implemented in the open source data mining software, namely Rapidminer 5.3 , based on the most widely used and cited SVM package, i.e. the LibSVM which was proposed in the work of (Chang & Lin 2011).

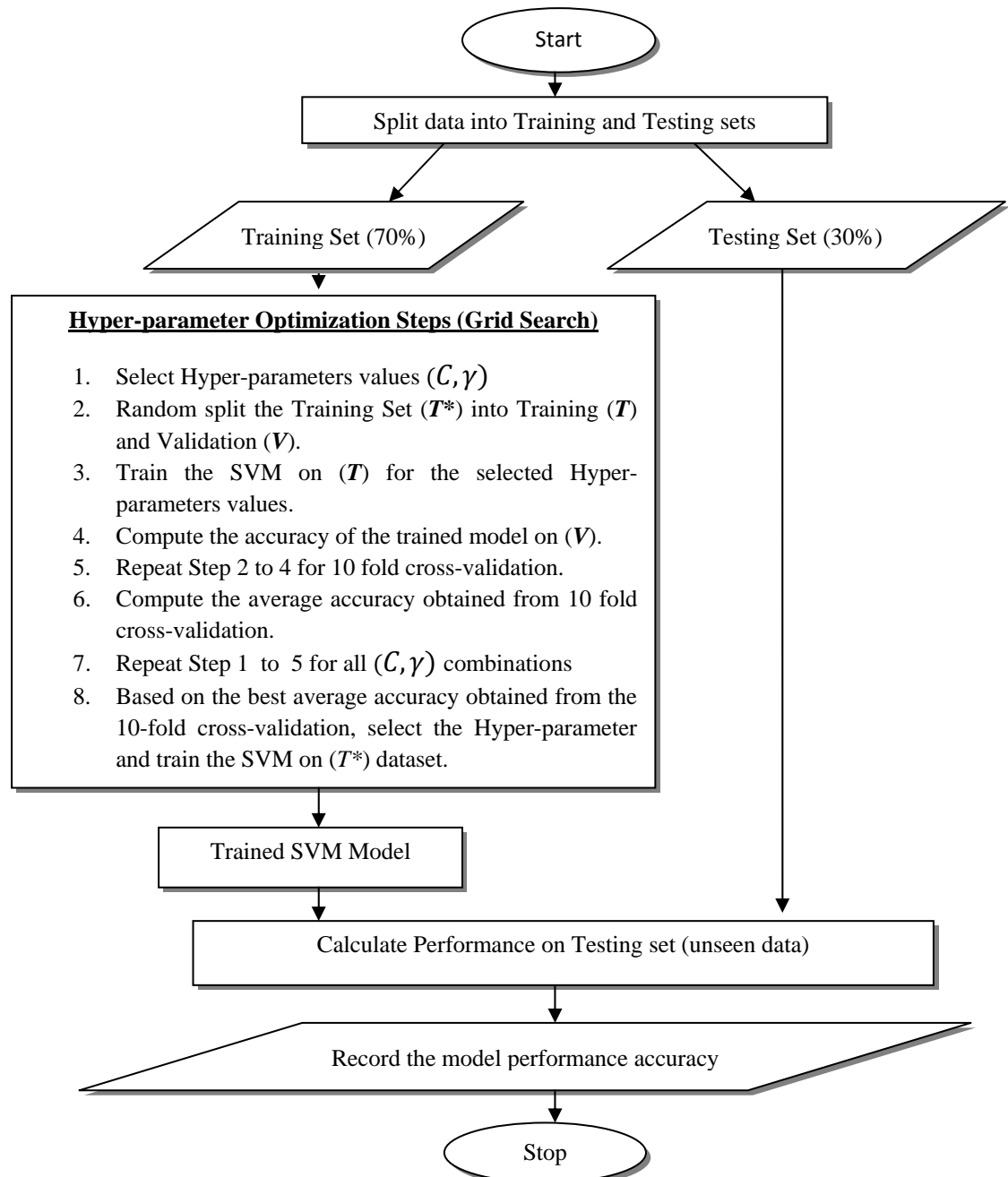


Figure 6.7: SVM model optimization, Training and Testing process

6.4.5 Dataset Balancing

Most classification exercises assume that training samples are evenly distributed among different categories. However, in practical applications, datasets often exist in an imbalanced form. Although the dataset used in this work was not significantly imbalanced (Chawla et al., 2011) where the minority: majority ratio of our dataset was approximately 1:3. However, oversampling the training dataset can consistently provide an improvement in the classification of test data. Moreover, it provides a more stable classifier in case of imbalanced classes. The oversampling technique, namely the Synthetic Minority Over-sampling Technique (SMOTE), was adopted to oversample the minority (*ALL*) class by introducing synthetic samples. SMOTE was solely applied on the *ALL* class in order to increase the sample size and to attain an acceptable number of samples in the majority (*AML*) class. In some cases, SMOTE technique produces an intersection between the classes' boundary, which in turn makes some samples of the majority class resides in the decision region of the minority class. In order to avoid the intersection between the *ALL* and *AML* samples created by the oversampling technique, the *ALL* samples were oversampled at different ratios such that the *ALL* were oversampled at 50%, 100% and 200%, respectively.

6.5 Summary

In this chapter, the final steps and implementation methods for classification of acute leukemia blast cell images were presented. Three different types of features were extracted from each blast cell and its nucleus. In total, 601 features were extracted, as follow: 31 shape-based features, 534 texture-based features and 36 color-based features. The SFS wrapper technique with the 10-fold cross-validation technique was employed to select the optimal set of features.

Two different classifiers, namely MLP-NN trained by back-propagation algorithm and the SVM were used to classify the acute leukemia blast cells. Intensive experiments to optimize the input parameters for both classifiers were carried out in order to choose the best classification model. Although the dataset is not severely imbalanced, the oversampling technique namely SMOTE was adopted for the purpose of studying the effect of oversampling the minority (*ALL*) class at three different rates on the test data.

One of the main contributions of this study is that it proposes a methodology that can assist the hematologists' in diagnosing cases of acute leukemia, by providing a computerized approach of image processing for classifying of blast cells.

CHAPETR 7

RESULTS AND DISCUSSION

7.1 Introduction

In this chapter, we discuss the evaluation results of the implemented CAD-AL, focusing on how the results resolve the problems mentioned in Chapter 1. A diagrammatic view of the test and evaluation results of the proposed CAD-AL is depicted in Figure 7.1.

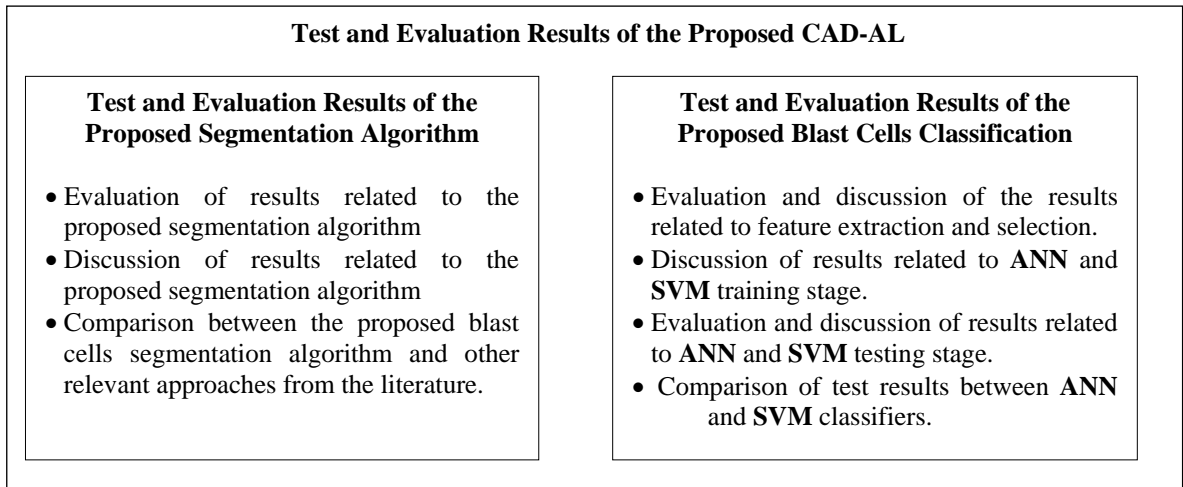


Figure 7.1: Overview of Test Results and Evaluation of the Proposed CAD-AL

7.2 Test and Evaluation Results of the Proposed Blast Cells Segmentation Algorithms

The test and evaluation results of the proposed segmentation algorithms are presented in two parts, Blast cell localization (**BCL**) and the completed blast cell segmentation algorithm (**CBCSA**). The results from both algorithms are described in Sections 7.2.1 and 7.2.2, respectively.

7.2.1 Test and Evaluation Results of the Blast Cell localization algorithm

The **BCL** algorithm presented in Section 5.2 is considered as the starting point towards a more sophisticated algorithm namely the **CBCSA** (Please Refer to Section 5.3).

This algorithm was the result of the initial experimentation on the blast cell segmentation stage carried out at the early stages of the work. The development of the blast cells localization (*BCL*) started right after acquiring the first batch of the PB images. This batch consisted of a 100 *ALL* images that contained 180 expert labelled Lymphoblast cells (**Initial-Dataset**). These 100 *ALL* images were used to test the performance of the localization algorithm. The localization algorithm was able to detect all of the 180 Lymphoblast cells with 100% accuracy. Samples of the localization results can be visually observed in Figure 7.2, 7.3 and 7.4, respectively. These figures below show the original PB image (right) with the processed PB image (left) where the red line represents the boundaries of each localized blast cell.

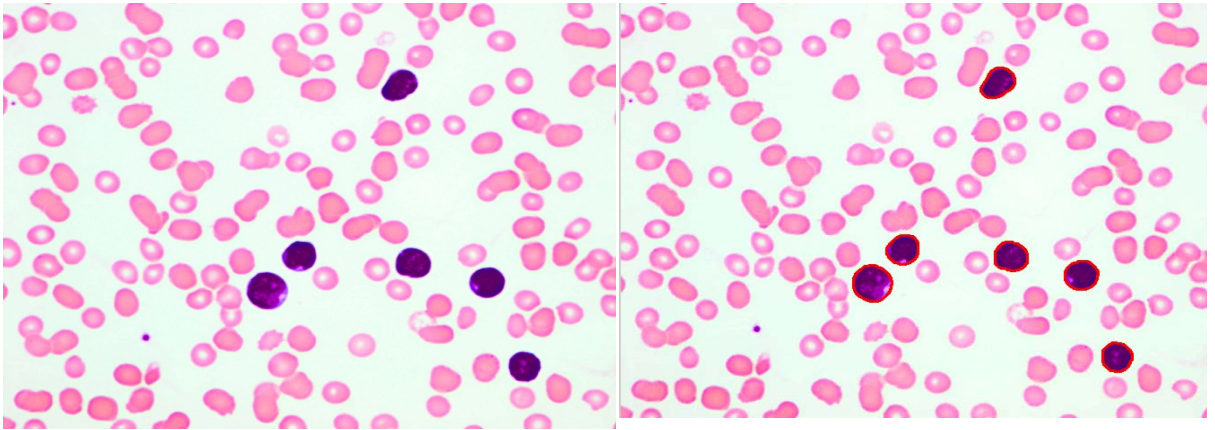


Figure 7.2: Localization of blast cells using *BCL* algorithm (example 1)

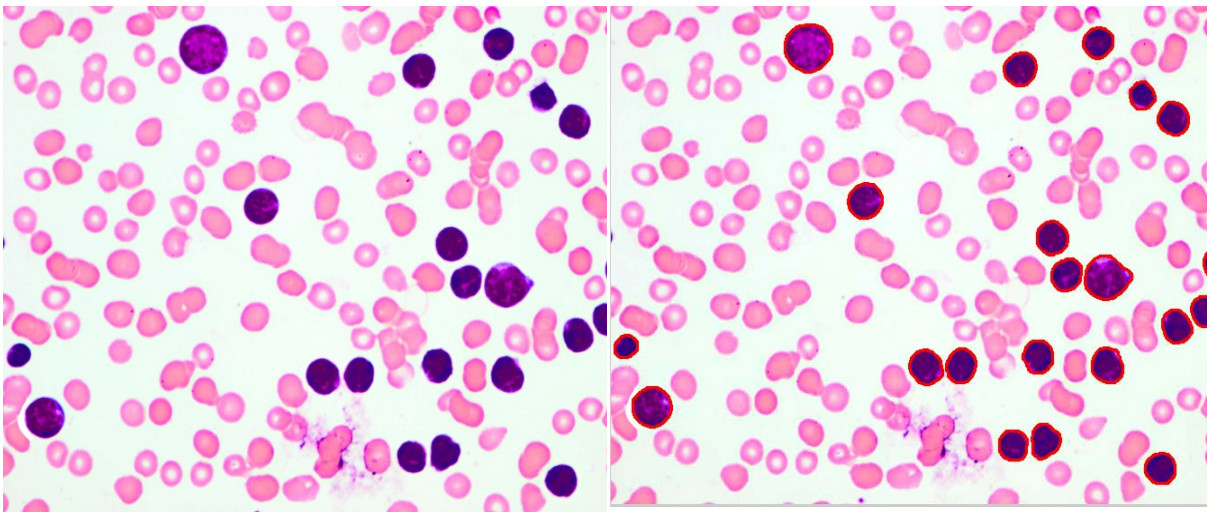


Figure 7.3: Localization of blast cells using *BCL* algorithm (example 2)

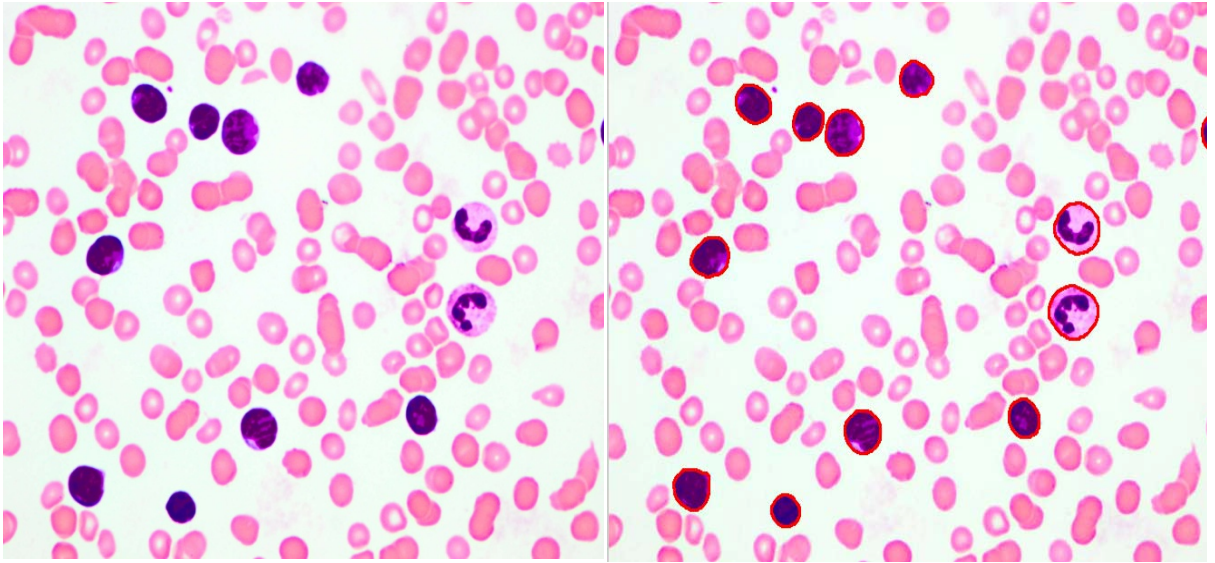
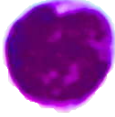











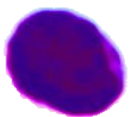
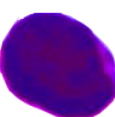


Figure 7.4: Localization of blast cells using **BCL** algorithm (example 3)

In order to measure the accuracy of the proposed **BCL** algorithm in a quantitative manner, a global quantitative method was used. For the purpose of evaluation, the ground truth (*GT*) of all of the 180 lymphoblast cells were localized and segmented manually by a hematologist as described in Section 4.2.3. The performance of the localization algorithm was evaluated by comparing the percentage of the mis-segmented pixels between the proposed localization algorithm and the ground truth. Table 7.1 shows the final result of the evaluation and how it compares to the results of the manual segmentation done by the expert. Table 7.1 only shows part of the testing results, namely, the number of pixels that are classified as cellular in the manual and the proposed method, as well as the error rate and the accuracy rate. The test results show an accuracy rate of 90-95% in restoring the Lymphoblast pixels from the original image. The main sources of error in the proposed **BCL** algorithm were the color inconsistency presented in the acquired PB images as well as the presence of other blood components that were directly touching the Lymphoblast cells. In other words, the difference in color between the other components (erythrocytes and plasma) and the Lymphoblast cells were hardly detected

Table 7.1: Evaluation of the proposed *BCL* Algorithm

Proposed localization algorithm		Ground Truth	
	No. of cellular pixels =12062		No. of cellular pixels =11136
Error Rate =7.677, Accuracy Rate=92.323			
	No. of cellular pixels =12640		No. of cellular pixels =11998
Error Rate =5.079, Accuracy Rate=94.921			
	No. of cellular pixels =10218		No. of cellular pixels =9532
Error Rate =6.714, Accuracy Rate=93.286			
	No. of cellular pixels =10591		No. of cellular pixels =10103
Error Rate =4.608, Accuracy Rate=95.392			
	No. of cellular pixels =11899		No. of cellular pixels =10931
Error Rate =8.135, Accuracy Rate=91.865			
	No. of cellular pixels =9238		No. of cellular pixels =8473
Error Rate =8.281, Accuracy Rate=91.719			
	No. of cellular pixels =7431		No. of cellular pixels =6729
Error Rate =9.447, Accuracy Rate=90.553			

7.2.1.1 Discussion of the Results Related to the Blast Cell Localization

In this research, two blast cells localization approaches have been proposed. The first approach, namely, the **BCL** was developed during the initial experimentation carried out at the early stage of this work. This approach was tested on a 100 **ALL** PB images acquired with a 400x microscope magnification which contained 180 labeled lymphoblast cells. The main goal was to clean up the PB image from the other blood components such as Erythrocytes, Platelets and Plasma while retaining only the blast cells on a white background. As shown in Figure 7.2, 7.3 and 7.4, the **BCL** was able to successfully identify and localize all the Lymphoblast cells in the test PB images. Overall, the **BCL** was able to detect all the 180 labeled Lymphoblast cells, and this finding was correlated with the results obtained from observations performed by the domain expert on the whole set of images. All the 180 localized cells were quantified based on the number of pixels successfully extracted and the results were compared against the ground truth obtained through manual quantification. In every case, the **BCL** algorithm results were able to match the ground truth with an accuracy of 90-95%. The difference in the number of pixels that was considered ‘cellular’ in the proposed method was due to the color inconsistency and blurriness, especially at the border between the Lymphoblast and the Plasma. A possible explanation for blurriness could be due to the low magnification power used to acquire the PB images. Normally a PB image shows a mixture of complex components, which have different intensities, chrominance, and boundary sharpness between all these blood components, which vary from one image to another. For that reason, the localization accuracy also varied based on each image scenario. Nevertheless, the localization accuracy was never below 90%. These findings suggest that the acquisition of PB images with a higher microscopic magnification power may give a better blast cell extraction results.

These findings led us to capture a new set of images comprising of 991 PB images, which we named as *Dataset-A*.

7.2.2 Test and Evaluation Results of the *CBCSA*

Image segmentation of PB images is the foundation for the CAD-AL. Therefore, in this research, an image segmentation based approach was performed. Although the statistical results of the *BCL* algorithm was fairly high, with a minimum segmentation accuracy rate of 90% when compared to the ground truth, we acknowledged that the *BCL* algorithm could not handle some of the newly acquired PB images from *Dataset-A*. This was due to the fact that many of the PB images in *Dataset-A* had color, illumination and staining variability, in addition to cells touching each other due to the heterogeneous cells distributions. To overcome these challenges, a new segmentation approach namely, *CBCSA* was developed (Please Refer to Section 5.3). The *CBCSA* was intensively tested on two different datasets of PB images (*Dataset-A* and *Dataset-B*) acquired from different sources with different specifications. *Dataset-A* was acquired from UMMC, Kuala Lumpur, Malaysia which consisted of 991 PB images with 1303 labeled blast cells, whereas *Dataset-B* was acquired from M. Tettamanti Research Center for childhood leukemia and hematological diseases, Monza, Italy (Labati et al., 2011) and consisted of 108 PB images with 267 labeled blast cells. To visualize the subjective performance of the *CBCSA* results, Figure 7.5 and 7.6 show the *CBCSA* results of *ALL* and *AML* PB samples from *Dataset-A*, respectively. Figure 7.7 shows the segmentation results of the *ALL* PB sample from *Dataset-B*. Each one these three figures is divided into six different parts. Part (a) shows the original sample whereas part (b) depicts the *CBCSA* blast cells localization results where each blast cell contour is highlighted with red color.

Columns (c) and (d) exhibit, the manually extracted blast cell regions and their corresponding manually extracted nuclei (Ground-Truth). While, columns (e) and (f) exhibit the **CBCSA** segmentation results of blast cell regions and their corresponding segmented nuclei.

The three figures illustrate three different scenarios; Figure 7.5 (a) shows an **ALL** PB sample with four solitary blast cells. Figure 7.6 (a) shows a more complex blast cells distribution as the blast cells are touching each other. Although, both PB samples shown in Figures 7.5 and 7.6 were taken from the same dataset (**Dataset-A**), it can be seen that both samples were different in terms of color and illumination, this is because these two samples were acquired from two different PB slides. We found that the condition of the PB images does not solely depend on the acquisition parameter, but it also relies on the quality of the staining procedure. Nevertheless, it can be seen in Figure 7.5 (b) and 7.6 (b) that the proposed **CBCSA** can overcome this variability. Besides, the **CBCSA** has the ability to split touching blast cells individually. Figure 7.7 shows a PB image sample taken from the foreign dataset namely **Dataset-B**, where the acquisition specification and PB slides of this dataset was totally different from the local dataset (**Dataset-A**). The PB image in Figure 7.7 (a) contains 22 blast cells; however, only 5 of these blast cells were numbered for illustration purposes. These 5 blast cells were selected in order to highlight the performance of the **CBCSA** in splitting the touching cells. As shown in Figure 7.7 (b) the cells labeled with the numbers (1-2) and (3-4-5) are touching each other. The boundaries of these blast cells are highlighted with a red color line in the processed image as shown in Figure 7.7 (b). The PB image presented in Figure 7.7 shows an even more complex scenario with smaller blast cells than what was depicted in Figures 7.5 and Figure 7.6.

The complexity of the image and size of the blast cells was attributed to the low microscope magnification power used to acquire the image. Moreover, the number of blast cells touching each other is higher than the PB image in Figure 7.6.

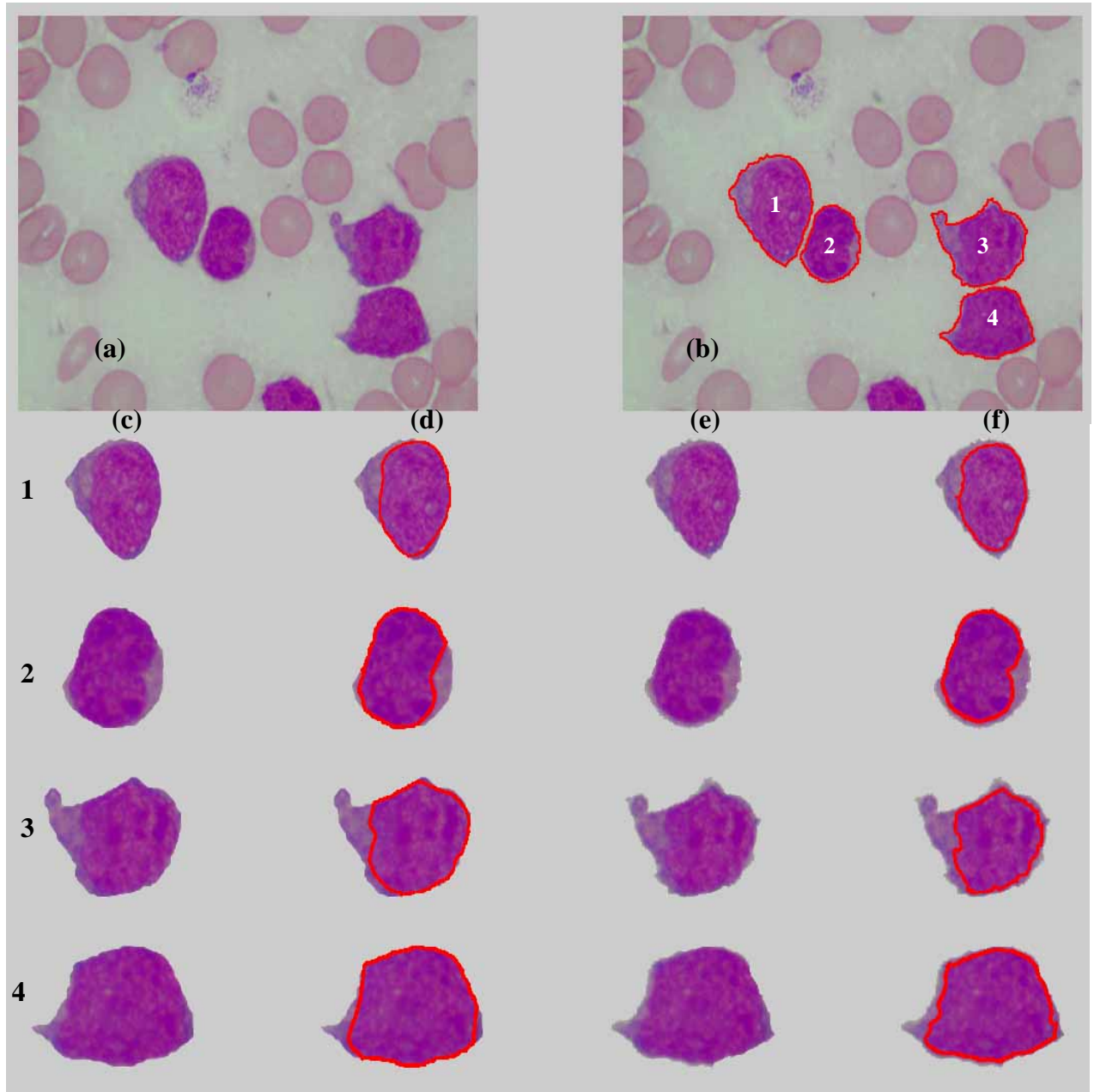


Figure 7.5: Experimental results of the proposed segmentation approach for *ALL* PB sample (a) original image (b) blast cells localized using *CBCSA* (c) Ground-truth of blast region (d) Ground-truth of nucleus region (e) Blast region obtained using the *CBCSA* (f) Nucleus region obtained using the proposed *CBCSA* approach

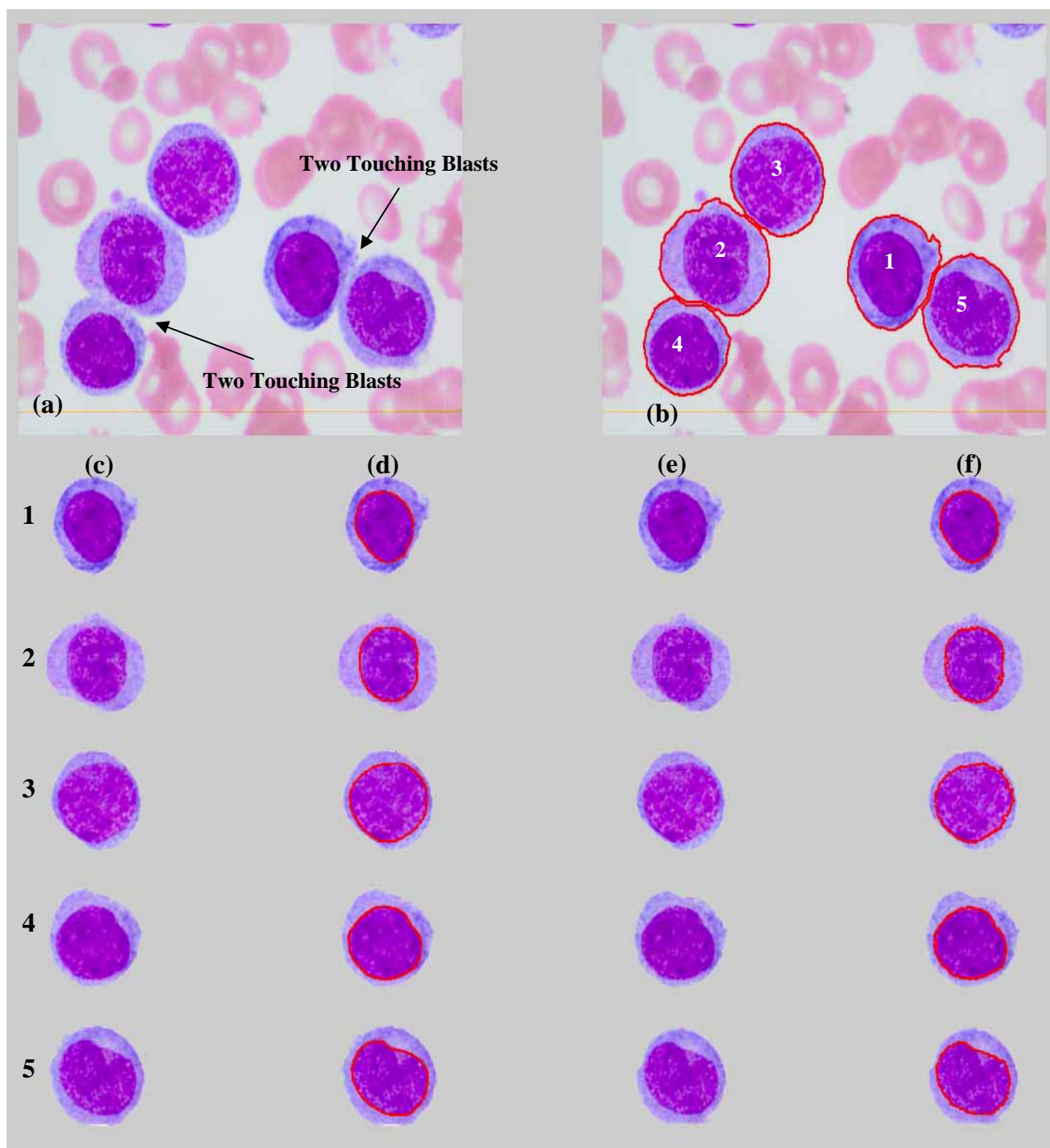


Figure 7.6: Experimental results of the proposed segmentation approach for *AML* PB sample (a) original image (b) blast cells localized using *CBCSA* (c) Ground-truth of blast region (d) Ground-truth of nucleus region (e) Blast region obtained using the *CBCSA* approach (f) Nucleus region obtained using the *CBCSA* approach

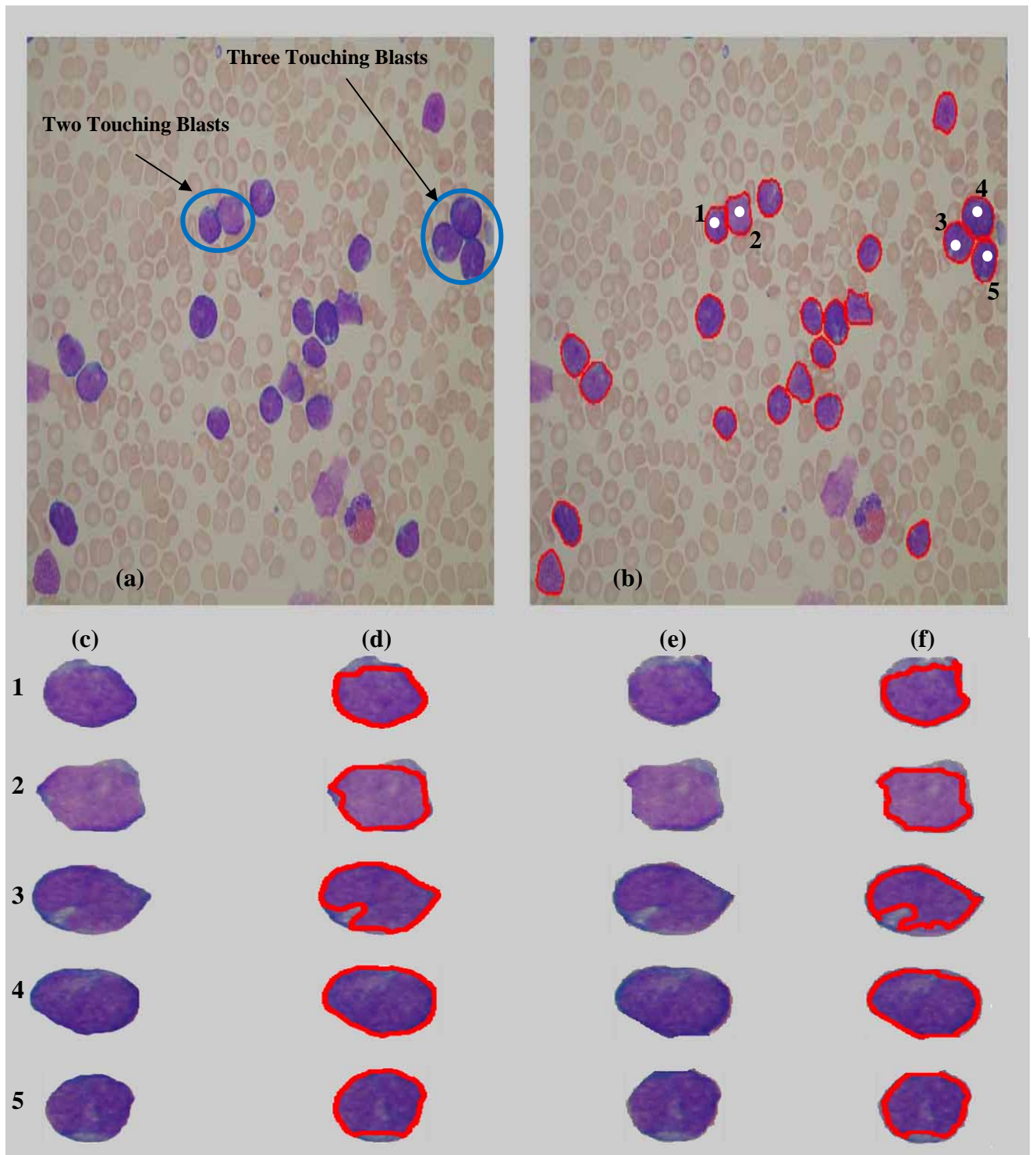


Figure 7.7: Experimental results of the proposed segmentation approach for *ALL* PB sample from Dataset-B where cells (1-2) and (3-4-5) are touching each other (a) original image (b) localized blast cells using *CBCSA* (c) Ground-truth of blast region (d) Ground-truth of nucleus region (e) Blast region obtained using *CBCSA* (f) Nucleus region obtained using *CBCSA*

The segmentation results presented in Figure 7.7 (b) corroborate the findings that the performance of the proposed **CBCSA** is invariant to color; illumination, staining and microscope magnification as well as having the ability to completely split the touching cells.

To visually compare each individual extracted blast cell region and its interrelated nuclear region against the ground truth in each Figure above, columns (c) and (d) were used to depict the manually extracted blast cell region and its nuclear region's border, respectively while columns (e) and (f) were used to depict the blast cells regions and its nuclear region's border extracted using the proposed **CBCSA**. Each row in column (c, d, e, f) are numbered according to the label given to each blast cell in part (b) of each Figure. Due to space limitation, only the results of five of the localized blast cells in Figure 7.7 were illustrated here for visual comparison. These blast cells were highlighted (Please Refer to Figure 7.7 (b)) with a white solid dot and extracted individually.

As the visual comparison between the segmented images produced by the **CBCSA** and the ground truth was not an objective nor a precise evaluation, we employed two well-known image segmentation evaluation protocols, that were highly ranked in Zhang's survey (Zhang,1996) namely the **RUMA** and the **ME** (Please Refer to Section 4.8.1) to evaluate our segmentation algorithm. The main goal of the **CBCSA** was to segment the blast cell and its nucleus in order to produce a maximum accuracy rate in terms of the segmented regional size. Hence, two features, namely the area (**A**) and the perimeter (**P**) of the segmented region were used in **RUMA** (represented as **RUMA_A** and **RUMA_P**), these two features were directly related to the size of the tested region. In **ME**, the numbers of the mis-segmented (cellular pixels classified as a background and vice versa) pixels in the tested region were compared to its manually segmented ground truth.

The differences between the ground truth and the outcome of the cells after using **CBCSA** for the blast cells (c-e) and the nuclear regions (d-f) in Figures 7.5, 7.6 and 7.7 were calculated using the **RUMA** and **ME**. These are shown in Tables 7.2, 7.3 and 7.4, respectively.

The quantitative evaluation of **RUMA** and **ME** are inversely proportional to the segmentation quality, meaning that the smaller the value the better the result. Hence, a value of 0 will result if both the manual and the segmented cell are completely identical, and a value of 100 will result if the segmented region is not detected.

Table 7.2: The difference between **GT** and the results of **CBCSA** for blast cells of Figure 7.5.

Cell Number	Blast region			Nucleus Region		
	$RUMA_A(\%)$	$RUMA_P(\%)$	$ME(\%)$	$RUMA_A(\%)$	$RUMA_P(\%)$	$ME(\%)$
1	4.65	1.48	3.46	3.70	0.54	2.59
2	6.00	2.02	5.45	0.15	1.40	3.16
3	2.89	0.05	4.70	5.82	1.48	4.05
4	1.42	3.19	5.64	4.76	1.45	6.62

Table 7.3: The difference between **GT** and the results of **CBCSA** for blast cells of Figure 7.6.

Cell Number	Blast region			Nucleus Region		
	$RUMA_A(\%)$	$RUMA_P(\%)$	$ME(\%)$	$RUMA_A(\%)$	$RUMA_P(\%)$	$ME(\%)$
1	3.33	1.26	2.50	0.98	0.58	0.81
2	0.21	1.49	2.11	0.61	4.91	1.00
3	1.55	0.61	1.47	0.61	6.50	2.33
4	0.23	0.62	2.06	1.64	1.76	1.20
5	0.14	1.13	4.03	2.08	1.27	2.92

Table 7.4: The difference between **GT** and the results of **CBCSA** for blast cells of Figure 7.7.

Cell Number	Blast region			Nucleus Region		
	$RUMA_A(\%)$	$RUMA_P(\%)$	$ME(\%)$	$RUMA_A(\%)$	$RUMA_P(\%)$	$ME(\%)$
1	9.52	9.36	7.39	5.65	5.99	5.74
2	4.43	4.23	5.99	7.39	3.85	7.06
3	3.47	2.81	4.41	6.86	4.41	7.97
4	2.97	5.26	4.14	3.72	1.21	4.32
5	1.69	5.41	5.37	5.69	2.84	5.79

This research involved 9 different blast cell subtypes (*ALL*: L1-L2-L3, *AML*: M1-M2-M3-M4-M5-M7). Thus, the difference between the ground truth and the result of *CBCSA* for all images from both datasets (*A* and *B*) was first computed for each blast cell subtype and its nucleus and then averaged over all the blast cells from each subtype. Later, the overall difference was obtained by averaging the results of all subtypes. The standard deviation was also computed because a systematic error with a small standard deviation is considered less harmful than an error with a large standard deviation. The final mean and standard deviation for the difference of each blast type and subtype are presented in Table 7.5.

Table 7.5: Mean \pm standard deviation for difference (%) between *GT* and *CBCSA* segmentation results for all sub-images extracted from *Dataset-A* and *Dataset-B*

Blast Cells Types	Blast Region			Nucleus Region		
	<i>RUMA_A</i> (%)	<i>RUMA_P</i> (%)	<i>ME</i> (%)	<i>RUMA_A</i> (%)	<i>RUMA_P</i> (%)	<i>ME</i> (%)
	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
<i>ALL</i> (L1-L2)	3.50 ± 4.31	2.86 ± 2.85	3.35 ± 3.47	3.02 ± 3.82	3.19 ± 2.42	3.20 ± 3.60
<i>ALL</i> (L3)	3.19 ± 2.31	1.82 ± 1.94	2.96 ± 1.55	8.80 ± 9.51	3.82 ± 7.22	4.70 ± 3.85
Overall (<i>ALL</i>) $\mu \pm \sigma$	3.35 ± 0.22	2.34 ± 0.74	3.16 ± 0.27	5.91 ± 4.08	3.51 ± 0.45	3.95 ± 1.05
<i>AML</i> (M1)	4.43 ± 3.38	1.95 ± 1.31	5.30 ± 2.70	3.74 ± 2.24	3.17 ± 1.55	4.85 ± 2.72
<i>AML</i> (M2)	4.12 ± 2.53	1.50 ± 2.33	3.59 ± 2.58	3.00 ± 2.81	2.25 ± 2.74	3.82 ± 2.50
<i>AML</i> (M3)	6.54 ± 3.97	2.43 ± 3.77	3.98 ± 1.89	4.22 ± 4.93	2.92 ± 5.26	3.55 ± 2.91
<i>AML</i> (M4)	1.31 ± 1.15	0.96 ± 0.75	1.93 ± 0.92	2.54 ± 4.67	2.80 ± 2.12	1.99 ± 2.06
<i>AML</i> (M5)	3.73 ± 2.59	1.35 ± 1.95	2.71 ± 2.08	2.78 ± 3.80	3.27 ± 1.75	2.76 ± 2.57
<i>AML</i> (M7)	8.33 ± 8.30	3.32 ± 3.93	4.93 ± 4.22	5.68 ± 7.57	4.63 ± 4.59	4.93 ± 4.78
Overall (<i>AML</i>) $\mu \pm \sigma$	4.74 ± 2.42	1.92 ± 0.85	3.74 ± 1.28	3.66 ± 1.17	3.17 ± 0.79	3.65 ± 1.15
<i>ALL</i> $\mu \pm \sigma$	3.46 ± 2.03	2.27 ± 2.05	4.14 ± 1.54	7.19 ± 4.46	6.77 ± 3.53	5.28 ± 1.59
Overall (<i>Dataset-A</i> and <i>Dataset-B</i>) $\mu \pm \sigma$	3.97 ± 0.98	2.18 ± 0.22	3.76 ± 0.52	5.59 ± 1.78	4.48 ± 1.98	4.29 ± 0.86

7.2.2.1 Discussion of Results Related to the *CBCSA*

The *CBCSA* performance test was conducted on two different dataset of images namely *Dataset-A* and *Dataset-B*. *Dataset-A* consisted of 991 high-resolution PB images (1303 blast cells) collected from the UMMC, Kuala Lumpur, Malaysia (Please Refer to Table 4.2). Whereas *Dataset-B* contained 108 PB images, 48 of which contained blast cells and the total number of labeled blast cells were 267. The ground truth of each blast cell was drawn manually under the supervision of the domain expert in UMMC (Prof. Hany Ariffin). It is worth mentioning that testing the *CBCSA* segmentation performance using PB images from different sources is beneficial. Testing a segmentation algorithm with images collected from different sources while obtaining highly accurate results, indicates that the performance of the segmentation algorithm is robust and reliable in the real-world application scenario.

Samples of the *CBCSA* experimentation results are visually presented in Figure 7.5, 7.6 and 7.7. The difference between the ground truth and the segmentation results of the localized blast cells and their nuclei of Figure 7.5, 7.6 and 7.7 are presented in Table 7.2, 7.3 and 7.4, respectively. The mean and standard deviation of the segmentation results of all the blast cells in *Dataset-A* and *Dataset-B* are given in Table 7.5.

The *CBCSA* shows good segmentation results, allowing the extraction of almost 1566 complete blast cells with their respective nuclei from 1024 different PB images at average error rate ($RUMA_A$) of 3.97% and 5.59% for the blast cell and the nucleus region respectively.

Considering the quantitative results presented in Table 7.2 which belong to the *ALL* sample presented in Figure 7.5 (a) and the results in Table 7.3, which belong to the *AML* sample presented in Figure 7.6 (a), it can be noted that the difference between the blast cell's

regions and the nuclear regions using **RUMA** and **ME** was generally higher in the **ALL** (maximum error rate 4.65%) than in the **AML** (maximum error rate 3.33%) sample.

This can happen when the Lymphoblast has a relatively high ratio of cytoplasm adjacent to the plasma. In general, in most of the **ALL** PB images, the intensity contrast between the cytoplasm and the plasma is low. Nevertheless, this is not always the case as Lymphoblast cells have a larger nuclear region adjoined to the plasma compared to the cytoplasm region, where the intensity contrast between the nuclear region and the plasma is high. For that reason, the extraction of the blast cell's region in **ALL** samples shows better results than in the **AML** samples (Cell number 2 in Figure 7.5 (b) with maximum error rate at 0.15) when most of the Lymphoblast boundary is occupied with nucleus. Touching cells are presented in the PB images as shown in Figures 7.6 (a) and 7.7(a). Figures 7.6 (b) and 7.7 (b) illustrate how the **CBCSA** can efficiently separate these touching cells with a high degree of accuracy such as cells number 2 and 4 in Figure 7.6 (b) where the maximum error rate according to **RUMA_A** was at 0.21% and 0.23% respectively . As shown in Table 7.3, which reflects the difference rate of the blast cells extracted of Figure 7.6, it can be seen that the maximum difference rate for the blast cell's region based on **RUMA_A** is 3.33% for the blast cell labeled with number (1) while the highest difference rate based on **ME** is 4.03% for the blast cell labeled with number (5). Table 7.4 demonstrates the extraction difference rate of 5 blast cells (highlighted with white solid dots in the center) in Figure 7.7 (b). It shows a higher difference rate for the blast cell's region based on **RUMA_A** and **ME** as compared to both samples of Figure 7.5(a) and 7.6 (a).

The difference rates produced by comparing the **CBCSA** against the ground truth using two segmentation evaluation metrics namely **RUMA** and **ME** are presented in Table 7.5. The blast cells in **Dataset-A** consisted of two acute leukemia types namely **ALL** and **AML**.

The blast cells from each type were further categorized by the domain expert in UMMC, Kuala Lumpur, Malaysia according to their subtypes (Please Refer Table 4.2) where each one of these acute leukemia subtypes was acquired from different PB slide. This is because an acute leukemia patient is usually diagnosed with one acute leukemia type and one subtype under the main type. **Dataset-B** contained only one type of blast cells, namely, the **ALL** without any categorization of the subtype's details (Labati, et al., 2011).

The main objective was to test the robustness of the proposed **CBCSA**, using two distinctive datasets acquired under different conditions. All the PB images in **Dataset-A** were captured using the same acquisition criteria. The segmentation results obtained by testing the **CBCSA** performance on **Dataset-A** was detailed to the subtype's level. The main intention was to get a clear figure of the segmentation accuracy not only for PB images collected using different acquisition criteria, but also for PB images collected from different PB slides using the same acquisition criteria, considering that there were many factors that could affect the quality of the acquired images such as storage time of the slide, staining quality, the quality of the PB smear preparation using the manual push-pull method with a spreader slide performed by a laboratory technologist. Moreover, the blast cells belonging to each acute leukemia subtype has different characteristics in terms of shape, texture, color, nucleus chromatic density, etc. (Please Refer Table 2.5 and Table 2.6).

As shown in Table 7.5, the **CBCSA** properly extracted the blast cell's region of the three **ALL** subtypes (L1-L2-L3), and the average extraction discrepancies in **RUM_A**, **RUM_P** and **ME** on 269 PB images containing 325 labeled blast cells were 3.35%, 2.34% and 3.16%, respectively. This is because the region that is adjacent to the surrounding blood components such as erythrocytes and plasma is mostly the nucleus, and the nucleus intensity contrast is significantly different from other adjoining blood components.

However, the average difference rate of *ALL* (L1-L2-L3) nuclear regions in *RUMA_A*, *RUMA_P* and *ME* were reduced to 5.91%, 3.51% and 3.95%, respectively. This was mainly caused by the deep-blue color cytoplasm of the (L3) subtype which is analogous to the color of the nucleus in many of the L3 samples (Estey et al., 2007).

Regarding the *AML (Dataset-A)* blast cells extraction, there are a few discussable points that can be inferred from Table 7.5. The overall average difference for the *AML* blast cell's region in *RUMA_A*, *RUMA_P* and *ME* were 4.74%, 1.92%, and 3.74% respectively. Generally, the result of the blast cell's region segmentation accuracy for the *AML* class was lower than that of the *ALL* class. The main reason was considerably related to the segmentation of the two *AML* subtypes namely M3 and M7 where the difference rate in *RUMA_A* for both were above 6%. As shown in Figure 7.8 (a) the Erythrocytes color is analogous to the M3 cytoplasm color. This is because the M3 cytoplasm region was stained as bright pink or red (Abdul-Hamid, 2011). In some other cases the M3 cytoplasm had a vitreous color as illustrated in Figure 7.8 (b). These two situations presented in Figure 7.8 (a-b) lead to low image gradient values along the cell edges. Therefore, the segmentation error for these *AML* subtypes was larger than the others. In fact, an effective separation between the WBC in general and adjacent Erythrocytes is the main problem for most segmentation methods (Bergen et al., 2008). This finding was highlighted in the works of (Won, 2005) and (Bergen et al., 2008) who found that the similarity in color and illumination of the touched Erythrocytes with the cytoplasm in the leukocyte (healthy WBC) makes their separation rather difficult. However, to our knowledge, this finding has not been previously addressed by many works which dealt with blast cell (unhealthy WBC) segmentation.

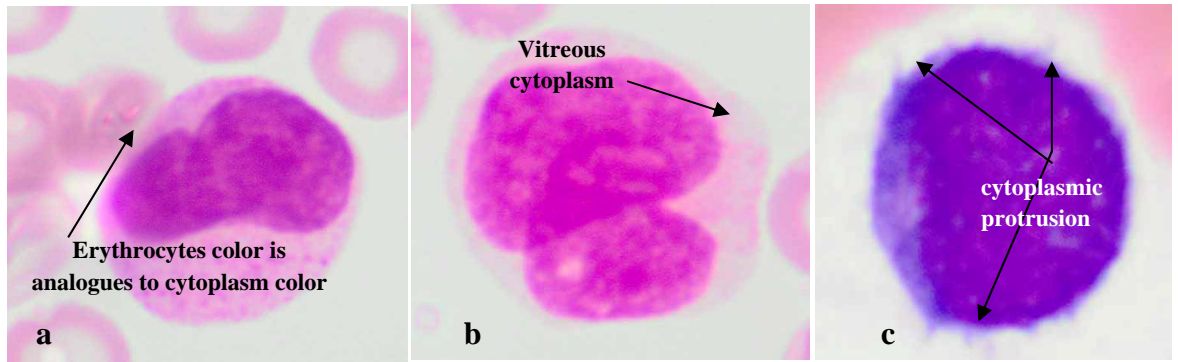


Figure 7.8: Blast region segmentation difficulties in *AML*. (a) Erythrocytes color is analogues to M3 cytoplasm color, (b) M3 blast with vitreous cytoplasm, (c) M7 with protrusion cytoplasm.

Concerning *AML* (M7), this subtype shows protrusion or budding (“blebs”) at the edge of the cytoplasm (Abdul-Hamid, 2011) as shown in Figure 7.8 (c). This cytoplasmic protrusion has not been detected properly. This is mainly caused by the smoothing process carried out using the median filter applied on the *b* channel image as well as by applying the morphological opening operation on the binary version of the **Hue** channel for the purpose of smoothing the blast cells contour and removing pixels that caused two objects to be touched by a thin connection (Please Refer to Section 5.3.1). However, these two smoothing operations contributed to the segmentation of the other blast cell subtypes tremendously. On the other hand, the *CBCSA* segmented the nuclear region of *AML* blast cells with high accuracy as all the *AML* subtypes had considerably distinguishable intensity contrasts between the nucleus and the cytoplasm, which can be easily be detected using the SRG algorithm. However, the M7 nuclear region showed lower segmentation accuracy than other *AML* subtypes. A possible explanation for this might be that the M7 is a Lymphoblast-like (National Cancer Institute, 2014) and some of the M7 blast cells have a dark cytoplasm region such as that in the *ALL* (L3), which make the separation between the nucleus and cytoplasm of this subtype difficult and error prone. Nevertheless, the overall average difference rate of the *AML* nuclear region in *RUMA_A*, *RUMA_P* and *ME* were only 3.66%, 3.17% and 3.65%, respectively. These results are quite promising.

Table 7.5 also reports the **CBCSA** performance on the blast cells from **Dataset-B**. As mentioned earlier, **Dataset-B** contained only the **ALL** blast cells without subtypes. The difference rate of the **ALL** blast cell's region from **Dataset-B** is quite comparable to that of **Dataset-A**. However, the nuclear region segmentation accuracy for the **ALL** in **Dataset-B** was lower than that of both types of blast cells in **Dataset-A**. This was mainly due to the low microscope magnification power used to acquire **Dataset-B** images, where the cellular details were not adequately distinguishable. As the magnification power decreases, the field of view appears bigger; however, the blood cells become smaller causing a decrease in the cellular details. The lower magnification objective lens is used for low-power observations, whereas the high magnification power enables cellular details to be observed (Addis & Larkcom, 1999).

The overall average difference rate of the proposed **CBCSA** applied on 1024 high-cell population PB images acquired from two different sources, was rather encouraging as the proposed algorithm was able to localize approximately all the blast cells presented. The overall average segmentation accuracy (100-extraction discrepancy) of all the localized blast cells from both datasets in **RUMAs** and **ME** for the blast cells and the nuclear region were as high as 96% and 94%, respectively.

The segmentation results and the related findings obtained by the proposed **CBCSA** answered the questions (b, e and g) posed in Section 1.4 clearly. The main three problems mentioned in Section 3.6.2 were properly addressed by the proposed **CBCSA**, and the presented results corroborated that **CBCSA** can efficiently segment blast cells regardless of the PB image variation in term of color, illumination, staining quality and also shape, texture, color of the blast cells.

Several points are inferred from the whole segmentation process as follow:

- It has been stated by (Díaz, & Manzanera, 2011) that for proper visual analysis of blood smear; images must be captured using a high microscopic magnification power. The current research found that high microscopic magnification power is not only beneficial for visual analysis, but also for image analysis. This finding confirms the association between high magnification power used to capture the images and the image segmentation accuracy.
- Variegation of the color, illumination and staining quality, in PB images and the diversity found in the shape, texture and color of the blast cells and its internal components makes blood image segmentation a challenging process. For that reason a direct segmentation technique is not adequate when segmenting acute leukemia blast cells.
- With proper choice of color channels, the proposed method can produce results that are highly accordant with those manually segmented by the domain expert.
- The localization and sub-imaging is highly desirable for better segmentation of the blast cell's internal components (i.e nucleus and cytoplasm).
- The success of segregating touching and overlapping blast cells is highly dependent on the overlapping degree. It has been found that watershed segmentation with distance transform can split adjacent blast cells overlapped with low to medium degree. However, a more sophisticated method is needed to split the blast cells that overlapped with high degree (Please Refer to Figure 7.11 (d-e)) such as marker-controlled watershed. More details about this point will be given in the next section.

7.3 Comparison with Other Blast Cell Segmentation Methods

We have shown in Table 7.5 that the developed segmentation algorithm produced robust, reliable and accurate results in comparison with the ground truth images segmented manually by domain expert from UMMC, Kuala Lumpur, Malaysia. The domain expert usually performs the manual morphological examination of the PB slide under the microscope. The results demonstrate that the proposed method, in spite of its simplicity, with a proper choice of suggested color channels is highly competitive with other well-known blast cells segmentation methods such as the works by (Putzu & Ruberto, 2013; Putzu & Ruberto, 2013a; Putzu & Ruberto, 2013b; Reta et al., 2010, Escalante et al, 2012; Sadeghian et al., 2009; Scotti, 2005). The proposed segmentation algorithm (**CBCSA**) addresses three main problems related to blast cells segmentation. These problems are (1) the localization and sub-imaging, (2) the segregation of touching cells and (3) diversity in color, illumination, staining variation. To our knowledge, most of the blast segmentation methodologies presented in the literature (Mohaptra et al., 2013; Escalante et al., 2012; Madhloom et al., 2012, Nee et al., 2012; Huey Nee 2012; Reta et al., 2010; Sadeghian et al., 2009) did not address these three problems simultaneously. Many researchers implemented their segmentation method directly on manually cropped sub-images which contained a single blast cell such as the work by (Reta et al., 2010; Escalante et al, 2012; Sadeghian et al., 2009). On the other hand, some other researchers proposed a solution to localize and crop each blast cell presented in high-cell population blood images such as the work by (Mohaptra et al., 2013; Madhloom et al, 2012). However, the problem of splitting touching cells was not considered. Only a few methods have reported both problems (1 and 2) such as the work by (Nee et al., 2012; Huey Nee et al., 2012).

Nevertheless, these methods proposed by (Nee et al., 2012; Huey Nee 2012) failed to completely localize blast cells causing inaccuracies in retaining the cytoplasm region and in some cases the whole cytoplasm region was missing. To our knowledge, none of the methods presented in the literature (Putzu & Ruberto, 2013; Putzu & Ruberto, 2013a; Putzu & Ruberto, 2013b; Mohaptra et al. 2013, Madhloom et al, 2012, Escalante et al, 2012, Nee et al., 2012; Huey Nee 2012; Reta et al., 2010, Sadeghian et al., 2009) considered validating the accuracy of their proposed segmentation methods on a second dataset of images that was acquired under different conditions from the main datasets. This could be because of the lack of medical data which is an intrinsic problem in research involving medical data (Mennicke et al., 2009) in the real-world scenario. Moreover, most of the previously proposed segmentation algorithms were not tested quantitatively against ground truth images segmented manually by a domain expert as the lack of a ground truth which require a domain expert to be involved is another intrinsic problem in research involving medical data (Babalola et al., 2008) in the real-world scenario.

The proposed segmentation method is compared with two state-of-the-art blast cell segmentation methods proposed by (Putzu & Ruberto, 2013; Putzu & Ruberto, 2013a; Putzu & Ruberto, 2013b) and (Scotti, 2005). The localization and segregation of touching blast cells are only compared with (Putzu & Ruberto, 2013) as the method presented by (Scotti, 2005) did not provide a solution to the issue of splitting the touching blast cells. However, the performance of nucleus/cytoplasm separation algorithm presented in the current research is compared with both. The reason for choosing these two methods as a benchmark is that, both of these methods used the same dataset reported in (Labati et al., 2011), which is available online at (<http://www.dti.unimi.it/fscotti/all/>) named as ***Dataset-B*** in our research.

We used this dataset for the purpose of testing the robustness and reliability of our proposed method. Besides, the method proposed by (Putzu & Ruberto, 2013) provided a scheme for localization and sub-imaging, splitting the touching blast cells as well as nucleus and cytoplasm separation. These reasons make the work by (Putzu & Ruberto, 2013) a good benchmark to be compared with. The performance of the benchmark (Putzu & Ruberto, 2013) method was tested with the first 33 PB images from ***Dataset-B*** which contained 267 labeled cells. The performance was reported based on the ability to localize and count the number of cells presented in the PB images. Some of these images had a high-cell population with many clumped blast cells such as Image005 which contained 24 ***ALL*** blast cells. We performed the same operation using the proposed ***CBCSA*** on these 33 PB images. The performance comparison between the ***CBCSA*** and the benchmark is reported in Table 7.6.

Table 7.6: Performance comparison between the proposed *CBCSA* and the benchmark (Putzu & Ruberto, 2013)

		(Putzu, & Ruberto, 2013)	Proposed <i>CBCSA</i>	((Putzu & Ruberto, 2013))	Proposed <i>CBCSA</i>
Image No.	Number of Manually localized cells	Number of localized cells	Number of localized cells	Accuracy (%)	Accuracy (%)
Image001	9	5	9	55	100
Image002	10	10	10	100	100
Image003	12	11	12	91	100
Image004	7	4	7	57	100
Image005	24	19	23	79	95
Image006	18	18	18	100	100
Image007	7	7	7	100	100
Image008	17	16	17	94	100
Image009	7	7	7	100	100
Image010	12	12	12	100	100
Image011	15	12	15	80	100
Image012	12	12	12	100	100
Image013	10	7	10	70	100
Image014	5	3	5	60	100
Image015	17	17	17	100	100
Image016	16	16	16	100	100
Image017	3	3	3	100	100
Image018	8	8	8	100	100
Image019	12	12	9	100	75
Image020	2	2	2	100	100
Image021	3	3	3	100	100
Image022	5	5	5	100	100
Image023	6	6	6	100	100
Image024	4	4	4	100	100
Image025	3	3	3	100	100
Image026	5	5	5	100	100
Image027	3	3	3	100	100
Image028	2	2	2	100	100
Image029	4	4	4	100	100
Image030	3	3	3	100	100
Image031	2	2	2	100	100
Image032	2	2	2	100	100
Image033	2	2	2	100	100
Total no. of localized cells	267	245	263		
Overall Accuracy				92%	99%

From 267 labeled cells, the proposed **CBCSA** was able to locate 263 cells resulting in an accuracy rate of 99% whereas the benchmark was only able to locate 245 cells with an accuracy rate of 92%.

The total number of the missing cells by the **CBCSA** was four; one cell was from Image005 (Figure 7.9 (a-b)) and the other three from Image019 ((Figure 7.10 (a-b))).

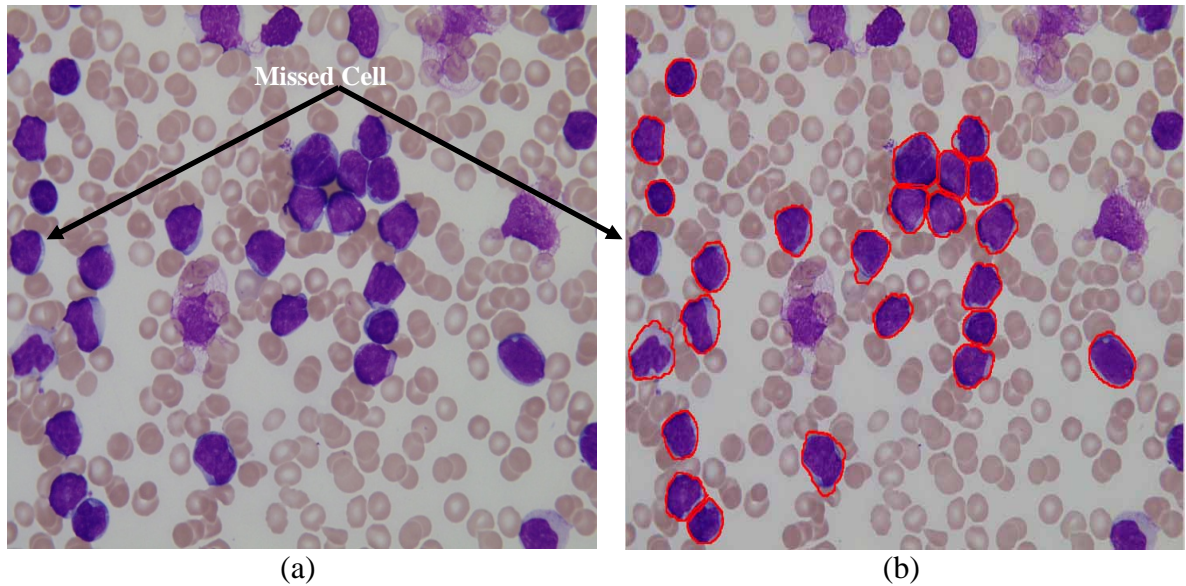


Figure 7.9: Segmentation result of Image005 (a) Original image, (b) Segmented image with blast cells border overlaid by red line

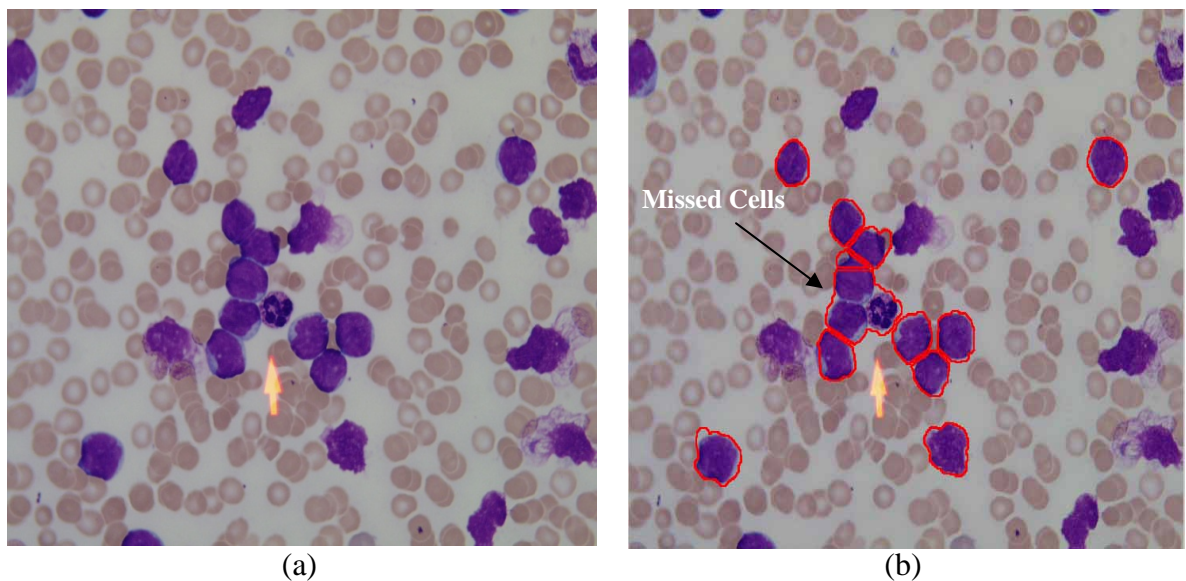


Figure 7.10: Segmentation result of Image019 (a) Original image, (b) Segmented image with blast cells border overlaid by red line

The cell in Image005 was missed (Please Refer to Figure 7.9 (a-b)) because it was very close to the image border, while the other three cells in Image019 were missed (Please Refer to Figure 7.10 (a-b)) because they overlapped with a high degree, such as in the scenario presented in Figure 7.11 (d-e) and the watershed transform based on the distance transform was not able to completely separate them. The touching cells can present in various forms such as chain, cluster or ring of touching cells as shown in Figure 7.11 (a-c). In these cases, the overlapping degree is low meaning that the cells form a concave region where they overlap. Using the distance transform, the connected set of central pixels (marker) can be easily found for each cell. Here is where the water starts to enter and rise as the distance from these central pixels to the pixels at the edge are the maxima of the distance transform. However, if the cells form clumps or rings and the internal holes are filled as shown in Figure 7.11 (d-e), the watershed transform will not be able to separate the cells properly as the maximum distance is greater than central pixels of the cells. For this reason, the watershed transform produces over-segmented or under-segmented regions as in the case of Image019. Other methods such as concavity analysis, which was used in the work of (Reta et al., 2010) may not give a proper splitting solution for scenarios such as of Figure 7.11 (d-e), as the concavity analysis requires that the individual cells are convex and that they form a concave region where they overlap (Bailey, 1992).

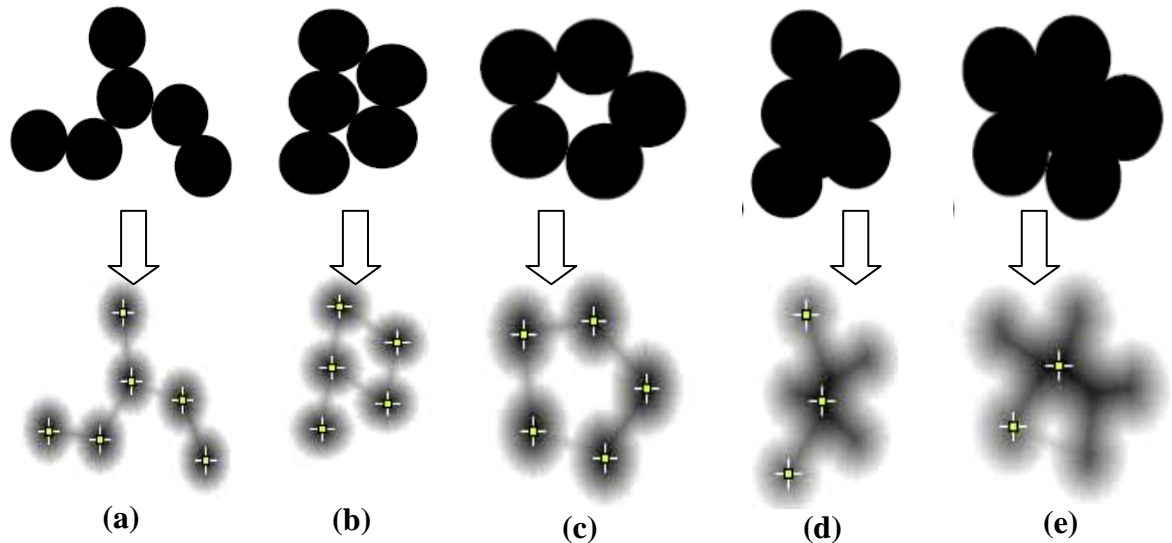


Figure 7.11: Various scenarios of touching (overlapping) cells. (a) Chain of cells, (b) Cluster, (c) Ring, (d) Cluster with filled holes, (e) Ring with filled holes.

In our work, there were 228 identified touching cells with different overlapping scenarios. We compared the watershed based on the distance transform with a marker-controlled watershed, as we proposed to use the nuclear region as a marker instead of the distance transform regional maxima where this approach can tremendously reduce the over-segmentation (under-segmentation) (Felkel et al, 2001). As shown in Table 7.7, we found that the marker-controlled watershed based on nucleus marker outperformed the watershed based on the distance transform and produced only 2 under-segmented cells out of 228. This is because the nuclei used as markers are themselves overlapped as shown in Figure 7.12 (a). However, the watershed based on the distance transform produced 9 over-segmented and 6-undersegmented cells.

Table 7.7: Evaluation of Touching Blast Cells Segmentation Results

Experiment Name	Watershed Based on Distance Transform	Marker-Controlled Watershed
Number of manually counted cells	228	
Correct	213(93.4%)	226(99.1%)
Missed	15(6.6%)	2 (0.9%)
Over segmented	9(3.94%)	0(0.0%)
Under segmented	6(2.63%)	2(0.9%)

Moreover, the marker-controlled watershed produced imprecise boundaries between overlapping cells mainly at the internal region, as shown in Figure 7.12 (b). There are many factors affecting the success of marker-controlled watershed segmentation results, such as the proper preparation of the marker size and the correlation between the image patterns, where the watershed transform was applied on and the desired object boundaries.

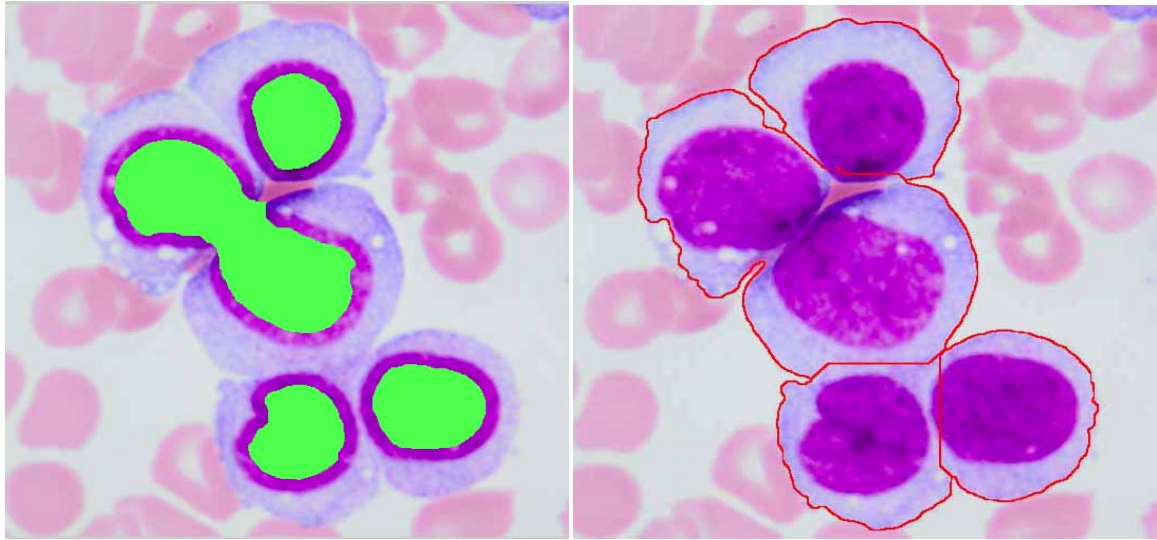


Figure 7.12: Sample result of marker-controlled watershed segmentation. (a) nucleus used as a marker, (b) watershed segmentation boundaries superimposed on the original image

The proposed nucleus/cytoplasm separation algorithm which is part of the **CBCSA** (Please Refer Section 5.3.4) was then compared with two adaptive methods presented in (Putzu & Ruberto, 2013) and (Scotti, 2005). Both of these methods were based on image thresholding using the Otsu threshold (Otsu, 1979). Figure 7.13 shows the nucleus segmentation results obtained by (Putzu & Ruberto, 2013), (Scotti, 2005) and the proposed SRG method using Saturation color channel with a combination of image arithmetic and morphological erosion to produce the seeded region. Figure 7.13 shows that the proposed **CBCSA** performs better than the other methods. Furthermore, the results obtained using the threshold-based segmentation requires some post-processing work in order to delete noisy regions and to fill holes.

In some other cases, the threshold-based segmentation fails to detect the nuclear region such as the blast cell presented in Figure 7.13(row 2). This result may be due to the fact that the cytoplasm region is as dark as the nuclear region.

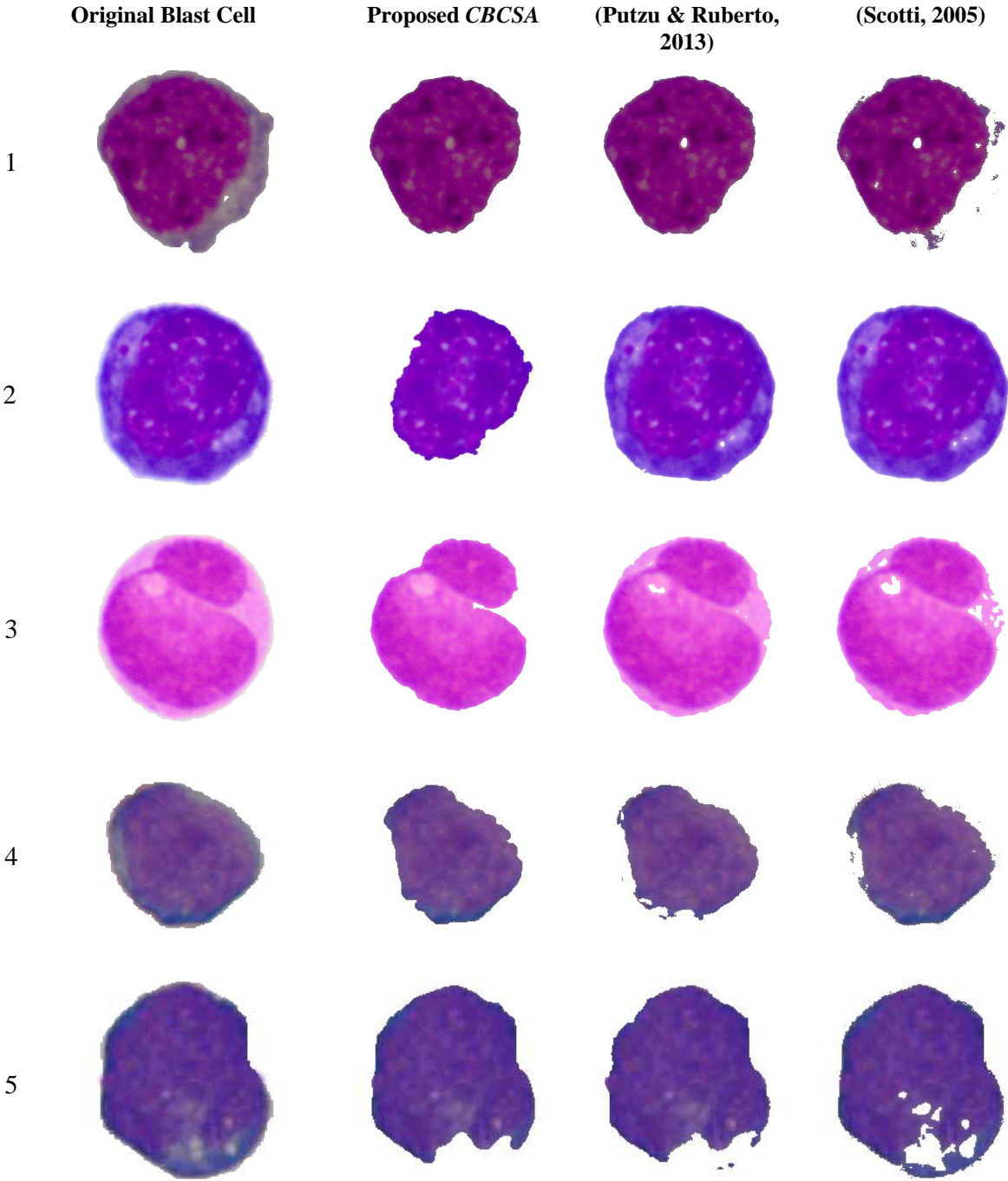


Figure 7.13: Comparison of nucleus segmentation results

Unlike healthy WBCs, the blast cells are characterized by an immature nuclear chromatin pattern (Rubin et al., 2008). As shown in our earlier work (Madhloom et al., 2010) the nucleus of healthy WBCs can be easily segmented using the threshold-based segmentation approach as the nucleus of healthy WBC is dense and homogenous (Gk & Pravati, 2006). However, this immaturity characteristic yields a nucleus with a non-uniform and non-homogeneous texture and color. Consequently, some of the nuclear pixels will be above the selected threshold and others will be below it, which may yield inconsistent results. The performance of the threshold-based segmentation drastically deteriorates for images with blurred object boundaries since they neglect the spatial information (Wang et al., 2013). Other methods such as the Clustering-based approach used in the work by (Mohaptra et al., 2013) and edge detection used in (Sadeghian et al., 2009) confronted similar difficulties as the threshold-based segmentation since these approaches also neglected the spatial information. Clustering-based approach assumes that the distinct regions in the image can be clustered based on the color or the intensity properties. This assumption is true when the image presents a prominent difference in the color or intensity between the regions. However, it performs poorly when a similar color is presented between different regions in the image, such as the blast cell shown in Figure 7.13 (row 2). Besides, clustering-based approaches encounter a great difficulty in deciding the suitable number of clusters of the image (Shafeeq & Hareesha, 2012). On the other hand, based on the experiments carried out in this research, it has been observed that boundaries between the nuclear region, and the cytoplasm region for many blast cell subtypes are blurred and smudged, where these features often deteriorate the performance of the edge detection based approach. Furthermore, the edge detection needs some subsequent edge pixels linking in order to close the contour of the detected regions.

Many of the segmentation algorithms presented in the literature such as the works by (Mohaptra et al., 2013; Sadeghian et al., 2009; Osowski et al., 2009) attempted to segment the blast cells from the blood image, while all the blood components such as erythrocytes, platelets and plasma were still presented in the image. This approach makes the segmentation process tremendously difficult as a lot of interference can come from all these blood components. However, in this research, we cleaned up the blood image from all unnecessary components and retained only the blast cells on an empty white background. This proposed approach made the subsequent step such as splitting the touching cells and nucleus/cytoplasm separation much easier. For instance, the cropped sub-image contains only a single blast cell as a foreground while the background is empty, so there is no need to clean up the sub-image again from other presented components. This feature makes the nucleus/cytoplasm separation process more objective because once the nuclear region is separated, then the cytoplasm region will be found.

The **b** color channel was used to identify the erythrocytes while the **Hue** channel was employed to identify the plasma. In a number of occasions, these two color channels cannot fully guarantee that the erythrocytes and the plasma are completely removed from the PB image. This is where the morphological reconstruction comes into the picture in order to reconstruct only the blast cells based on their nuclei, which were used as Markers to remove all the debris left from the erythrocytes and the plasma. From the results presented earlier, it has been found that the proposed method is robust and reliable in the real-world scenario of PB images.

7.4 Results and Discussion Related to Feature Extraction and Selection

The use of relevant features to characterize an output class is essential for any classification problem. In this study, three types of features were extracted from 1303 blast cells, which belong to *Dataset-A* namely shape, texture and color. The blast cells from *Dataset-B* were not used in the feature extraction and classification stages. This is because of the low magnification power used during image acquisition as well as the blurriness of the PB images from *Database-B*, which make their quality quite inferior when compared to *Dataset-A*. This difference in the quality can produce a huge variation in the extracted features. For that reason, *Dataset-B* was not used to predict the classification performance of acute leukemia.

As discussed earlier in Section 6.2, a total number of 601 features were extracted from each segmented blast cell region and the corresponding nucleus, divided into: 31 shape-based features, 534 texture-based features and 36 color-based features. Figure 7.14 illustrates the results related to the selection of the optimal subset of features using the SFS algorithm. As discussed in Section 4.4 and as can be seen from the results shown in Figure 7.14, the SFS algorithm started with an empty set of features and then began to incorporate the features into larger and larger subsets. As indicated in Figure 7.14, at first, the overall evaluation accuracy of the induction algorithm increased. This meant that the features that were added to the optimal subset improved the performance of the induction algorithm, and then the accuracy rate remained almost unchanged as the number of the selected features reached 14. When more than 18 features were selected, the overall accuracy rate started to decrease. It can be seen that the overall accuracy rate did not increase substantially after the addition of the 19th feature. Based on these, only the top 18 features were selected and utilized to construct the MLP-NN and the SVM classification model.

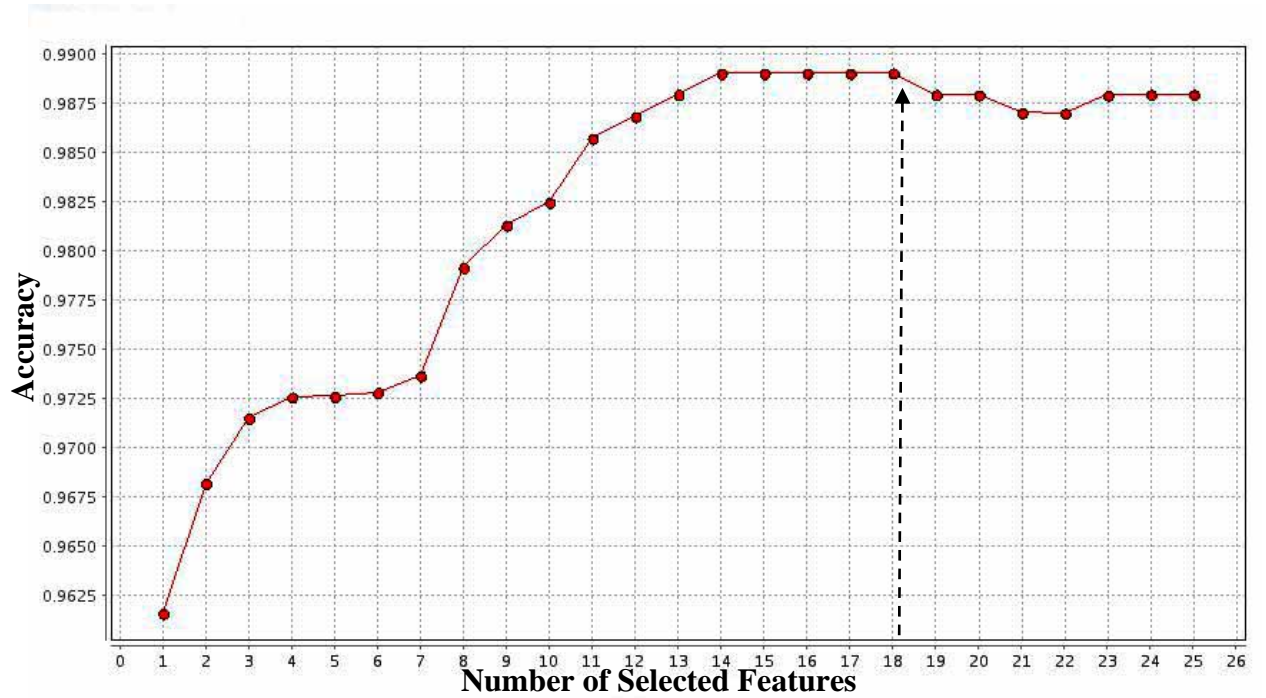


Figure 7.14: SFS Performance

The 18 optimal features which were used to classify blast cells into either *ALL* or *AML* are summarized into mean and standard deviation as shown in Tables 7.8.

Table 7.8: Morphological features extracted from blast cell region and nucleus region of *ALL* and *AML*

Feature Index	Feature Name	<i>ALL</i> $\mu \pm \sigma$	<i>AML</i> $\mu \pm \sigma$
F1	Nucleus Eccentricity	0.496 ± 0.194	0.581 ± 0.199
F2	Nucleus Perimeter	0.292 ± 0.164	0.287 ± 0.095
F3	Nucleus elliptical area	0.500 ± 0.192	0.580 ± 0.199
F4	Ratio between nucleus and the whole cell	0.716 ± 0.246	0.502 ± 0.180
F5	Histogram Standard Deviation	0.361 ± 0.197	0.294 ± 0.139
F6	GLCM_Difference Variance (Quantization=8,distance=1,Angle=0)	0.374 ± 0.177	0.177 ± 0.089
F7	GLCM_Information measure of correlation1 (Quantization=8,distance=1,Angle=0)	0.267 ± 0.164	0.214 ± 0.115
F8	GLCM_Cluster Shade (Quantization=64,distance=1,Angle=0)	0.206 ± 0.239	0.126 ± 0.045
F9	GLCM_Correlation (Quantization=128,distance=1,Angle=45)	0.680 ± 0.178	0.873 ± 0.133
F10	GLCM_Inverse Difference Moment Normalized (Quantization=128,distance=1,Angle=0)	0.734 ± 0.210	0.890 ± 0.060

F11	GLCM_Inverse Difference Moment Normalized (Quantization=128,distance=1,Angle=45)	0.743 ± 0.232	0.871 ± 0.071
F12	GLCM_Correlation (Quantization=256,distance=1,Angle=90)	0.681 ± 0.177	0.871 ± 0.135
F13	GLCM_Cluster Shade (Quantization=256,distance=1,Angle=135)	0.190 ± 0.218	0.129 ± 0.045
F14	Red_Entropy	0.437 ± 0.226	0.624 ± 0.123
F15	Blue_Mean	0.297 ± 0.162	0.482 ± 0.199
F16	Blue_Entropy	0.427 ± 0.246	0.445 ± 0.158
F17	Hue_Mean	0.700 ± 0.195	0.592 ± 0.211
F18	Value_Skew	0.429 ± 0.178	0.550 ± 0.129

The feature selection result indicates that all the three types of the extracted features namely shape, texture and color are important to differentiate between the two acute leukemia types (*ALL* and *AML*). It can be seen from Table 7.8 that the set of optimal features (18 features) comprises of four shape features, nine texture features and five color features. Three out of the four selected shape features are directly related to the nuclear region's shape, while the forth is related to the ratio size between the nuclear region and the whole blast cell region, This emphasizes the importance of the nucleus shape features in the performance of the classification algorithm. This finding indicates that the nuclear region's shape features exhibit higher discrimination power between *ALL* and *AML*. Moreover, it has been found that the color characteristic of the abnormality of the nuclear chromatin causes an increase in the staining capacity of nuclei. Such modification in the DNA content of the nuclei is visible in the form of change in the color intensity in both *ALL* and *AML*. However, this DNA modification and hence, the chromatic color changes affects the *ALL* and the *AML* differently. These modified features demonstrated a tremendous difference between both types of acute leukemia when they were extracted from several color channels, namely, red, blue, hue and value color channels.

The nuclear chromatin distribution pattern, which can be captured in the form of texture, describes the DNA organization structure in the blast cell's nucleus and is an essential diagnostic feature used to differentiate between *ALL* and *AML*. The GLCM was used to compute the statistical and spatial distribution of each blast cell's nuclear chromatin. In this research, a total number of 22 nucleus texture features were extracted from the GLCM (regarding four orientations and one distance) (Please Refer to Section 6.2.2.2) and were assessed to distinguish between *ALL* and *AML*. The GLCM features were calculated using six different gray-levels quantization: 8, 16, 32, 64, 128, and 256. The goal was to investigate the effects and contribution of these 22 GLCM features in the differentiation between *ALL* and *AML* when calculated at different quantization levels. Surprisingly, during the feature selection process, it has been found that the GLCM feature extracted from gray-level quantization 8, 64, 128 and 256 have been included in the optimal set of features (Please Refer Table 7.8). The performance of the GLCM features using different gray-level quantization was used to diagnose various diseases such as breast cancer (Gomez et al., 2012) and glaucoma (Karthikeyan & Rengarajan, 2014). In this research, we found that the GLCM features extracted from gray-level quantization 16 and 32 had no contribution to the differentiation between the two acute leukemia types. However, to our knowledge, extracting 22 GLCM features using different gray-level quantization for the purpose of acute leukemia diagnosis has not been investigated before. Most of the works done in the past, investigated only the performance of a subset of GLCM features using the highest gray-level (256) such as the works carried out by (Reta et al., 2010; Mohapatra & Patra, 2010; Mohapatra et al., 2010b; Mohapatra et al., 2011b; 2013; Madhloom et al., 2012b; and Putzu & Ruberto, 2013). A possible explanation for this is that, these previous researchers assumed that a higher gray-level quantization can provide more information,

which can lead to a better classification performance using the GLCM statistics. Similarly, a lower gray-level quantization would result in the loss of information and a poorer classification. However, this is not always the case for many GLCM statistics, as shown in Table 7.8; some of the GLCM statistics demonstrated a high discrimination power between *ALL* and *AML* such as F6 and F7 at low gray-level quantization. In terms of selecting appropriate GLCM features, the results in this research seemed to support Clausi's (2002) observation, namely, that some GLCM features can produce high discrimination power between classes using high gray-level quantization, while some others can show high discrimination power using low gray-level quantization.

7.5 Experimentation Result of the MLP-NN Architecture Selection

One of the major difficulties with MLPs lies in the selection of the optimal network architecture for a given problem (Andersen & Martinez, 1999). The results of the MLP-NN can vary significantly due to the infinite number of possibilities of the network's architecture. As explained in Section 6.4.3, a single hidden layer structure was chosen for this experiment, while the number of input neurons was decided to be 18 (derived from the feature selection stage), and the number of output classes was set to 2 which corresponded to 2 output classes as either *ALL* or *AML*. Various number of hidden layer neurons were tested (2, 4, 6, 8, 10, and 12). The neurons on each network layer used the *sigmoid* activation functions. Three different learning rates were used (0.001, 0.01, and 0.1). Each network step was trained with a 10-fold cross validation and the accuracy was recorded at every hundred epochs with a maximum number of 1200 epochs.

Figure 7.15 compares the cross-validation accuracy achieved using three different learning rates with different numbers of hidden nodes. The results indicate that the best MLP-NN model was obtained using the learning rate 0.01 and four hidden neurons in the hidden layer. Figure 7.16 compares the best cross-validation accuracy achieved using three learning rates with four hidden neurons and different numbers of training cycles (epoch). The results presented in Figure 7.16 show that the best cross-validation accuracy rate was achieved when the learning rate was 0.01 and at 800 epochs.

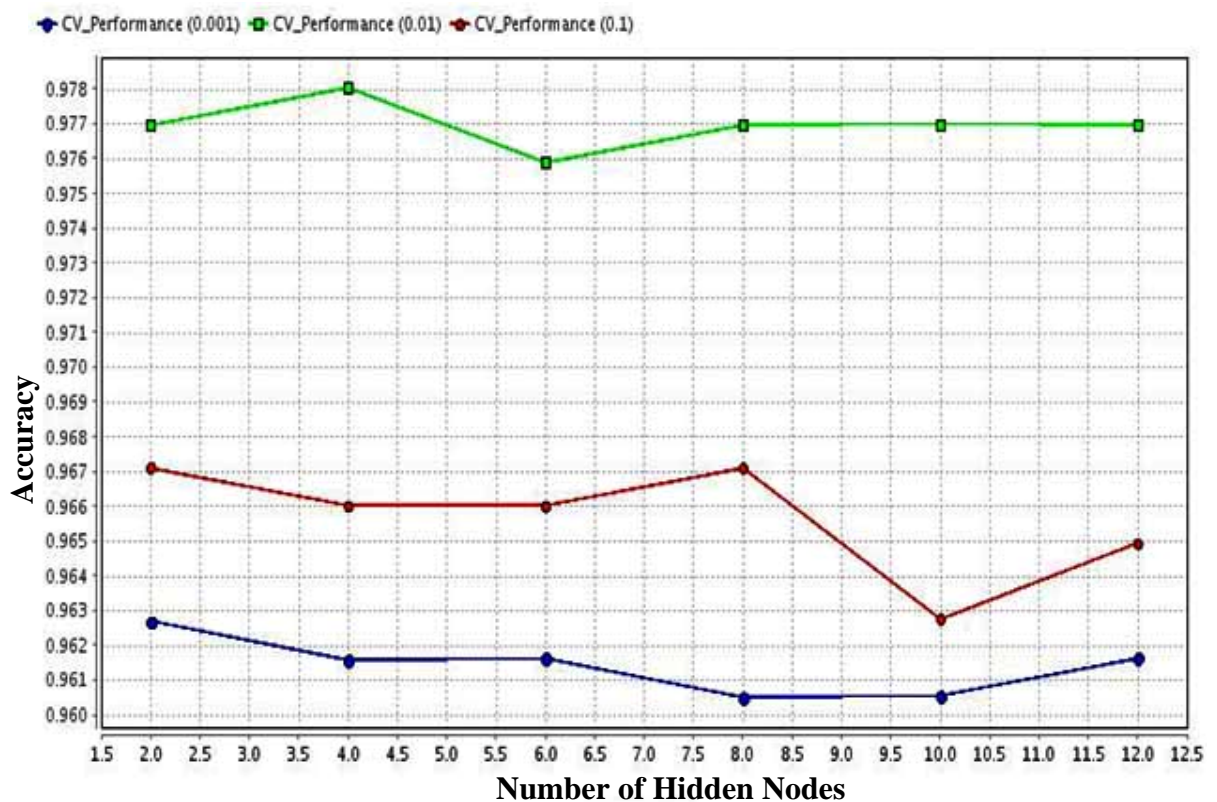


Figure 7.15: Validation Accuracy versus hidden nodes for three different learning rates

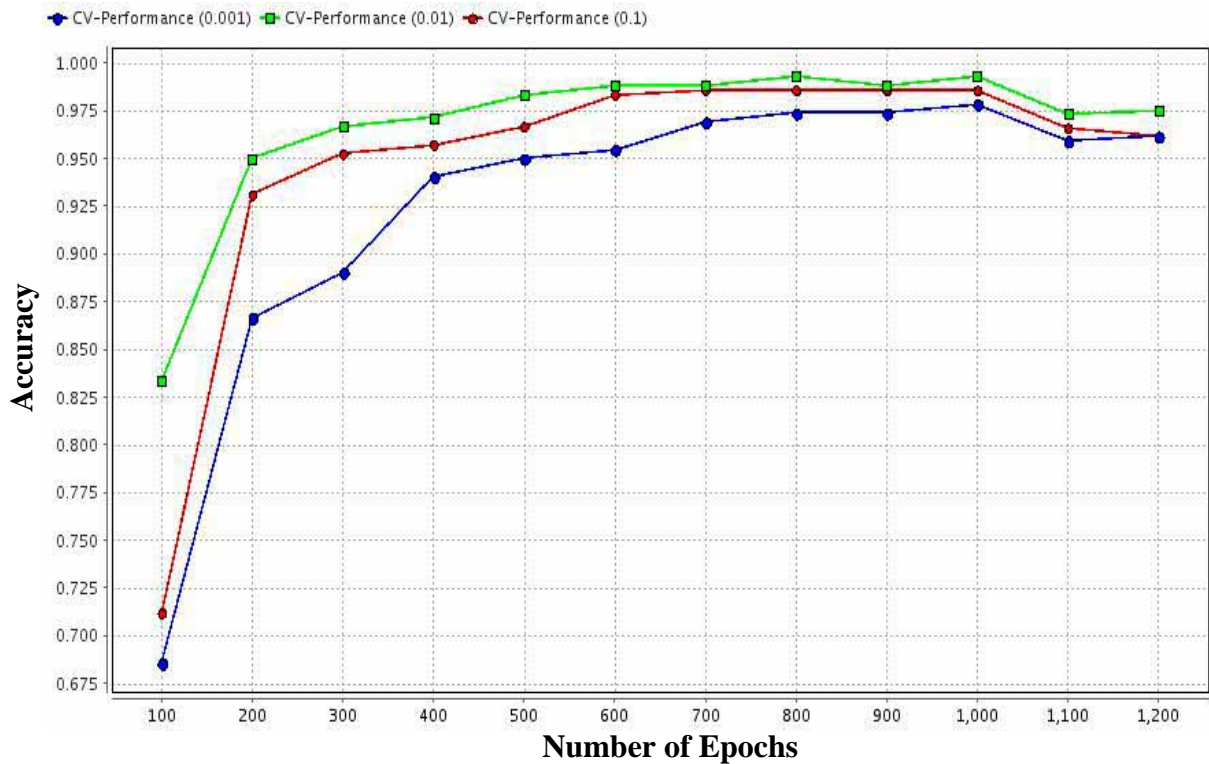


Figure 7.16: Validation Accuracy versus number of training cycles (epochs) for three different learning rates and four hidden nodes

The final MLP-NN model parameters setting is tabulated in Table 7.9, whereas Figure 7.17 depicts the graphical representation of the chosen MLP-NN model. Experimentally, the MLP-NN model with a single hidden layer and the parameters given in Table 7.9 has yielded the highest accuracy results. This conclusion about the design is illustrated in the results shown in Figure 7.15 and 7.16. The chosen MLP-NN model was then trained with the full training dataset (912 samples) and then a blind test (391 samples) was performed on the testing dataset. The training and testing dataset distribution are given in Table 6.7.

Table 7.9: MLP-NN final Parameters setting

Name of Parameter	Input Layer	Hidden Layer	Output Layer
Number of Nodes	18	4	2
Training Function	gradient descent (MLP-Backpropagation Algorithm)		
Learning Rate	0.01		
Number of Epochs	800		
Performance Function	MSE (Mean Square Error)		
Transfer Function	-----	Sigmoid	Sigmoid

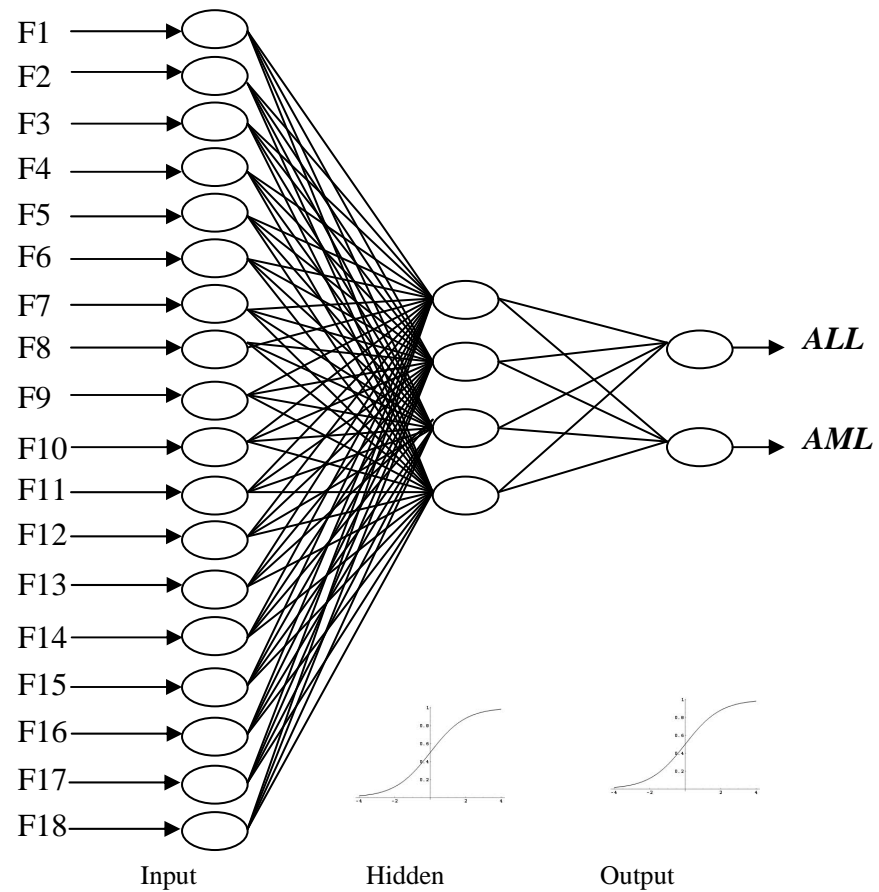


Figure 7.17: Graphical representation of the selected MLP-NN Architecture

7.6 Experimentation Result of the SVM Hyper-Parameters Selection

In this research, the SVM with the RBF kernel was used to classify acute leukemia blast cells. Based on the experiment described in Section 6.4.4, the RBF kernel hyper-parameters were experimentally tuned using a grid search with a 10-fold cross validation approach. A total number of 88 different cross-validation experiments were performed on various combination of C and γ values. It was observed that the cross validation accuracy decreased as the value of C (error penalty) increased to more than 8. Some experiments failed to converge for large values of C ($C > 512$). The optimum value of $C = 8$ was subsequently chosen.

As shown in Figure 7.18, we found that $C = 8$ kept the cross validation error small, while at the same time generalized well on the test set. The performance of the RBF kernel largely depended on the value of γ which was the radius of the RBF kernel. Moreover, the value of $\gamma=0.125$ was found to be best in combination with $C = 8$. As indicated in Figure 7.18, the highest cross-validation accuracy and testing accuracy (Accuracy values highlighted with red color in Figure 7.18) were obtained at (C, γ) equal to $(8, 0.125)$.

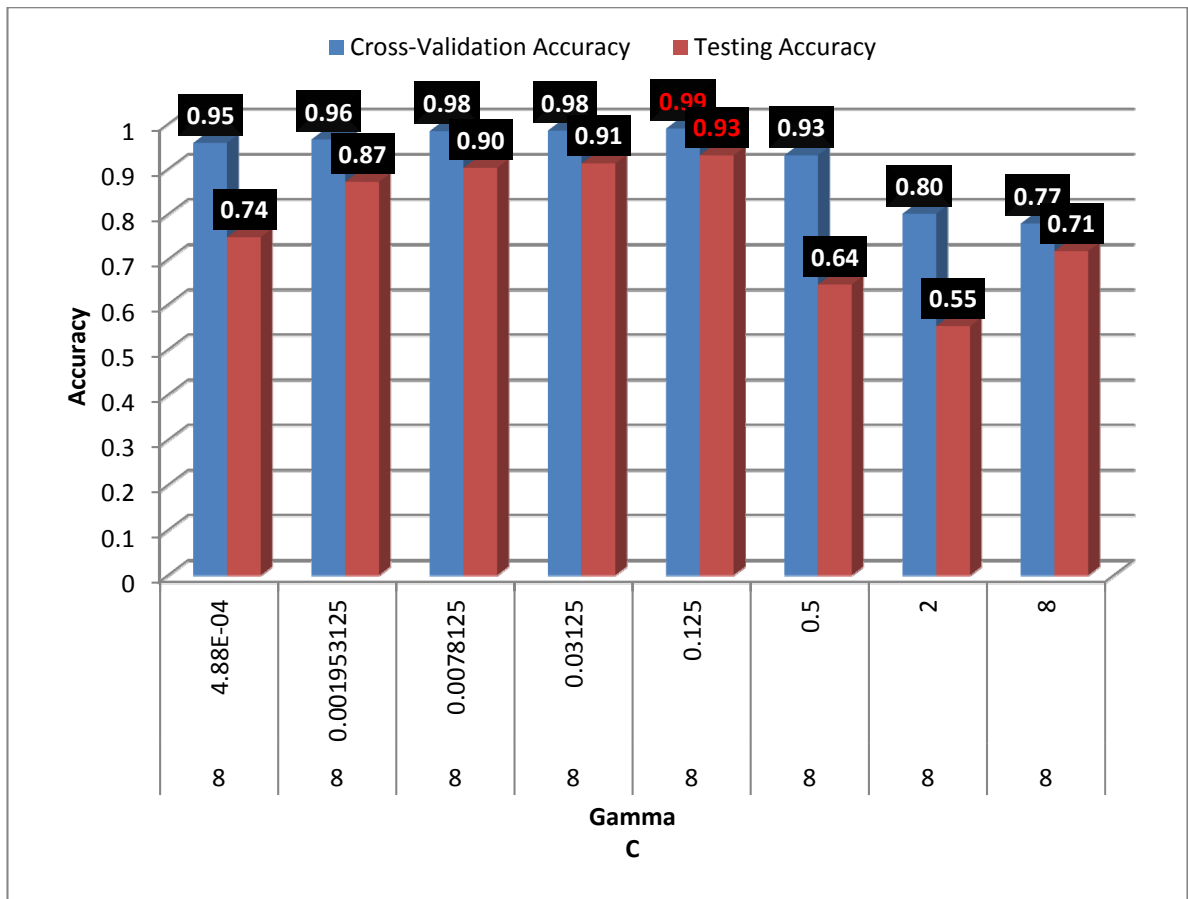


Figure 7.18: Validation and testing accuracy using SVM with various combination of C and γ

7.7 Results and Discussion of Acute Leukemia Classification

In order to meet the objectives of this research outlined in Section 1.3, several experiments were carried out to test the performance of the two classification techniques, namely the MLP-NN and the SVM. As presented in Section 7.5 and 7.6, respectively. The best MLP-NN and SVM models were chosen for subsequent experiments. The highest cross-validation performance of the MLP-NN was obtained at four hidden neurons, 0.01 learning rate and 800 epochs, whereas the best SVM model trained with RBF kernel was obtained at (C, γ) equal to $(8, 0.125)$.

To test the performance of the selected MLP-NN and SVM models for classifying acute leukemia types (**ALL** versus **AML**) in an unbiased way, the full training set (912 samples) and the test set (391 samples) earlier presented in Table 6.7, were used. The test set (blind test set) was considered as an unseen independent dataset where it was not used during the model selection process. Furthermore, each blast cell feature vector consisted of 18 features founded using SFS (Please Refer Table 7.8). It is worth mentioning that all the extracted features were calculated using the blast cell region and the nuclear region segmented using the proposed **CBCSA**. On the other hand, in order to assess the effectiveness of the selected feature, another classification experiment was carried out using the full feature set (without feature selection). The classification performance results of both MLP-NN and SVM tested on 391 samples (**ALL**=98, **AML**=293) with and without feature selection are tabulated in Tables 7.10 and 7.11, respectively. The classification performance results of **ALL** and **AML** are reported using the metrics presented in Section 4.8.2.

Table 7.10: Classification performance using the MLP-NN as the learning machine

Classification Performance Metrics	MLP-NN with FS	MLP-NN Without FS
TPs	291	203
FPs	24	2
FNs	2	90
TNs	74	96
Precision	92.38%	99.02%
Sensitivity	99.32%	69.28%
Specificity	75.51%	97.96%
Accuracy	93.35%	76.47
AUC	0.973	0.979
G-Mean	86.6	82.3

Table 7.11: Classification performance using the SVM as the learning machine

Classification Performance Metrics	SVM with FS	SVM Without FS
TPs	293	268
FPs	25	26
FNs	0	25
TNs	73	72
Precision	92.14%	91.16
Sensitivity	100.00%	91.47
Specificity	74.49%	73.47
Accuracy	93.61%	86.96
AUC	0.981	0.906
G-Mean	86.3	81.8

The testing results of the MLP-NN trained with 18 selected features (Please Refer to Table 7.10) shows that 291 out of the total 293 *AML* samples are correctly classified, whereas 74 out of 98 *ALL* samples are successfully recognized. This experiment gives a classification accuracy of 93.35%. On the other hand, the testing results of the MLP-NN trained with the full set of features shows that 203 out of the total 293 *AML* samples are classified correctly, whereas 96 out of 98 *ALL* samples are classified with an accuracy of 76.47%.

The classification results tabulated in Table 7.11 indicate that the SVM classifier, trained with 18 selected features, was able to classify all the *AML* samples accurately, whereas only 73 out of 98 *ALL* samples were properly recognized achieving an overall accuracy of 93.61%. However, the classification performance of the SVM classifier, trained with the full feature set, achieved a lower accuracy rate of 86.96% in comparison to the results obtained by the SVM trained after feature selection.

Based on these findings, we can conclude that both classification engines obtain better classification performance when they are trained and tested after performing the feature selection step.

A possible explanation for this is that the dataset, with the full set of features, contains many uninformative and irrelevant features which could degrade the classification engine's performance. This finding indicates that the feature selection step can dramatically improve the classification accuracy by eliminating all the irrelevant non-useful features. These findings are consistent with those of (Osowski et al., 2009; Siroic et al., 2007) who conducted a similar experiment to classify 11 types of blast cells taken from the BM with and without feature selection using the SVM classification engine. They concluded that the feature selection step can highly improve the classification performance.

Both the MLP-NN and the SVM tested, with the optimal set of features, demonstrated a highly comparable results as both achieved an accuracy of approximately 93%. Moreover, both classifiers obtained an acceptable range of AUC ($AUC > 0.9$), as many researchers have shown that the acceptable range of the AUC value should be 0.8 or higher (Speight & Hammond, 1998; Ohno-Machado, 2000; Abdul-Kareem, 2002).

The current research does not justify the conclusion that the SVM outperforms ANN in general. This is probably because, in the experiments carried out, the architecture of the feed forward back-propagation neural network had been optimized in terms of hidden neurons, learning rate and number of training cycles before comparing it to the performance of standard SVM with RBF kernel implementation. Nevertheless, our results indicate that solutions obtained by SVM training seem to be more robust with a smaller standard error as compared to MLP-NN training. The robustness of SVM over the MLP-NN is clearly shown when testing the performance of the SVM classifier with the full set features, as the difference between the SVM's performance for full set features and that with feature selection is minimal when compared to the MLP-NN's performance for the full set features and with feature selection.

Although the classification results, obtained from testing the SVM classifier with the full feature set, is not higher than testing it after feature selection. However, the results emphasize the efficiency of the SVM classifier when dealing with a very large number of features (in our case 601 features), due to the exploitation of kernel functions, which is the attraction of SVM classifier. One of the main advantages of SVM is “sparseness of the solution”. This means that an SVM classifier depends only on the support vectors, and the classifier’s function is not influenced by the whole data set, as in the case of many neural network systems (Byvatov et al., 2003).

We also observed from the results in Tables 7.10 and 7.11, respectively that both classifiers tested with optimal feature set (18 features) produced a high FPs rate (*ALL* classified as *AML*) at 24 and 25, respectively. This indicates that both classifiers wrongly recognized the FPs (*ALL* samples) as *AML* samples. The high FPs rates are reflected in the sensitivity and specificity values as it can be noted that the difference between sensitivity and specificity is high.

Clinically, a high difference between sensitivity and specificity could be acceptable when classifying for instance, cells from a leukemic patient versus cells from a healthy person. In this case, high FPs or high difference between sensitivity and specificity will mean a healthy person will be identified as a leukemic patient. This is less dangerous than identifying a leukemic patient as a healthy person. A healthy person who has been identified as a leukemic patient could undergo further investigations and ultimately no pathology will be found. This kind of situation exists with some cancer screening tests such as the cervical smear testing (Lalkhen & Mccluskey, 2008).

On the other hand, in this particular research, both types of classes used (*ALL* and *AML*) were of leukemic (unhealthy) cases. Since each type of acute leukemia requires a different type of treatment, it is important to classify both classes accurately.

There are two possible explanations for the high difference between sensitivity and specificity in our research. Firstly, it could be explained by the high morphological similarity between the two acute leukemia blast cells. For example, the high morphological similarity between the (*ALL*) and the *AML* (M7) (National Cancer Institute, 2014). Secondly, this result may be explained by the fact that the dataset in use is unevenly distributed, where there are more *AML* samples than *ALL* ones, and hence, the bigger class (*AML*) dominates the smaller class (*ALL*) and causes confusion in the performance of both classifiers (MLP-NN and SVM). Nevertheless, our dataset is not significantly imbalanced as the ratio of the minority to the majority class is approximately 1:3. In the work by (Chawla et al., 2004) the significant imbalance ratio between the majority and the minority classes was defined to be as low as 1:100, 1:1000 or 1:10000. However, this small imbalance ratio can still affect the classification performance.

The G-mean value was used as an indicator to deal with the problem of imbalanced data because of its independency from the sample distribution between classes (Barandela et al., 2003; Kubat & Matwin, 1997). The G-mean is commonly utilized when the performance of each class is of concern and expected to be high simultaneously (Sun et al., 2006; Yuan & Liu, 2012).

Both the MLP-NN and the SVM produced almost the same G-mean value of 86%. Normally, a classification model that is constructed from an imbalanced dataset tends to draw the boundaries between classes away from the ideal place and towards the minority side. Thus, the model is likely to classify some *ALL* blasts (minority) into the *AML* blasts

(majority) class, leading to disparity between specificity and sensitivity values. Based on the obtained G-mean value, the imbalanced data issue was further investigated in order to reduce the difference between sensitivity and specificity while increasing the G-mean and the accuracy rate simultaneously. For this purpose, the oversampling technique, namely Synthetic Minority Over-sampling Technique (SMOTE), was adopted.

It is also worth mentioning that during the initial experimentation of acute leukemia classification, it was observed that the sequence of steps carried out starting from feature selection, model selection, rebalancing the training data and then classification resulted in the best classification performance. We noticed that oversampling the training data before feature selection produced inferior classification performance. Furthermore, It was also observed that rebalancing the training data set before model optimization produced the same set of parameters for both the MLP-NN and the SVM. This is because the degree of imbalance in our dataset was small.

7.8 Results and Discussion of Acute Leukemia Classification After Oversampling

The oversampling technique, namely the Synthetic Minority Over-sampling Technique (SMOTE), was adopted to oversample the minority (*ALL*) class by introducing new synthetic samples. In some cases, SMOTE technique produces an intersection between the classes boundary. In order to avoid this, we sought to oversample the minority class at different rates such that the *ALL* samples were oversampled at 50%, 100% and 200%, respectively. Six experiments were conducted and the training set was rebalanced at each oversampling rate. Subsequently, the classifiers (MLP-NN and SVM) were constructed with the rebalanced training set and tested with the blind test set.

The testing results of the MLP-NN and the SVM at the three defined oversampling rates are presented separately in Tables 7.12 and Table 7.13 to give the reader a better idea of the performance about each classifier.

Table 7.12: Classification performance using the MLP-NN at three different oversampling rates

Classification Performance Metrics	MLP-NN (original-imbalance)	MLP-NN (balance at 50%)	MLP-NN (balance at 100%)	MLP-NN (balance at 200%)
TPs	291	291	290	285
FPS	24	16	10	11
FNs	2	2	3	8
TNs	74	82	88	87
Precision	92.38%	94.79%	96.67%	96.28%
Sensitivity	99.32%	99.32%	98.98%	97.27%
Specificity	75.51%	83.67%	89.80%	88.78%
Accuracy	93.35%	95.40%	96.68%	95.14%
AUC	0.973	0.980	0.979	0.970
G-Mean	86.6	91.15	94.27	92.92

Table 7.13: Classification performance using the SVM at three different oversampling rates

Classification Performance Metrics	SVM (Original-imbalance)	SVM (balance at 50%)	SVM (balance at 100%)	SVM (balance at 200%)
TPs	293	289	293	288
FPS	25	12	12	10
FNs	0	4	0	5
TNs	73	86	86	88
Precision	92.14%	96.01%	96.07%	96.64%
Sensitivity	100.00%	98.63%	100.00%	98.29%
Specificity	74.49%	87.76%	87.76%	89.80%
Accuracy	93.61%	95.91%	96.93%	96.16%
AUC	0.981	0.979	0.987	0.978
G-Mean	86.3	93.03	93.68	93.94

Analyzing the results of MLP-NN presented in Table 7.12, it is apparent that oversampling the training data with SMOTE produced a better classification results when compared to the original dataset. However, the classification results at different oversampling rates showed that the best results were achieved when the training dataset was oversampled at 100% rate. It can be seen that the difference between sensitivity and specificity was improved by around 14%, whereas the G-mean and accuracy were increased by nearly 7% and 3%, respectively.

In addition, we observed that oversampling the training set at 50% was not sufficient to significantly reduce the FPs samples, whereas it can be seen that oversampling the training data at 100% improved the classification accuracy. However, at this rate, the effects of intersection between the classes boundary started to appear as the TPs samples were reduced by one. This issue appeared more seriously when rebalancing the training dataset at 200% as the number of FNs samples increased to eight. This indicated that the intersection between the classes boundary became highly overlapped.

SVM classification results demonstrated almost the same behavior as the MLP-NN where the best classification performance was obtained when the training dataset was rebalanced at 100%. It can be seen that the oversampling technique (SMOTE) improved the difference between sensitivity and specificity by around 13%, while the G-mean and accuracy were increased by approximately 7% and 3%, respectively, for both the SVM and the MLP-NN. (Akbari et al.,2004) argued that the SVM has the ability to give more accurate results on moderately imbalanced data. The reason is that only the Support Vectors (SVs) are used for the classification and many majority samples which are far from the decision boundary can be removed without affecting classification.

In a previous comparison of the SVM to several machine learning methods carried out by (Burbidge et al., 2001) to classify drug/non-drug samples, it was observed that an SVM classifier outperformed other classification methods such as the ANN and the Nearest Neighbor. Another study carried out by (Mohaptra et al., 2013) to classify lymphoblasts and lymphocyte showed that the SVM obtained better classification results when compared to other classification methods. However (Burbidge et al., 2001; Gazzah & Amara, 2008) showed that the performance of a specially designed and structurally optimized neural network was comparable to the SVM model. This observation is supported by the finding in our research as both results obtained from the SVM and the well optimized MLP-NN architecture for classifying acute leukemia blast cells were comparable in overall accuracy and produced almost similar results, although not identical sets of correctly and misclassified blast cells.

7.9 Comparison between the Proposed Acute Leukemia Classification Approach and Other Approaches in the Literature.

The problem of acute leukemia classification based on the blast cell morphology has been addressed by several recent research works in the literature as shown earlier in Table 3.6. A comparison of the proposed acute leukemia classification approach with other relevant methods, in terms of accuracy is shown in Table 7.14.

Table 7.14: Performance comparison between the proposed method and other state-of-the-art methods

Reference	Subtypes Included	Classifier	Performance Accuracy
Reta et al., 2010	<i>ALL</i> (L1,L2) <i>AML</i> (M2, M3,M5)	SVM by Sequential Minimal Optimization	92%
Halim et al., 2011	<i>ALL AML</i> Number of subtypes are not reported	MLP- Scaled Conjugate Gradient Fuzzy ARTMAP Network	94.51% 90.27%
Harun et al., 2011	<i>ALL</i> (L1) <i>AML</i> (M1)	Hybrid Multilayered Perceptron Network	97.72%
Supardi e tal., 2012	<i>ALL AML</i> Number of subtypes are not reported	K-Nearest Neighbor	86%
Escalante et al., 2012	<i>ALL</i> (L1,L2) <i>AML</i> (M2, M3,M5)	Ensemble particle swarm model selection (Best-per-iteration)	96.66
Nasir et al., 2013	<i>ALL, AML</i> Number of subtypes are not reported	MLP-Bayesian Regulation MLP-Levenberg Marquardt Simplified Fuzzy ARTMAP	95.70% 95.55% 92.43%
Proposed Method	<i>ALL</i> (L1,L2,L3) <i>AML</i> (M1,M2.M3,M4.M5,M7)	MLP-NN SVM	96.63% 96.93%

It can be noted from Table 7.14 that the proposed approach included nine different acute leukemia subtypes including the L3 and M7 which were segmented with higher error rate at 6% due to its complex apperance compared to the other subtypes (Please Refer to Table 7.5), resulted in a more complex dataset compared to the ones reported in Table 7.14. Nevertheless, the accuracy obtained using the proposed method was as high as 96%, this outperforms most of the other reported methods.

Most of the studies reported in Table 7.14 used BM samples. However, in this research, all the image samples were taken from the PB. The clinical advantage of using the PB smear

rather than BM is that the microscopic morphological examination of the PB sample is often the first step in the diagnostic process of Leukemia.

This diagnostic procedure is still considered the most economical procedure for initial screening of acute leukemia (Angulo et al., 2006). Thus, the initial diagnosis usually starts by examining the PB smear, and prior to exposing the patient to more invasive procedures, such as BM biopsy. Hence, a PB smear screening is of particular importance because it facilitates rapid diagnosis and specifies treatment (Bain, 2005). Besides, the PB smear screening can be done at any general clinic or medical laboratory.

The clinical impact of our work is that it will hopefully aid primary care physicians and general practitioners who are usually the first contact of the leukemia patients. As leukemia is cancer of both the adults and children, and in Malaysia, a country of 13 states of 30 million people, there are only four tertiary-referral centers for childhood cancer and less than 30 trained pediatric hemato-oncologists. Hence, having a tool to facilitate the initial screening of children suspected of having acute leukemia would be very beneficial to clinicians and laboratories located outside of major hospitals.

Accurate cancer classification is important in order to save the human's life. Despite using common diagnostic tools, most researchers nowadays are interested in using ML classification techniques to classify cancer. This research was conducted in order to compare the performance of two ML classification techniques, which were the SVM and the MLP-NN in classifying acute leukemia blast cells. Both techniques can be used effectively in acute leukemia classification.

7.10 Summary

In this chapter, we presented the testing and evaluation of the CAD-AL. The CAD-AL covered the main diagnostic components of segmentation, feature extraction, feature selection and classification. Two segmentation methods were proposed namely **BCL** and **CBCSA**. The **CBCSA** is an advance and robust segmentation algorithm which was tested and evaluated using two different datasets of PB images acquired from two different sources. The feature extraction stage involved a generation of categorized features, extracted from (i) shape-based techniques, (ii) texture-based technique and (iii) color-based technique. The feature selection showed that all the three categories of features are useful for the classification of acute leukemia blast cells.

A comprehensive set of experiments was conducted in this study to obtain the best parameters setting for both the MLP-NN and the SVM.

Furthermore, two classifiers; namely, the Support Vector Machine, and the Multi-layer Perceptron were employed in this research. The results showed that both classifiers were highly effective in the classification of acute leukemia blast cells using the optimal set of features.

The SMOTE approach was also used to provide an improvement in the classification of the imbalanced class datasets. At 100% oversampling, the minority and majority classes were optimally recognized by both classification techniques and the intersections between the classes regions were avoided.

CHAPTER EIGHT

CONCLUSION AND FUTURE WORK

8.1 Conclusion

Acute leukemia is a group of heterogeneous deadly diseases that affects all ages, and its accurate diagnosis is remarkably important for the reduction of its morbidity and mortality. Despite recent developments in technology and investigation modalities such as flow cytometry and cytogenetics, the clinical diagnosis of acute leukemia remains challenging and suffers from inherent subjectivity, particularly for primary healthcare practitioners, who are the initial contact point for hematology patients. As a result, CAD-AL has become a major research focus in recent years (Cornet et al., 2008) with the aim of assisting general practitioners and clinical laboratory practitioners in Leukemia diagnosis, through the provision of quantitative reproducible analysis of the PB smear. CAD-AL can play a useful role in producing quantitative results, recording patient follow-ups, and monitoring therapeutic progress. However, this type of CAD system has not been practically adopted yet and the work for such a diagnostic system is still in progress. For that to happen, a diagnostic system with higher accuracy and reliability compared to what has been achieved so far, is needed.

This research presents new acute leukemia methodology, with the aim of improving some of the existing methods and developing new techniques to facilitate an accurate and reliable acute leukemia classification. The proposed methodology involves several stages including image acquisition, image segmentation, feature extraction, feature selection, and classification. In this chapter we conclude the thesis by explaining the contributions of this research with respect to each stage of the proposed methodology. This is followed by a number of recommendations for future work on this topic.

In this research we proposed a new methodology and subsequently developed a system that segments and classifies acute leukemia blast cells in PB smear images. A graphical illustration of the segmentation and classification process is given in Figure 1.4 and 4.1 (a-c). The modeling and development of the proposed system is presented in Chapter 5 and 6. The proposed system involved four main stages namely: the image acquisition stage, the segmentation stage, the feature extraction/selection stage and classification stage. The most difficult and challenging part was the PB image segmentation which was further divided into two parts: the single blast cell localization and extraction, and the nucleus/cytoplasm separation. A segmentation algorithm called **CBCSA** was developed in order to extract a single blast cell from the high-cell population PB image and to separate the nucleus region from the cytoplasm region. Various image processing techniques were combined to localize a single blast cell such as color space analysis, image thresholding, mathematical morphology and watershed, while SRG was used to separate the nucleus from the cytoplasm. We used two different datasets of PB images acquired from two different sources to test the robustness of the proposed algorithm. The segmentation results presented in Section 7.2.2 (Table 7.2, 7.3, 7.4 and 7.5) showed several key advantages of the proposed **CBCSA** such as the robustness and accuracy of the algorithm in comparison to the manually segmented blast cells performed by a clinical expert. The results are very promising, and should open new paths for leukemia diagnosis. In addition, the present algorithm can also be used to detect normal WBC, including lymphocytes, indicating that it can be used for differential blood count systems.

601 blast cell related features from three different categories namely, shape, texture and color were extracted and evaluated in various experiments.

18 features were subsequently selected (Please Refer to Table 7.8) based on their performance using SFS. These 18 features were then used to classify blasts cells into *ALL* or *AML* using SVMs and MLP-NN.

The parameters of both classifiers were tuned using a 10-fold cross validation, the model produced the highest accuracy on validation set was constructed and trained with the full training set and tested on an unseen test set. The experimental results presented in Section 7.7 (in Table 7.10 and 7.11) indicated that SVM and MLP-NN achieved high accuracy of 93.61% and 93.35% respectively, in classifying acute leukemia blasts. Furthermore, it has been concluded that the results obtained, using specially designed and structurally optimized neural network were comparable to the SVM model, where both provided almost comparable classification performance.

8.2 Main Contribution

The key contributions of this thesis are summarized below:

a. Color-channels analysis for blast cells detection in PB microscopic images

Color information plays a significant role in PB image segmentation. Variegation in the color of various blood components and the diversity found in the staining quality of different PB slides make blast cell localization and extraction a challenging process. Although numerous blood cell segmentation methods have been developed, only a few studies focused on determining the most effective color space and color channels which can highly differentiate between various blood components. In this research, we proposed a comprehensive color channel analysis procedure, which showed that (1) the enhancement of the original RGB image using color contrast stretching makes the blood components (RBC, nucleus, cytoplasm, plasma) highly different from each other in terms of color (as illustrated in Figure

5.14 (b)). (2) The **b** channel of the Lab color space effectively highlights almost all the erythrocytes in the PB image which can then be easily removed using optimal threshold (as illustrated in Figure 5.15-5.17). (3) The HSV color space was found to be very efficient in highlighting the plasma background through the **Hue** channel (as shown in Figure 5.18 (b-c)), while the Saturation channel was found to be the optimal color channel that was used to highlight the initial seed points of the blast cell nucleus. The determined color channels were the main building block for the blast cells detection and extraction since it provided a remarkable added value to the final segmentation result.

b- Blast cells localization and extraction in PB images.

The goal of the localization and extraction process is to separate each blast cell from the background components such as erythrocytes, platelets and plasma. It is considered a fundamental step towards the development of the CAD-AL, and its accuracy is essential for precise implementation of the subsequent steps. In various methods available in the literature (Please Refer to Table 3.4), the idea was to initially identify the nuclei which are the most prominent regions in the PB image, (due to the blood staining which highlight the nucleus with dark purple color) and then to detect the entire cell membrane such as in the work of (Mohapatra et al., 2013, Sadeghian et al., 2009). However, the proposed **CBCSA** first cleans up the PB image from all the unwanted blood components such as erythrocytes and plasma by applying Otsu's global histogram thresholding on the optimal color channels derived from the above-mentioned color spaces analysis process. This is followed by the application of morphological reconstruction to reconstruct a completed blast cell (Please Refer to Section 5.2.3) and watershed transform to separate touching

blast cells. This method resulted in blast cell localization and extraction with a maximum error of approximately 4% when compared to the ground truth (Please Refer to Table 7.5).

c- Nucleus region segmentation using SRG.

The accuracy of blast cell nucleus segmentation is remarkably important for acute leukemia blast cells classification. This is because the nucleus region carries invaluable characteristics of blast cell such as nuclear shape and chromatin pattern (Rubin et al., 2008). It has been shown in our earlier work (Madhloom et al., 2010) that the nucleus of healthy WBC can be easily segmented using a combination of arithmetic operation and automatic thresholding. However, due to the inhomogeneous nature of the blast cell nucleus chromatic pattern, it has been concluded that segmentation techniques that do not consider spatial information such as thresholding could not produce satisfactory results. For that reason, in this research, the SRG was used to separate the nucleus region from the cytoplasm region. This method resulted in nucleus region extraction with a maximum average error of approximately 5.59% according to the $RUMA_A$ metric (Please Refer to Table 7.5).

d- GLCM feature extraction for blast cell nucleus with various quantization levels

In a CAD-AL, feature extraction is applied to extract the features that accurately characterize a blast cell. These features, which are extracted from the ROI determined through the segmentation process, are similar to those visually detected by the hematologist in their clinical practice. By studying the morphological examination of the PB smear under the microscope as well as consulting the hematologist from UMMC, a comprehensive feature extraction strategy was

developed. The proposed method captures features regarding (1) the characteristics of the external attribute of the blast cell and its corresponding nucleus such as shape and border characteristics (F1-F4 Table 7.8) and (2) the color and texture characteristics of the nucleus chromatin pattern (F5,F14-18 Table 7.8).

In this research, 22 GLCM texture features were extracted from the blast cell nucleus based on six different gray-level quantization (Please Refer to Table 6.2, 6.3,6,4 and 6.5). To our knowledge, this approach has not been investigated before in any CAD-AL.

Experimental results demonstrated that some GLCM features (F6-F13 Table 7.8) can produce high discrimination power between *ALL* and *AML* using high gray-level quantization, while some others can show high discrimination power using low gray-level quantization. It can be concluded that GLCM features extracted using different quantization levels prominently contribute to the final optimal feature set (Please Refer to Table 7.8).

e- Machine learning approach to classify acute leukemia blast into *ALL* and *AML* from PB images

Feature selection is the process of eliminating irrelevant, redundant, or noisy features from the initial feature set created during the feature extraction step. In this research the Sequential forward selection (SFS) method which has the ability to consider features dependency was used. SFS has been successfully applied in various applications, yet to the best of our knowledge it has not been applied in previous acute leukemia studies. In the literature, there has not yet been a clear agreement on which feature sets are the most suitable for this task. However, this research has established that the three extracted categories of features, namely shape, texture and color are important for the final classification of acute leukemia

blast cells. Furthermore, it has been observed that the number of features used is also crucial for the classification accuracy. As shown in Table 7.10 and 7.11, the classification accuracy improved by 17% and 7% for both MLP-NN and SVM respectively, after performing feature selection.

This highlights the importance of the feature selection. Hence, feature selection not only improves the classification complexity by minimizing the utilized number of features, but also enhances the classification accuracy.

The secondary contribution of this research provides the basis for conducting a comparative analysis between different ML classification technologies, such as MLP-NN and SVM. It has been concluded that a well optimized MLP-NN architecture can produce comparable results to SVM classifiers. This indicates that both classifiers have a good generalization capability for the classification of *ALL* and *AML*.

In addition to the above novel contributions, the current research found that high microscopic magnification power such as 1000x (used to capture PB images in *Dataset-A*) is not only beneficial for blood smear visual analysis (Díaz, and Manzanera., 2011), but also for blood smear image analysis. Lower magnification power such as 300x-500x (used to capture images in *Dataset-B*) makes blood cells look smaller, which in turn, reduces the appearance of cellular details. As an example lower magnification power reduces the appearance of the boundaries between the nucleus and the cytoplasm and hence, affects the nucleus region segmentation accuracy. Table 7.5 demonstrates that the nucleus region segmentation accuracy obtained from *Dataset-A* is higher than that obtained from *Dataset-B*. This finding confirms the association between high magnification power, used to capture the images, and the image segmentation accuracy.

8.3 Achievement of Research Objectives

As discussed in Section 1.3, the goal of this research is to increase the diagnostic accuracy of image processing and machine learning techniques for optimum classification between *ALL* and *AML* as well as the reproducibility of PB smear morphological examination. In order to achieve this goal, a new acute leukemia diagnostic methodology was carefully analyzed and designed, as shown in Figure 4.1 (a-c). The research objectives outlined in Section 1.3 are discussed as follows:

- 1- A new acute leukemia diagnostic methodology was carefully analyzed and designed as shown in Figure 4.1 (a-c), taking into consideration the current problems which are mainly related to the segmentation stage such as color variation in PB image, cell localization, and separation of touching and overlapping cells. This point complies with the first objective of the research.
- 2- The optimal selected feature set constructed using the SFS (Please Refer to Table 7.8) has shown to reduce the classification error for both classifiers (MLP-NN and SVM) when compared to the use of the full feature set for classification. This complies with the second objective of the research.
- 3- A reliable and robust blast cells segmentation algorithm, namely, *CBCSA* (Please Refer to Section 5.3) was developed to localize and extract blast cells and their nuclei with high accuracy (Please Refer to Table 7.5). The performance of the proposed *CBCSA* was tested on two distinctive datasets collected from different sources namely, *Dataset-A* and *Dataset-B*. Moreover, the proposed feature set has shown to be very useful for the differentiation of acute leukemia blast cells. The classification performance of the proposed feature set was evaluated using two

classifiers called MLP-NN and SVM. This point complies with the first, second, and the third objectives of the research.

8.4 Impact and Significance to the Medical Field

The proposed methodology (Please Refer to Figure 4.1 (a-c)) can be implemented as an intelligent image based diagnostic system (CAD-AL) to assist laboratory practitioners, primary care physicians, as well as hematologists in their daily practice by acting as a second reader. This issue highlights the demand for receiving a non-invasive in vivo second opinion, which increases the diagnostic accuracy and decreases the human error rate. The implementation of such a system in real-world practice can lead to the following advantages:

1. It saves more lives and reduces financial and emotional costs imposed on patients, such as eliminating the need for performing unnecessary bone marrow biopsy procedures.
2. It is highly desirable and advantageous for hematologists to have a diagnostic system which could provide quantitative and objective evaluation of acute leukemia blast cells.
3. CAD-AL would allow for reproducible diagnosis by diminishing the inter-observer variability that could be found in the diagnosis of acute leukemia.
4. It reduces the amount of repetitive and tedious work done by physicians allowing the inspection of more PB slides and reporting the findings in a shorter time frame. This would be very useful for people who need regular follow-ups, thus saving time and cost of multiple travels, particularly for those who live in rural areas.
5. CAD-AL can also be used for educational purposes where the trainee can use the CAD-AL to validate his/her findings.

Despite the existence of highly accurate leukemia diagnostic methods such as flow cytometry and cytogenetics, these technologies are very expensive and only a few hospitals (in developing countries) can afford to buy them. Besides, these equipment need highly trained laboratory personnel and, often, an exchange of information among the clinicians (Anlike et al., 2013). Therefore, a computer based morphological blood analysis method for acute leukemia identification offers a better tradeoff between low cost and accuracy, and we believe it may have a broad impact. In a real world scenario, for example, in Malaysia, a country comprising of 13 different states with an approximate population of 30 million, there are only four tertiary-referral centers for childhood cancer and less than 30 trained pediatric hemato-oncologists. Hence, having a CAD-AL system that facilitates the counting and classification of blast cells of children having acute leukemia, would be beneficial to clinicians and laboratories in terms of speeding up the process and verifying the results, especially if it is done by a junior personnel.

8.5 Future Expansion and Recommendations

Throughout the thesis, we have presented new ideas and approaches to address the key components of an acute leukemia diagnostic process using image processing and ML techniques. We now canvass several possible directions that can be pursued to further advance the state of this research. The following are some suggestions that may serve as the foundation for future research.

- 1- The proposed methodology achieved an acceptable segmentation and classification accuracy of over 90%, as indicated in Chapter 7. However, it is still required to improve the system by extensively testing it on a larger number of PB images with different conditions, i.e., on thousands of PB images acquired with different staining

methods and different microscopic and camera settings. This would require a well-organized participation of several medical centers. The CAD-AL should be tailored to satisfy the needs of: (i) expert users, including hematologists, who may seek a quantitative analysis of a given PB slide as a second opinion to their clinical diagnosis, and (ii) non-expert users, including general practitioners or laboratory practitioners, who may need to increase the accuracy of their diagnosis and gain more confidence in leukemia diagnosis.

- 2- This research uses GLCM texture features for quantifying the nucleus chromatic pattern. The reason behind using GLCM in this research is attributed to the recent success of GLCM in blood cells analysis. However, this does not mean that amongst the texture based techniques, only GLCM can provide good computation results. Thus, other texture based techniques can be evaluated in this research to perform a comparative study. The following texture analysis techniques have gained recent success in pattern classification problems and can be considered:

- (a) Local binary patterns
- (b) Tamura Texture features
- (c) Gray Level Run Length Matrix (GLRLM)

A comparative study evaluating other texture feature estimation approaches will benefit this research, with a possibility of further improving the performance and accuracy of the developed system.

- 3- Leukemia diagnosis using morphological examination of PB smear under the microscope is still considered the most economical procedure for initial screening of acute leukemia (Angulo., et al, 2006).

One of the main goals of this research is to provide a reliable and accurate computer-based diagnostic system for acute leukemia using PB smear images. However, it is worthwhile to further extend the proposed methodology to be used for morphological examination of both PB images and BM images.

8.6 Summary

In this research, we have focused on a new methodology for acute leukemia diagnosis and classification using image processing and ML techniques. The proposed methodology involved four main stages starting with image acquisition, image segmentation, feature extraction/selection and finally classification. We subsequently developed a system, which we named CAD-AL (computer-aided diagnosis for acute leukemia) to perform this diagnostic procedure. In this chapter, we concluded this thesis by highlighting our major contributions and discussing the achievement of the research objectives with respect to the experimental results obtained along with the key findings and significance of the research. In addition, this chapter discusses the impact and significance of our research to the hematology community in particular, and to the society in general. Furthermore, a number of future activities to extend the work are highlighted in this chapter.

REFERENCES

- Abdul-Hamid, G. (2011). Classification of Acute Leukemia. *Acute Leukemia-The Scientist's Perspective and Challenge*. Rijeka: InTech. doi: 10.5772/19848
- Abdul-Kareem, S. (2002) Application of artificial neural network for the prognosis of nasopharyngeal carcinoma. Doctoral thesis, University Of Malaya, Kuala Lumpur Malaysia).
- Abdul-Kareem, S., Baba, S., Zubairi, Y. Z., & Wahid, M. I. A. (2000). ANN as a tool for medical prognosis. *Health Informatics Journal*, 6(3), 162-165.
- Abou-Alfa, G. K., & DeMatteo, R. (2011). 100 questions & answers about liver cancer: Jones & Bartlett Publishers.
- Adams, R., & Bischof, L. (1994). Seeded region growing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(6), 641-647. doi: 10.1109/34.29591
- Adds, J., & Larkcom, E. (1999). *Tools, techniques and assessment in biology: a course guide for students and teachers*: Nelson Thornes.
- Aehlert, B., & Vroman, R. (2011). *Paramedic Practice Today: Above and Beyond, Volume 1*: Jones & Bartlett Publishers.
- Aimi Salihah, A. N., Mashor, M. Y., Harun, N. H., & Rosline, H. (2010). *Colour image enhancement techniques for acute Leukaemia blood cell morphological features*. Paper presented at the Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on. Istanbul, Turkey. doi: [10.1109/ICSMC.2010.5641867](https://doi.org/10.1109/ICSMC.2010.5641867)
- Aimi Salihah, A. N., Mustafa, N., & Fazli, N. (2009). *Application of Thresholding Technique in Determining Ratio of Blood Cells for Leukemia Detection*. Paper presented at the Proceedings of the International Conference on Man-Machine Systems, Batu Ferringhi, Penang, Malaysia.
- Aimi Salihah, A., Mashor, M., Harun, N. H., & Rosline, H. (2010). *Colour image enhancement techniques for acute leukaemia blood cell morphological features*. Paper presented at the Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on.
- Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets *Machine Learning: ECML 2004* (pp. 39-50): Springer.
- Akilandeswari, U., Nithya, R., & Santhi, B. (2012). Review on feature extraction methods in pattern classification. *European Journal of Scientific Research*, 71(2), 265-272.
- Almarzooqi, S., Crumbacher, J., Firgau, E., & Kahwash, S. (2011). Comparison of Peripheral Blood versus Bone Marrow Blast Immunophenotype in Pediatric Acute Leukemias. *Ibnosina Journal of Medicine and Biomedical Sciences*, 3(6), 195-204.

- American Childhood Cancer Organization. (2012). Childhood Cancer Statistics. Retrieved 19/3/2012, from <http://www.acco.org/Information/AboutChildhoodCancer/ChildhoodCancerStatistics.aspx>
- American Cancer Society. (2013). Childhood Leukemia Retrieved 23/10/2013, from <http://www.cancer.org/cancer/leukemiainchildren/detailedguide/childhood-leukemia-diagnosis>
- American Cancer Society (2012). How is childhood leukemia classified. Retrieved 3/4/2012, from <http://www.cancer.org>.
- Andersen, T., & Martinez, T. (1999). Cross validation and MLP architecture selection. In Neural Networks, 1999. IJCNN'99. International Joint Conference on (Vol. 3, pp. 1614-1619). IEEE.
- Angelescu, S., Berbec, N. M., Colita, A., Barbu, D., & Lupu, A. R. (2012). Value of Multifaced Approach Diagnosis and Classification of Acute Leukemias. *Mædica*, 7(3), 254.
- Angulo, J., Klossa, J., & Flandrin, G. (2006). Ontology-based lymphocyte population description using mathematical morphology on colour blood images. *Cellular and Molecular Biology*, 52(6), 3-16
- Angulo, J., Klossa, J., & Flandrin, G. (2006). Ontology-based lymphocyte population description using mathematical morphology on colour blood images. *Cellular and Molecular Biology*, 52(6), 3-16.
- Angulo, J., & Flandrin, G. (2003). Automated Detection of Working Area of Peripheral Blood Smears Using Mathematical Morphology. *Analytical Cellular Pathology*, 25(1), 37-49.
- Ariffin, H. (2012). Private Meeting with the Hemato-oncologist Prof. Dr. Hany Ariffin from the University Malaya Medical Center.
- Babalola, K., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., . . . Rueckert, D. (2008). Comparison and Evaluation of Segmentation Techniques for Subcortical Structures in Brain MRI. In D. Metaxas, L. Axel, G. Fichtinger & G. Székely (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008* (Vol. 5241, pp. 409-416): Springer Berlin Heidelberg.
- Bailey, D. G. (1992). *Segmentation of touching objects*. Paper presented at the 7th NZ Image Processing Workshop (pp. 1-6).
- Bain B.J. (1991). *Leukemia Diagnosis*: Blackwell Publishing.
- Bain, B. (2003). *Leukaemia diagnosis* (Third ed.): John Wiley & Sons.

- Bain, B. J. (2005). Diagnosis from the blood smear. *New England Journal of Medicine*, 353(5), 498-507.
- Bain, B. (2006). *Blood Cells: A Practical Guide* (Fourth ed.): Published by Wiley-Blackwell.
- Bain, B. J. (2008). *A beginner's guide to blood cells*: John Wiley & Sons.
- Bain, B. J. (2010). *Leukaemia Diagnosis*: John Wiley & Sons.
- Ballabio, D., & Todeschini, R. (2009). Chapter 4 - Multivariate Classification for Qualitative Analysis. In D.-W. Sun (Ed.), *Infrared Spectroscopy for Food Quality Analysis and Control* (pp. 83-104). San Diego: Academic Press.
- Barandela, R., Sánchez, J. S., García, V., & Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3), 849-851.
- Barua, S., Islam, M. M., & Murase, K. (2011). A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning. In B.-L. Lu, L. Zhang & J. Kwok (Eds.), *Neural Information Processing* (Vol. 7063, pp. 735-744): Springer Berlin Heidelberg.
- Batuwita, R., & Palade, V. (2013). Class Imbalance Learning Methods for Support Vector Machines *Imbalanced Learning* (pp. 83-99): John Wiley & Sons, Inc.
- Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, 3(10), 27-38.
- Bellman, R., Bellman, R. E., Bellman, R. E., & Bellman, R. E. (1961). *Adaptive control processes: a guided tour* (Vol. 4): Princeton university press Princeton.
- Bennett, J. M., Catovsky, D., Daniel, M. T., Flandrin, G., Galton, D., Gralnick, H. t., & Sultan, C. (1976). Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. *British journal of haematology*, 33(4), 451-458.
- Bennett, J. M., Catovsky, D., Daniel, M. T., Flandrin, G., Galton, D., Gralnick, H. t., & Sultan, C (1980). A variant form of acute hypergranular promyelocytic leukemia (M3). *British journal of haematology*, 44(1) 169-170.
- Bennett, J., Catovsky, D., Daniel, M.-T., Flandrin, G., Galton, D., Gralnick, H., & Sultan, C. (1981). The morphological classification of acute lymphoblastic leukaemia: concordance among observers and clinical correlations. *British journal of haematology*, 47(4), 553-561.

- Bennett, J. M., Catovsky, D., Daniel, M. T., Flandrin, G., Galton, D. A., Gralnick, H. R., & Sultan, C. (1985). Proposed revised criteria for the classification of acute myeloid leukemia A report of the French-American-British Cooperative Group. *Annals of internal medicine*, 103(4), 620-625.
- Bennett, J. M., Catovsky, D., Daniel, M.-T., Flandrin, G., Galton, D. A., Gralnick, H. R., & Sultan, C. (1985a). Criteria for the diagnosis of acute leukemia of megakaryocyte lineage (M7) A report of the French-American-British Cooperative Group. *Annals of internal medicine*, 103(3), 460-462.
- Bennett, J., Catovsky, D., Daniel, M. T., Flandrin, G., Galton, D., Gralnick, H., & Sultan, C. (1991). Proposal for the recognition of minimally differentiated acute myeloid leukaemia (AML-MO). *British journal of haematology*, 78(3), 325-329.
- Bergen, T., Steckhan, D., Wittenberg, T., & Zerfaß, T. (2008). *Segmentation of leukocytes and erythrocytes in blood smear images*. Paper presented at the 30th Annual International IEEE EMBS Conference, British Columbia, Canada.
- Bernasconi, A., Antel, S. B., Collins, D. L., Bernasconi, N., Olivier, A., Dubeau, F., . . . Arnold, D. L. (2001). Texture analysis and morphological processing of magnetic resonance imaging assist detection of focal cortical dysplasia in extra-temporal partial epilepsy. *Annals of neurology*, 49(6), 770-775.
- Beucher, S., & Lantuéjoul, C. (1979). *Proc. Int. Workshop Image Processing, Real-time Edge and Motion Detection/Estimation*. Paper presented at the Proc. Int. Workshop Image Processing, Real-time Edge and Motion Detection/Estimation.
- Billard, M., Lainey, E., Armoogum, P., Alberti, C., Fenneteau, O., & Da Costa, L. (2010). Evaluation of the CellaVision™ DM automated microscope in pediatrics. *International Journal of Laboratory Hematology*, 32(5), 530-538.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*: Oxford university press.
- Bonilha, L., Kobayashi, E., Castellano, G., Coelho, G., Tinois, E., Cendes, F., & Li, L. M. (2003). Texture analysis of hippocampal sclerosis. *Epilepsia*, 44(12), 1546-1550.
- Bottou, L., & Lin, C.-J. (2007). Support vector machine solvers. Large scale kernel machines, 301-320.
- Bouatmane, S., Roula, M., Bouridane, A., & Al-Maadeed, S. (2011). Round-Robin sequential forward selection algorithm for prostate cancer classification and diagnosis using multispectral imagery. *Machine Vision and Applications*, 22(5), 865-878. doi: 10.1007/s00138-010-0292-x
- Boundless. (2013). *Anatomy and Physiology*: Boundless.
- Bourke, P. (1995). RGB colour space. Retrieved 25/3/2012, from http://paulbourke.net/texture_colour/colourspace/

- Briggs, C., Longair, I., Slavik, M., Thwaite, K., Mills, R., Thavaraja, V., . . . Machin, S. J. (2009). Can automated blood film analysis replace the manual differential? An evaluation of the CellaVision DM96 automated image analysis system. *International Journal of Laboratory Hematology*, 31(1), 48-60. doi: 10.1111/j.1751-553X.2007.01002.x
- Brooks, M. L. (2008). *Exploring Medical Language: A Student-Directed Approach*: Elsevier Science Health Science Division.
- Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & chemistry*, 26(1), 5-14.
- Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2), 121-167.
- Burger, W., & Burge, M. J. (2009). Digital image processing: an algorithmic introduction using Java: Springer.
- Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences*, 43(6), 1882-1889.
- Cairo, M. S., & Perkins, S. L. (2012). *Hematological Malignancies in Children, Adolescents and Young Adults*: World Scientific.
- Caselato GR, Kobayashi E, Bonilha L, Castellano G, Rigas AH, Li LM, . . . F., C. (2003). Hippocampal texture analysis in patients with familial mesial temporal lobe epilepsy. *Arq Neuropsiquiat*, 61(1), 83-87.
- Castellano, G., Bonilha, L., Li, L. M., & Cendes, F. (2004). Texture analysis of medical images. *Clinical Radiology*, 59(12), 1061-1069.
- Ceelie, H., Dinkelaar, R., & van Gelder, W. (2007). Examination of peripheral blood films using automated microscopy; evaluation of Diffmaster Octavia and Cellavision DM96. *Journal of clinical pathology*, 60(1), 72-79.
- Celik, C., Aksel, J., & Karaoglan, B. (2006). Comparison of the Orpington Prognostic Scale (OPS) and the National Institutes of Health Stroke Scale (NIHSS) for the prediction of the functional status of patients with stroke. *Disability & Rehabilitation*, 28(10), 609-612.
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*.

- Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1), 1-6.
- Chen, D.-R., Chang, R.-F., Kuo, W.-J., Chen, M.-C., & Huang, Y.-L. (2002). Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks. *Ultrasound in medicine & biology*, 28(10), 1301-1310.
- Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, 17(1), 113-126.
- Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., & Lin, C. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, 60(1), 59-70. doi: <http://dx.doi.org/10.1016/j.neuroimage.2011.11.066>
- Ciesla, B. (2007). *Hematology in Practice*: Published by F.A. Davis Co.
- Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, 28(1), 45-62.
- Cohen, M. E., & Hudson, D. (1999). *Neural networks and artificial intelligence for biomedical engineering*: Wiley-IEEE Press.
- COLORROTATE. (2012). Retrieved 25/3/2012, from <http://learn.colorotate.org/color-models/#.VJo8jeCAG>
- Cornet, E., PEROL, J. P., & Troussard, X. (2008). Performance evaluation and relevance of the CellaVisionTM DM96 system in routine analysis and in patients with malignant hematological diseases. *International Journal of Laboratory Hematology*, 30(6), 536-542.
- Costa, E., Lorena, A., Carvalho, A., & Freitas, A. (2007). *A review of performance evaluation measures for hierarchical classifiers*. Paper presented at the Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop (pp. 1-6).
- Costa, L. d. F. D., & Cesar Jr, R. M. (2000). *Shape analysis and classification: theory and practice*: CRC Press, Inc.
- Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*: Cambridge university press.
- Das, K. (2013). *Clinical Medicine*: JP Medical Ltd.
- Deepajothi, S., & Selvarajan, S. (2013). Performance Evaluation of SVM–RBF Kernel for Classifying ECoG Motor Imagery. *International Journal of Computer Science and Telecommunications*, 4(5).
- Demir, C., & Yener, B. (2005). Automated cancer diagnosis based on histopathological images: a systematic survey. *Rensselaer Polytechnic Institute, Tech. Rep.*

- Désir, C., Petitjean, C., Heutte, L., & Thiberville, L. (2010). Using a Priori Knowledge to Classify in Vivo Images of the Lung. In D.-S. Huang, X. Zhang, C. Reyes García & L. Zhang (Eds.), *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence* (Vol. 6216, pp. 207-212): Springer Berlin Heidelberg.
- Dheeba, J., & Selvi, S. T. (2011). A CAD System for Breast Cancer Diagnosis Using Modified Genetic Algorithm Optimized Artificial Neural Network. In B. Panigrahi, P. Suganthan, S. Das & S. Satapathy (Eds.), *Swarm, Evolutionary, and Memetic Computing* (Vol. 7076, pp. 349-357): Springer Berlin Heidelberg.
- Diamond Diagnostics. (2013). Sysmex KX21N Hematology Analyzer. Retrieved 24/10/2013, from http://www.diamonddiagnostics.com/equipment/Hem/Sysmex_KX21n.htm
- Díaz, G., & Manzanera, A. (2011). Automatic Analysis of Microscopic Images in Hematological Cytology Applications. *Clinical Technologies: Concepts, Methodologies, Tools and Applications*, 325.
- Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02), 185-205.
- Döhner, H., Estey, E. H., Amadori, S., Appelbaum, F. R., Büchner, T., Burnett, A. K., . . . Larson, R. A. (2010). Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood*, 115(3), 453-474.
- Downey, K., Riches, S. F., Morgan, V. A., Giles, S. L., Attygalle, A. D., Ind, T. E., . . . deSouza, N. M. (2013). Relationship between imaging biomarkers of stage I cervical cancer and poor-prognosis histologic features: quantitative histogram analysis of diffusion-weighted MR images. *American Journal of Roentgenology*, 200(2), 314-320.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.
- Dunn, C. D. R., & Pallister, C. J. (1998). *Progress in Haematology*: Cambridge University Press.
- Dugdale, D. C., (2010). *Bone Marrow Aspiration*. Retrieved 2/3/2012 from: <http://health.allrefer.com/pictures-images/bone-marrow-aspiration.html>.
- Edward D. Ball, G. A. L. (2002). *100 Questions & Answers about Leukemia*: Jones & Bartlett Publishers.

- Erber, W. N. (2010). *Diagnostic Techniques in Hematological Malignancies* Cambridge University Press.
- Escalante, H. J., Montes-y-Gómez, M., González, J. A., Gómez-Gil, P., Altamirano, L., Reyes, C. A., . . . Rosales, A. (2012). Acute leukemia classification by ensemble particle swarm model selection. *55*(3), 163-175.
- Esposito, F., & Malerba, D. (2001). Guest Editorial: Machine Learning in Computer Vision. *Applied Artificial Intelligence: An International Journal*, *15*(8), 693-705.
- Esteridge, B. H., Reynolds, A. P., & Walters, N. J. (2000). *Basic medical laboratory techniques*: Cengage Learning.
- Estey, E. H., Faderl, S. H., & Kantarjian, H. M. (2007). *Hematologic Malignancies: Acute Leukemias*: Springer.
- Estridge, B. H., & Reynolds, A. P. (2011). *Basic clinical laboratory techniques*: Cengage Learning.
- Fabijańska, A., & Sankowski, D. (2009). Computer vision system for high temperature measurements of surface properties. *Machine Vision and Applications*, *20*(6), 411-421. doi: 10.1007/s00138-008-0135-1
- Falini, B., Tiacci, E., Martelli, M. P., Ascani, S., & Pileri, S. A. (2010). New classification of acute myeloid leukemia and precursor-related neoplasms: changes and unsolved issues. *Discovery medicine*, *10*(53), 281-292.
- Fausett, L. (1994). *Fundamentals of neural networks: architectures, algorithms, and applications*: Prentice-Hall, Inc.
- Felkel, P., Bruckschwaiger, M., & Wegenkittl, R. (2001). *Implementation and Complexity of the Watershed-from-Markers Algorithm Computed as a Minimal Cost Forest*. Paper presented at Computer Graphics Forum (Vol. 20, No. 3, pp. 26-35). Blackwell Publishers Ltd.
- Forsyth, D. A., & Ponce, J. (2003). *Computer Vision: A Modern Approach* (1st ed.): Prentice Hall.
- Gazzah, S., & Amara, N. B. (2008). Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script. *International Arab Journal of Information Technology (IAJIT)*, *5*(1).
- Greer, J. P., Arber, D. A., Glader, B., List, A. F., Means, R. T., Paraskevas, F., & Rodgers, G. M. (2013). *Wintrobe's Clinical Hematology* (13th ed.): Wolters Kluwer.
- GK, P., & Pravati, P. (2006). *Textbook Of Practical Physiology* (Vol. 2nd): Orient Blackswan.

- Gletsos, M., Mougiakakou, S. G., Matsopoulos, G. K., Nikita, K. S., Nikita, A. S., & Kelekis, D. (2003). A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier. *IEEE Transactions on Information Technology in Biomedicine*, 7(3), 153-162.
- Gökbuget, N., & Hoelzer, D. (2009). *Treatment of adult acute lymphoblastic leukemia*. Paper presented at the Seminars in hematology (Vol. 46, No. 1, pp. 64-75), Frankfurt, Germany.
- Gomez, W., Pereira, W. C. A., & Infantosi, A. F. C. (2012). Analysis of Co-Occurrence Texture Statistics as a Function of Gray-Level Quantization for Classifying Breast Ultrasound. *IEEE Transactions on Medical Imaging*, 31(10), 1889-1899. doi: 10.1109/tmi.2012.2206398
- Gonzalez, & Woods. (2002). *Digital Image Processing 2nd Edition* Prentice Hall
- Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2003). *Digital Image Processing Using MATLAB*: Published by Pearson Prentice Hall.
- Gonzalez, R. C., & Woods, R. E. (2008). *Digital Image Processing* (3rd ed.): Prentice Hall.
- Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R Journal*, 2(1), 30-38.
- Guyon., I., & Elisseeff., A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(15), 1157-1182.
- Hagen, H., Linsen, L., & Hamann, B. (2008). *Visualization in Medicine and Life Sciences*: Springer.
- Häggström, M. (2009). MedicineNet.com Leukemia (cont.), Symptoms. Retrieved 3/11/2009 from http://upload.wikimedia.org/wikipedia/commons/3/3a/Symptoms_of_leukemia.png
- Haifeng, S., & Lanlan, L. (2010, 4-15, August 2010). *Local Adaptive Image Enhancement Based on HSI Space* Paper presented at the Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCST '10), Jiaozuo, P. R. China.
- Halim, N. H. A., Mashor, M. Y., Abdul Nasir, A. S., Mokhtar, N. R., & Rosline, H. (2011, 4-6 March 2011). *Nucleus segmentation technique for acute Leukemia*. Paper presented at the Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on. Penang, Malaysia. doi: [10.1109/CSPA.2011.5759871](https://doi.org/10.1109/CSPA.2011.5759871)
- Halim., N. H. A., Mashor., M. Y., & Hassan., R. (2011b). Automatic Blasts Counting for Acute Leukemia Based on Blood Samples. *International Journal of Research and Reviews in Computer Science*, 2(4), 971-976.

- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.
- Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5), 786-804. doi: 10.1109/proc.1979.11328
- Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6), 610-621. doi: 10.1109/tsmc.1973.4309314
- Haralick, R., & Shanmugam, K. (1973). Computer classification of reservoir sandstones. *IEEE Transactions on Geoscience Electronics*, 11(4), 171-177.
- Harris, N. L., Jaffe, E. S., Diebold, J., Flandrin, G., Muller-Hermelink, H. K., Vardiman, J., . . . Bloomfield, C. D. (1999). World Health Organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Committee meeting—Airlie House, Virginia, November 1997. *Journal of Clinical Oncology*, 17(12), 3835-3849.
- Harun, N. H., M.Y.Mashor, Nasir, A. S. A., & H.Rosline. (2011). Automated Classification Of Blasts In Acute Leukemia Blood Samples Using HMLP Network. Paper presented at the Proceedings of the 3rd International Conference on Computing and Informatics, Bandung, Indonesia.
- Harun., N. H., Mashor, M. Y., Mokhtar, N. R., Aimi Salihah, A. N., Hassan, R., Raof, R. A. A., & Osman, M. K. (2010). Comparison of acute leukemia Image segmentation using HSI and RGB color space. Paper presented at the Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on.
- Hayati, M., & Shirvany, Y. (2007). Artificial neural network approach for short term load forecasting for Illam region. *World Academy of Science, Engineering and Technology*, 28, 280-284.
- He, X., & Liao, Q. (2008). *A novel shape prior based segmentation of touching or overlapping ellipse-like nuclei*. Paper presented at the Medical Imaging.
- Henseler, J. 1995. Back propagation. Artificial neural networks. Springer.
- Hirschmugl, M., Ofner, M., Raggam, J., & Schardt, M. (2007). Single tree detection in very high resolution remote sensing data. *Remote Sensing of Environment*, 110(4), 533-544.
- Hoffbrand, V., & Moss, P. (2011). *Essential Haematology* (Vol. 28): Wiley. Com
- Hoffbrand, V., Pettit, J.E., and Moss, P.A.H., (2001) *Essential Haematology*, Fourth Edition. Blackwell Science.

- Holli, K. K., Harrison, L., Dastidar, P., Wäljas, M., Liimatainen, S., Luukkaala, T., . . . Eskola, H. (2010). Texture analysis of MR images of patients with mild traumatic brain injury. *BMC medical imaging*, 10(1), 8.
- Hossain, M. R., Oo, A. M. T., & Ali, A. B. M. S. (2013). The Effectiveness of Feature Selection Method in Solar Power Prediction. *Journal of Renewable Energy*, 2013, 9. doi: 10.1155/2013/952613
- Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification. *Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003*. Retrieved from <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>
- Huey Nee, L., Mashor, M. Y., & Hassan, R. (2012). *White blood cell segmentation for acute leukemia bone marrow images*. Paper presented at the Biomedical Engineering (ICoBE), 2012 International Conference on. Penang, Malaysia. doi: 10.1109/ICoBE.2012.6179038
- Hwang, S., Oh, J., Tavanapong, W., Wong, J., & de Groen, P. C. (2007). Automatic polyp region segmentation for colonoscopy images using watershed algorithm and ellipse segmentation. Paper presented at the Medical Imaging (pp. 65141D-65141D). International Society for Optics and Photonics.
- Ichihashi, T., Naoe, T., Kuriyama, K., Sasada, M., & Ohno, R. Acute Lymphocytic Leukemia Retrieved 24/10/ 2013, from <http://pathy.med.nagoya-u.ac.jp/atlas/img/t6/img028.jpg>
- Image Processing Toolbox (R2014b), MATLAB (Matrix Laboratory). Mathworks, 2014. Retrieved 24/11/ 2014, from <http://www.mathworks.com/help/toolbox/images/>
- Intel Developer Zone, (2012). Color Models. Retrieved 25/3/2012, 2012, from <https://software.intel.com/en-us/node/503873>
- Ismail, W., Hassan, R., & Swift, S. (2010). Detecting Leukaemia (AML) Blood Cells Using Cellular Automata and Heuristic Search. In P. Cohen, N. Adams & M. Berthold (Eds.), *Advances in Intelligent Data Analysis IX* (Vol. 6065, pp. 54-66): Springer Berlin Heidelberg.
- Jack, K. (2005). *Video demystified: a handbook for the digital engineer*. Newnes.
- Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1), 4-37.
- James, D., Clymer, B. D., & Schmalbrock, P. (2001). Texture detection of simulated microcalcification susceptibility effects in magnetic resonance imaging of breasts. *Journal of Magnetic Resonance Imaging*, 13(6), 876-881.

- Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429-449.
- Jaroszeski., M. J., & Heller., R. (1998). *Flow cytometry protocols*: Springer.
- Jayaraman S, Esakkirajan S, & Veerakumar T. (2011). *Digital image processing*: Tata McGraw-Hill Education.
- Ji, Q., Engel, J., & Craine, E. (2000). Texture analysis for classification of cervix lesions. *Medical Imaging, IEEE Transactions on*, 19(11), 1144-1149.
- Jiang, J., Trundle, P., & Ren, J. (2010). Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 34(8), 617-631.
- Jiao, L., Chen, Q., Li, S., & Xu, Y. (2013, January). Colon Cancer Detection Using Whole Slide Histopathological Images. Paper Presented at the World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China (pp. 1283-1286). Springer Berlin Heidelberg.
- John, G. H. (1994). Cross-Validated C4.5: Using Error Estimation for Automatic Parameter Selection: Technical Report Stanford University. Retrieved from <http://i.stanford.edu/pub/cstr/reports/cs/tn/94/12/CS-TN-94-12.pdf>
- Johnson, P. (2010). History of Leukemia. Retrieved 15/3/2010, from <http://www.buzzle.com/articles/history-of-leukemia.html>
- Joshi, M. M. D., & Karode, A. (2013). Detection of Acute Leukemia Using White Blood Cells Segmentation Based on Blood Samples. *International journal of Electronics and Communication Engineering & Technology*, 4(1), 148-153.
- Kampen, K. R. (2012). The discovery and early understanding of leukemia. *Leukemia Research*, 36(1), 6-13. doi: <http://dx.doi.org/10.1016/j.leukres.2011.09.028>
- Kaensar, C. (2013). A Comparative Study on Handwriting Digit Recognition Classifier Using Neural Network, Support Vector Machine and K-Nearest Neighbor. Paper presented at the The 9th International Conference on Computing and Information Technology (IC2IT2013).
- Karthikeyan, S., & Rengarajan, N. (2014). Performance analysis of gray level co-occurrence matrix texture features for glaucoma diagnosis. *Am. J. Applied Sci*, 11, 248-257.
- Kawthalkar, S. M. (2012). *Essentials of haematology*: Jaypee Brothers Medical P.
- Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural computation*, 15(7), 1667-1689.

- Khashman, A., & Al-Zgoul, E. (2009). *Image segmentation of blood cells in leukemia patients*. Paper presented at the Proceedings of the 4th WSEAS international conference on Computer engineering and applications, Cambridge, USA.
- Khot S.T, Bhalekar, S., Jaggi, D., & Rani, D. (2013). An Innovative Approach in Myelogenous Leukemia Detection using Attribute Analysis. *International Journal of Advanced Electrical and Electronics Engineering*, 2(5), 43-46.
- Kizrak, M. A., & Ozena, F. (2012). Automatic Acute Lymphocytic Leukemia Diagnosis Based on Kernel Ridge Regression Method *AWERProcedia Information Technology & Computer Science*, 2, 50-55.
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1), 273-324.
- Kothari., R., Cualing., H., & Balachander., T. (1996). Neural Network Analysis of Flow Cytometry Immunophenotype Data. *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, 43(8), 803-810.
- Kubat, M., & Matwin, S. (1997). *Addressing the curse of imbalanced training sets: one-sided selection*. Paper presented at the In Proceedings of the Fourteenth International Conference on Machine Learning.
- Kumar, A. (2008). Computer-vision-based fabric defect detection: a survey. *Industrial Electronics, IEEE Transactions on*, 55(1), 348-363.
- Kumar, R. S., Verma, A., & Singh, J. (2006). Color Image Segmentation and Multi-Level Thresholding by Maximization of Conditional Entropy. *International Journal of Signal Processing*, 3(1), 15-16.
- Kumar., V., Abul K. Abbas, Fausto., N., & Mitchell., R. (2007). *Robbins Basic Pathology* (8th Edition ed.).
- Kuncheva, L. (2004). *Combining Pattern Classifiers: Methods and Algorithms*: John Wiley & Sons.
- Labati, R. D., Piuri, V., & Scotti, F. (2011). *All-IDB: The acute lymphoblastic leukemia image database for image processing*. Paper presented at the 18th IEEE International Conference on Image Processing (ICIP). Brussels, Belgium. doi: [10.1109/ICIP.2011.6115881](https://doi.org/10.1109/ICIP.2011.6115881)
- Lai, D. T., Pakkanen, J., Begg, R., & Palaniswami, M. (2008). Computational Intelligence and Sensor Networks for Biomedical Systems. *Encyclopedia of healthcare information systems*.
- Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain*, 8(6), 221-223.

- Lantuéjoul, C., & Beucher, S. (1981). On the use of the geodesic metric in image analysis. *Journal of Microscopy*, 121(1), 39-49.
- Lavelle., P. (2004). Leukaemia. Retrieved 19/3/2012, from <http://www.abc.net.au/health/library/stories/2004/10/18/1830091.htm>
- Le, M.-T., Bretschneider, T., Kuss, C., & Preiser, P. (2008). A novel semi-automatic image processing approach to determine Plasmodium falciparum parasitemia in Giemsa-stained thin blood smears. *BMC Cell Biology*, 9(1), 15.
- Li, J., Chen, Z., Wei, L., Xu, W., & Kou, G. (2007). Feature selection via least squares support feature machine. *International Journal of Information Technology & Decision Making*, 6(04), 671-686.
- Liao, Q., & Deng, Y. (2002). An accurate segmentation method for white blood cell images. *IEEE International Symposium on Biomedical Imaging, Proceedings*, 245-248. doi: 10.1109/ISBI.2002.1029239
- Lim, G. C. C., Rampal, S., & Yahaya, H. (2008). *Cancer Incidence in Peninsular Malaysia, 2003-2005: The Third Report of the National Cancer Registry, Malaysia*: National Cancer Registry.
- Lin, K.-M., & Lin, C.-J. (2003). A study on reduced support vector machines. *Neural Networks, IEEE Transactions on*, 14(6), 1449-1459.
- Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4), 491-502.
- Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(2), 539-550.
- Lofsness. (2008). Blood Cell Maturation. Retrieved 21/10/2013, from <http://www1.umn.edu/hema/pages/matchart.html>
- Luo, S.-T., & Cheng, B.-W. (2012). Diagnosing breast masses in digital mammography using feature selection and ensemble methods. *Journal of Medical Systems*, 36(2), 569-57
- Madhloom, H., Kareem, S., & Ariffin, H. (2012). An Image Processing Application for the Localization and Segmentation of Lymphoblast Cell Using Peripheral Blood Images. *Journal of Medical Systems*, 36(4), 2149-2158. doi: 10.1007/s10916-011-9679-0

- Madhlloom, H. T., Kareem, S. A., & Ariffin, H. (2012b). *A Robust Feature Extraction and Selection Method for the Recognition of Lymphocytes versus Acute Lymphoblastic Leukemia*. Paper presented at the Advanced Computer Science Applications and Technologies (ACSAT), 2012 International Conference on. Kuala Lumpur, Malaysia. doi: [10.1109/ACSAT.2012.62](https://doi.org/10.1109/ACSAT.2012.62)
- Madhlloom, H. T., Kareem, S. A., Ariffin, H., Zaidan, A. A., Alanazi, H. O., & Zaidan, B. B. (2010). An Automated White Blood Cell Nucleus Localization and Segmentation using Image Arithmetic and Automatic Threshold. *Journal of Applied Sciences*, 10(11), 959-966.
- Madhukar, M. (2012). *Microscopic Image Segmentation and Classification of Acute Leukemia*. (MSc.), University of Texas at San Antonio.
- Madhukar, M., Agaian, S., & Chronopoulos, A. T. (2012b). *Deterministic model for Acute Myelogenous Leukemia classification*. Paper presented at the Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on
- Manfred R. Koller, Bernhard Palsson, & Masters, J. R. W. (1999). *Primary hematopoietic cells*: Springer.
- Mao-jun, S., Zhao-bin, W., Hong-juan, Z., & Yi-de, M. (2008). *A new method for blood cell image segmentation and counting based on PCNN and autowave*. Paper presented at the 3rd International Symposium on Communications, Control and Signal Processing., St Julians, Malta. doi: [10.1109/ISCCSP.2008.4537182](https://doi.org/10.1109/ISCCSP.2008.4537182)
- Marcano, C., x00F, o, A., Quintanilla-dominguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). *Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network*. Paper presented at the IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society.
- Markey, M. K., Boland, M. V., & Murphy, R. F. (1999). Toward Objective Selection of Representative Microscope Images. *Biophysical Journal*, 76(4), 2230-2237. doi: [http://dx.doi.org/10.1016/S0006-3495\(99\)77379-0](http://dx.doi.org/10.1016/S0006-3495(99)77379-0)
- Markiewicz, T., Osowski, S., Marianska, B., & Moszczynski, L. (2005). *Automatic recognition of the blood cells of myelogenous leukemia using SVM*. Paper presented at the Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on.
- Markiewicz., T., & Osowski., S. (2006). *Data Mining Techniques for Feature Selection in Blood Cell Recognition*. Paper presented at the ESANN'2006 proceedings, European Symposium on Artificial Neural Networks, Bruges (Belgium).
- Markiewicz., T., & Osowski., S. (2006). *Data Mining Techniques for Feature Selection in Blood Cell Recognition*. Paper presented at the ESANN'2006 proceedings, European Symposium on Artificial Neural Networks, Bruges, Belgium.

- Marques, O. (2011). *Practical image and video processing using MATLAB*: John Wiley & Sons
- Mcgaufflin, S., Munger, J and Nelson, R,. (2005). Chronic Myelogenous Leukemia. Retrieved (22 /10/2013). from <http://rebeccanelson.com/leukemia/cml.html>.
- Mennicke, J., Münzenmayer, C., Wittenberg, T., & Schmid, U. (2009). An optimization framework for classifier learning from image data for computer-assisted diagnosis. In J. Sloten, P. Verdonck, M. Nyssen & J. Haueisen (Eds.), *4th European Conference of the International Federation for Medical and Biological Engineering* (Vol. 22, pp. 629-632): Springer Berlin Heidelberg.
- Mohapatra, S., & Patra, D. (2010, 16-18 Dec. 2010). *Automated cell nucleus segmentation and acute leukemia detection in blood microscopic images*. Paper presented at the Systems in Medicine and Biology (ICSMB), 2010 International Conference on. Kharagpur, India. doi: [10.1109/ICSMB.2010.5735344](https://doi.org/10.1109/ICSMB.2010.5735344)
- Mohapatra, S., Patra, D., & Kumar, K. (2011). *Blood microscopic image segmentation using rough sets*. Paper presented at the Image Information Processing (ICIIP), 2011 International Conference on.
- Mohapatra, S., Patra, D., & Kumar, K. (2011c). *Blood microscopic image segmentation using rough sets*. Paper presented at the Image Information Processing (ICIIP), 2011 International Conference on.
- Mohapatra, S., Patra, D., & Satpathi, S. (2010b). *Image analysis of blood microscopic images for acute leukemia detection*. Paper presented at the Industrial Electronics, Control & Robotics (IECR), 2010 International Conference on, Orissa, India. doi:10.1109/IECR.2010.5720171
- Mohapatra, S., Patra, D., & Satpathy, S. (2011). *Automated leukemia detection in blood microscopic images using statistical texture analysis*. Paper presented at the Proceedings of the 2011 International Conference on Communication, Computing & Security, Rourkela, Odisha, India.
- Mohapatra, S., Patra, D., & Satpathy, S. (2013). An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Computing and Applications*, 1-18.
- Mohapatra, S., Samanta, S. S., Patra, D., & Satpathi, S. (2011b). *Fuzzy Based Blood Image Segmentation for Automated Leukemia Detection*. Paper presented at the Devices and Communications (ICDeCom), 2011 International Conference on.
- Mona Lisa, (2012). Retrieved 20/3/2012, from http://commons.wikimedia.org/wiki/Category:Mona_Lisa

- Mora-González, M., Muñoz-Maciel, J., Casillas, F. J., Peña-Lecona, F. G., Chiu-Zarate, R., & de Guevara, H. P. L. (2011). Image Processing for Optical Metrology. *Chap, 25*, 523-546.
- Najarian., K., & Splinter., R. (2012). *Biomedical signal and image processing*: CRC Press.
- Nandagopalan, S., Dhanalakshmi, C., Adiga, B. S., & Deepak, N. (2010). *Color doppler echocardiographic image analysis via shape and texture features*. Paper presented at the Bioinformatics and Biomedical Technology (ICBBT), 2010 International Conference on, Chengdu, China. doi: 10.1109/ICBBT.2010.5478992
- Nasir, A. A., Mashor, M. Y., & Hassan, R. (2013). Classification of Acute Leukaemia Cells Using Multilayer Perceptron and Simplified Fuzzy ARTMAP Neural Networks. *International Arab Jornal of Information Technology, 10*(5).
- Nasir, A. A., Mustafa, N., & Nasir, N. M. (2009). *Application of Thresholding Technique in Determining Ratio of Blood Cells for Leukemia Detection*. Paper presented at the 2009 Proceedings of the International Conference on Man-Machine Systems (ICoMMS). Batu Ferringhi, Penang, Malaysia.
- National Cancer Institute. Classification of Adult Acute Myeloid Leukemia. Retrieved 23/3/2014, from <http://www.cancer.gov/cancertopics/pdq/treatment/adultAML/healthprofessional/page2>
- Nee, L. H., Mashor, M. Y., & Hassan, R. (2012). White Blood Cell Segmentation for Acute Leukemia Bone Marrow Images. *Journal of Medical Imaging and Health Informatics, 2*, 278-284. doi: <http://dx.doi.org/10.1166/jmihi.2012.1099>
- Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*: Pearson Education.
- Nemane., J. B., & Chakkarwar., V. A. (2013). A Novel Method of White Blood Cell Segmentation and Counting *International Journal on Advanced Computer Engineering and Communication Technology, 1*(1), 44-49.
- Ng, H. P., Su, H., Ong, S. H., Foong, K. W. C., Goh, P. S., & Nowinski, W. L. (2008). *Medical image segmentation using watershed segmentation with texture-based region merging*. Paper presented at the Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, Vancouver, BC. doi: 10.1109/IEMBS.2008.4650096
- Ng, H. P., Ong, S. H., Foong, K. W. C., Goh, P. S., & Nowinski, W. L. (2006). *Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm*. Paper presented at the Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on, Denver, CO. doi: 10.1109/SSIAI.2006.1633722

- Nguyen, G. Hoang., Bouzerdoum, A. & Phung, S. (2009). Learning pattern classification tasks with imbalanced data sets. In P. Yin (Eds.), *Pattern recognition* (pp. 193-208). Vukovar, Croatia: In-Teh.
- Nikoo, H., Talebi, H., & Mirzaei, A. (2011, 16-17 Nov. 2011). *A Supervised Method for Determining Displacement of Gray Level Co-Occurrence Matrix*. Paper presented at the Machine Vision and Image Processing (MVIP), 2011 7th Iranian, Tehran, Iran. doi: 10.1109/IranianMVIP.2011.6121563
- Nithya, R., & Santhi, B.. (2011). Comparative Study On Feature Extraction Method For Breast Cancer Classification. *Journal of Theoretical and Applied Information Technology*, 33(2), 220-226.
- Norman, B. (2009). *Diagnostic Hematology* (1st ed.): Springer.
- Ochiai, F., & Eguchi, M. (1986). Morphometrical evaluation of acute leukemic cells by electron microscopy. *Virchows Archiv B*, 52(1), 403-411. doi: 10.1007/BF02889981
- Orazi, A., O'Malley, D. P., & Arber, D. A. (2006). *Illustrated pathology of the bone marrow*: Cambridge University Press
- Orozco-Monteagudo, M., Geerinck, T., Mihai, C., Sahli, H., & Taboada-Crispi, A. (2013). Pap-Smear Cell Nucleus Extraction by Using Morphological Watershed. In J. Folgueras Méndez, T. Y. Aznielle Rodríguez, C. F. Calderón Marín, S. B. Llanusa Ruiz, J. Castro Medina, H. Vega Vázquez, M. Carballo Barrera & R. Rodríguez Rojas (Eds.), *V Latin American Congress on Biomedical Engineering CLAIB 2011 May 16-21, 2011, Habana, Cuba* (Vol. 33, pp. 1070-1073): Springer Berlin Heidelberg.
- Ortiz, F., Torres, F., De Juan, E., & Cuenca, N. (2002). Colour mathematical morphology for neural image analysis. *Real-Time Imaging*, 8(6), 455-465.
- Osowski, S., Robert Siroi, c., Markiewicz, T., & Siwek, K. (2009). Application of Support Vector Machine and Genetic Algorithm for Improved Blood Cell Recognition. *IEEE Transactions on Instrumentation And Measurement*, 58(7), 2159-2168.
- Osowski, S., Markiewicz, T., Marianska, B., & Moszczyński, L. (2004). *Feature Generation For the Cell Image Recognition of Myelogenous Leukemia* Paper presented at the IEEE International Conference EUSIPCO, Vienna, Austria
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62-66.
- Pandey, S., & Singh, M. (2010). *Study and implementation of morphology for image segmentation*. (Master Dissertatiation , Thapar University, Patiala, India). Retrieved from <http://dspace.thapar.edu:8080/dspace/bitstream/10266/1276/3/1276.pdf>

- Patel, P. B., & Marwala, T. (2006). *Neural networks, fuzzy inference systems and adaptive-neuro fuzzy inference systems for financial decision making*. Paper presented at the Neural Information Processing.
- Patil., R., Sohani., M., Bojewar., S., & Tatiya., P. (2012). *Acute Leukemia blast counting using RGB, HSI color spaces*. Paper presented at the IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012)
- Paul D. O'Malley. (2006). *New developments in sickle cell disease research*: Nova Publishers.
- Pauline's Cancer Awareness Foundation. (2013). General Overview of Leukaemia. Retrieved 22/10/2103, 2013, from <http://www.paulinecancerfoundation.org/s/resources/p/knowledge-base/b/overview-of-leukaemia.html>
- Pavlidis, T. (1978). A review of algorithms for shape analysis. *Computer graphics and image processing*, 7(2), 243-258.
- Pilgrim., G. (2010). Leukemia Prognosis. Retrieved 25/3/2012, from <http://www.buzzle.com/articles/leukemia-prognosis.html>
- Pise, S. J., Kekre, H. B., Sarode, T., Thepade, S., & Halarnkar, P. Kekre's (2010) fast codebook generation in VQ with various color spaces for colorization of grayscale images *Thinkquest~2010* (pp. 102-108): Springer India.
- Piuri, V., & Scotti, F. (2004). *Morphological classification of blood leucocytes by microscope images*. Paper presented at the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications.
- Plissiti, M. E., Nikou, C., & Charchanti, A. (2010). *Watershed-based segmentation of cell nuclei boundaries in Pap smear images*. Paper presented at the Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on.
- Pommerville, J. C. (2009). *Alcamo's Fundamentals of Microbiology: Body Systems*: Jones & Bartlett Publishers.
- Pui, C.-H. (2012). *Childhood leukemias*: cambridge university press.
- Pui, C.-H. (2003). *Treatment of acute leukemias: new directions for clinical research*: Humana Press.
- Putzu, L., & Di Ruberto, C. (2013). *White Blood Cells Identification and Classification from Leukemic Blood Image*. Paper presented at the Proceedings of the IWBBIO International Work-Conference on Bioinformatics and Biomedical Engineering.

- Putzu, L., & Ruberto, C. (2013a). Investigation of Different Classification Models to Determine the Presence of Leukemia in Peripheral Blood Image. In A. Petrosino (Ed.), *Image Analysis and Processing – ICIAP 2013* (Vol. 8156, pp. 612-621): Springer Berlin Heidelberg.
- Putzu, L., & Ruberto, C. D. (2013b). White Blood Cells Identification and Counting from Microscopic Blood Image. *International Journal of Medical, Pharmaceutical Science and Engineering*, 73(1), 15 - 22.
- Qian, C. (2004). *Using Classification and Regression Tree to Detect Hematology Abnormalities* (Master dissertation, Florida Atlantic University). Retrieved from <http://digitool.fcla.edu>.
- Rajendran, S., Arof, H., Ibrahim, F., & Yegappan, S. (2008). Review on Technical Aspects of Image Acquisition, Analysis and Retrieval for Leukaemia Cells. In N. Abu Osman, F. Ibrahim, W. Wan Abas, H. Abdul Rahman & H.-N. Ting (Eds.), *4th Kuala Lumpur International Conference on Biomedical Engineering 2008* (Vol. 21, pp. 230-233): Springer Berlin Heidelberg.
- Reaman., G. H. (2011). *Childhood Leukemia: A Practical Handbook*: Springer.
- Ren, J., Qiu, Z., Fan, W., Cheng, H., & Yu, P. (2008). Forward Semi-supervised Feature Selection. In T. Washio, E. Suzuki, K. Ting & A. Inokuchi (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 5012, pp. 970-976): Springer Berlin Heidelberg.
- Reta, C., Altamirano, L., Gonzalez, J. A., Diaz, R., & S.Guichard, J. (2010). *Segmentation of Bone Marrow Cell Images for Morphological Classification of Acute Leukemia*. Paper presented at the Twenty-Third International Florida Artificial Intelligence Research Society Conference, Florida,USA.
- Rezatofghi, S. H., & Soltanian-Zadeh, H. (2011). Automatic recognition of five types of white blood cells in peripheral blood. *Computerized Medical Imaging and Graphics*, 35(4), 333-343.
doi:<http://dx.doi.org/10.1016/j.compmedimag.2011.01.003>
- Riley, R. S., James, G. W., Sommer, S., & Martin, M. J. (2012). How to Prepare & Interpret Peripheral Blood Smears. Retrieved 19/2/2012, from <http://www.pathology.vcu.edu/education/PathLab/pages/hematopath/pbs.html>
- Riley, R. S., & Ben-Ezra, J. (1999). Leukemias and Multiple Myeloma. Retrieved 24/9/2013, from <http://www.pathology.vcu.edu/education/PathLab/pages/hematopath/leukho99.html>
- RL Bijlani, M., & S Manjunatha, M. (2010). *Understanding Medical Physiology: A Textbook for Medical Students*: JP Medical Ltd.

- Rocha, H. (2008). Model parameter tuning by cross validation and global optimization: application to the wing weight fitting problem. *Structural and Multidisciplinary Optimization*, 37(2), 197-202.
- Rodak, B., Fritsma, G., & Doig, K. (2007). *Hematology: Clinical Principles & Applications*: Elsevier Health Sciences.
- Rodenacker, K., & Bengtsson, E. (2003). A feature set for cytometry on digitized microscopic images. *Analytical Cellular Pathology*, 25(1), 1-36.
- Rubin, R., Strayer, D. S., Rubin, E., & McDonald, J. M. (2008). *Rubin's pathology: clinicopathologic foundations of medicine*: Lippincott Williams & Wilkins.
- Rüger., S. (2010). *Multimedia Information Retrieval*: Morgan & Claypool Publishers
- Sabino, D. M. U., Costa, L. F., Rizzatti, E. G., & Zago, M. A. (2004, 15-18 April 2004). *Toward leukocyte recognition using morphometry, texture and color*. Paper presented at the Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on.
- Sabino, D., Costa, L. d. F., Martins, S., Calado, R., & Zago, M. (2003). Automatic leukemia diagnosis. *Acta Microscopica*, 12(1), 1-6.
- Sachdeva, M., Ahluwalia, J., Das, R., Varma, N., & Garewal, G. (2006). Role of FAB classification of acute leukemias in era of immunophenotyping. *Indian journal of pathology & microbiology*, 49(4), 524-527.
- Sadeghian., F., Seman., Z., Ramli., A. R., Kahar., B. H. A., & Saripan., M.-I. (2009). A Framework for White Blood Cell Segmentation in Microscopic Blood Images Using Digital Image Processing. *Biological Procedures Online*, 11(1), 196-206. doi: 10.1007/s12575-009-9011-2
- Saeys Y, Inza I, & Larrañaga P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.
- Sahin, S., & Sumnu, S. G. (2006). *Physical Properties of Foods*: Springer.
- Schrier, R. W. (2007). *The Internal Medicine Casebook: Real Patients, Real Answers* (Third ed.): Published by Lippincott Williams & Wilkins.
- Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*: MIT Press.
- Scotti, F. (2006). *Robust Segmentation and Measurements Techniques of White Cells in Blood Microscope Images*. Paper presented at the Instrumentation and Measurement Technology Conference, 2006. IMTC 2006. Proceedings of the IEEE. Sorrento. doi: [10.1109/IMTC.2006.328170](https://doi.org/10.1109/IMTC.2006.328170)

- Scotti, F. (2005). *Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images*. Paper presented at the 2005 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications. doi: 10.1109/CIMSA.2005.1522835
- Serra, J. P. (1982). *Image analysis and mathematical morphology* (Vol. 1): Academic Press.
- Selvarajah, S., & Kodituwakku, S. (2011). Analysis and comparison of texture features for content based image retrieval. *International Journal of Latest Trends in Computing*, 2(1).
- Shah, S. K., McNitt-Gray, M. F., Rogers, S. R., Goldin, J. G., Suh, R. D., Sayre, J. W., . . . Aberle, D. R. (2005). Computer Aided Characterization of the Solitary Pulmonary Nodule Using Volumetric and Contrast Enhancement Features< sup> 1</sup>. *Academic Radiology*, 12(10), 1310-1319.
- Shafait, F., Reif, M., Kofler, C., & Breuel, T. M. (2010). *Pattern Recognition Engineering*. Paper presented at the RapidMiner Community Meeting and Conference.
- Shafeeq, A., & Hareesha, K. (2012). Dynamic clustering of data with modified k-means algorithm. Paper presented at the Proceedings of the 2012 conference on information and computer networks.
- Shaikh, S. H., Maiti, A. K., & Chaki, N. (2011). On Creation of Reference Image for Quantitative Evaluation of Image Thresholding Method Computer Information Systems–Analysis and Technologies (pp. 161-169): Springer.
- Sharif, J. M., Miswan, M. F., Ngadi, M. A., Salam, M. S. H., & Mahadi bin Abdul Jamil, M. (2012, 27-28 Feb. 2012). *Red blood cell segmentation using masking and watershed algorithm: A preliminary study*. Paper presented at the Biomedical Engineering (ICoBE), 2012 International Conference on, Penang , Malaysia. doi: 10.1109/ICoBE.2012.6179016
- Sheikh., H., Zhu., B., & -Tzanakou, E. M. (1996). Blood Cell Identification Using Neural Networks
- Shamsuddin, S. M., Sallehuddin, R., & Yusof, N. M. (2008). Artificial Neural Network Time Series Modeling for Revenue Forecasting. *Chiang Mai J. Sci*, 35(3), 411-426.
- Shenouda, E. M. A. (2006). A Quantitative Comparison of Different MLP Activation Functions in Classification. In J. Wang, Z. Yi, J. Zurada, B.-L. Lu & H. Yin (Eds.), *Advances in Neural Networks - ISNN 2006* (Vol. 3971, pp. 849-857): Springer Berlin Heidelberg.
- Shih, F., King, C. T., & Pu, C. C. (1995). Pipeline architectures for recursive morphological operations. *Image Processing, IEEE Transactions on*, 4(1), 11-18.

- Sigurdsson, S., Hansen, L. K., & Drzewiecki, K. T. (2003). Color segmentation of skin lesions with the generalizable Gaussian mixture model *MM, Informatik og Matematisk Modellering, DTU*.
- Sinha, S., Lucas-Quesada, F. A., Debruhl, N. D., Sayre, J., Farria, D., Gorczyca, D. P., & Bassett, L. W. (1997). Multifeature analysis of Gd-enhanced MR images of breast lesions. *Journal of Magnetic Resonance Imaging*, 7(6), 1016-1026.
- Silverstein, A., Silverstein, V. B., & Nunn, L. S. (2006). *Cancer: Twenty-First Century Books*.
- Sinha, N., & Ramakrishnan, A. (2003). *Automation of differential blood count*. Paper presented at the TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region. doi: [10.1109/TENCON.2003.1273221](https://doi.org/10.1109/TENCON.2003.1273221)
- Sivanandam, S., & Deepa, S. (2006). *Introduction to neural networks using Matlab 6.0: Tata McGraw-Hill Education*.
- Siroic, R., Osowski, S., Markiewicz, T., & Siwek, K. (2007). *Support Vector Machine and Genetic Algorithm for Efricient Blood Cell Recognition*. Paper presented at the IEEE Instrumentation and Measurement Technology Conference Proceedings. Warsaw, Poland. doi: [10.1109/IMTC.2007.379321](https://doi.org/10.1109/IMTC.2007.379321).
- Skarbek, W., Koschan, A., & Veroffentlichung, Z. (1994). Colour image segmentation-a survey. *Institute for Technical Informatics, Technical University of Berlin*.
- Smith, A. R. (1978, August 1978). *Color gamut transform pairs*. Paper presented at the Proceeding SIGGRAPH '78 Proceedings of the 5th annual conference on Computer graphics and interactive techniques New York, NY, USA.
- Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222. doi: [10.1023/B:STCO.0000035301.49549.88](https://doi.org/10.1023/B:STCO.0000035301.49549.88)
- Srinivasan, G., & Shobha, G. (2008). Statistical Texture Analysis. *Proceedings of World Academy of Science: Engineering & Technology*, 48.
- Soh, L.-K., & Tsatsoulis, C. (1999). Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(2), 780-795.
- Soloman, S. (2009). *Sensors and Control Systems in Manufacturing* (2nd ed): McGraw Hill Professional.
- Sonka, M., Hlavac, V., & Boyle, R. (2014). *Image processing, analysis, and machine vision: Cengage Learning*.
- Sormunen, C. (2009). *Terminology For Health Professionals: Cengage Learning*.

- Sovierzoski, M. A., Schwarz, L., & Azevedo, F. M. (2010). Evaluation of Benchmark Indexes to Determine the Best Performance of a Binary Neural Classifier. In O. Dössel & W. Schlegel (Eds.), *World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany* (Vol. 25/4, pp. 464-467): Springer Berlin Heidelberg.
- Starr, C., Evers, C. A., & Starr, L. (2007). *Biology Today and Tomorrow: With Physiology*: CengageBrain. com.
- Stevens, M. L. (1997). *Fundamentals of clinical hematology* (1e ed.): Medical
- Suematsu, N., Ishida, Y., Hayashi, A., & Kanbara, T. (2002). *Region-based image retrieval using wavelet transform*. Paper presented at the Proc. 15th international conf. on vision interface.
- Supardi, N. Z., Mashor, M. Y., Harun, N. H., Bakri, F. A., & Hassan, R. (2012, 23-25 March 2012). *Classification of blasts in acute leukemia blood samples using k-nearest neighbour*. Paper presented at the Signal Processing and its Applications (CSPA), 2012 IEEE 8th International Colloquium on. Melaka, Malaysia. doi: [10.1109/CSPA.2012.6194769](https://doi.org/10.1109/CSPA.2012.6194769)
- Sun, Y. (2011). Neural Networks for Emotion Classification. *arXiv preprint arXiv:1105.6014*.
- Sun, H., & Luo, Y. (2009). Adaptive watershed segmentation of binary particle image. *Journal of Microscopy*, 233(2), 326-330.
- Sun, Y., Kamel, M. S., & Wang, Y. (2006). *Boosting for Learning Multiple Classes with Imbalanced Class Distribution*. Paper presented at the Proceedings of the Sixth International Conference on Data Mining.
- Szczepański, T., van der Velden, V. H. J., & van Dongen, J. J. M. (2003). Classification systems for acute and chronic leukaemias. *Best Practice & Research Clinical Haematology*, 16(4), 561-582. doi: [http://dx.doi.org/10.1016/S1521-6926\(03\)00086-0](http://dx.doi.org/10.1016/S1521-6926(03)00086-0)
- Tecklin, J. S. (2008). *Pediatric physical therapy*: Lippincott Williams & Wilkins.
- Theera-Umpon, N., & Dhompongsa, S. (2007). Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification. *IEEE Transactions on Information Technology in Biomedicine*, 11(3), 353-359.
- Theml, H., Diem, H., & Haferlach, T. (2004). *Color Atlas of Hematology*. New York: Thieme Stuttgart.
- Theodoridis, S., Pikrakis, A., Koutroumbas, K., & Cavouras, D. (2010). *Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach*: Academic Press.

- Tian., Y., Guan., T., & Wang., C. (2010). Real-Time Occlusion Handling in Augmented Reality Based on an Object Tracking Approach. *Sensors*, 10(4), 2885–2900.
- Tkachuk, D. C., Hirschmann, J. V., & Wintrobe, M. M. (2007). *Wintrobe's Atlas of Clinical Hematology*: Lippincott Williams & Wilkins.
- Tsagaris, V., & Anastassopoulos, V. (2004, February 5, 2004). *Multispectral image fusion method using perceptual attributes*. Paper presented at the Proc. SPIE 5238 Image and Signal Processing for Remote Sensing IX 357, Barcelona, Spain. doi:10.1117/12.510651
- Tkachuk, D. C., Hirschmann, J. V., & Wintrobe, M. M. (2007). *Wintrobe's Atlas of Clinical Hematology*: Lippincott Williams & Wilkins.
- Truchon, J.-F., & Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *Journal of chemical information and modeling*, 47(2), 488-508.
- Tuceryan, M., & Jain, A. K. (1993). Texture analysis. *Handbook of pattern recognition and computer vision*, 2, 207-248.
- Umbaugh, S. E. (2010). *Digital Imaging Processing and Analysis: Human and Computer Vision Applications with CVIPtools, Second Edition*: CRC Press.
- Ushizima, D. M., Lorena, A. C., & De Carvalho, A. C. P. L. F. (2005). *Support vector machines applied to white blood cell recognition*. Paper presented at the Hybrid Intelligent Systems, 2005. HIS '05. Fifth International Conference on. Rio de Janeiro, Brazil. doi: [10.1109/ICHIS.2005.100](https://doi.org/10.1109/ICHIS.2005.100)
- Vapnik, V. (2000). *The nature of statistical learning theory*: springer.
- Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley and Sons, Inc., New York.
- Vapnik, V. N. (1995). *The nature of statistical learning theory*: Springer-Verlag New York, Inc.
- Walczak, S., & Cerpa, N. (1999). Heuristic principles for the design of artificial neural networks. *Information and Software Technology*, 41(2), 107-117. doi: [http://dx.doi.org/10.1016/S0950-5849\(98\)00116-5](http://dx.doi.org/10.1016/S0950-5849(98)00116-5)
- Wang, X. M. (2014). Advances and issues in flow cytometric detection of immunophenotypic changes and genomic rearrangements in acute pediatric leukemia. *Translational Pediatrics*, 3(2), 149-155.
- Wang, Z., Song, Q., Soh, Y. C., & Sim, K. (2013). An adaptive spatial information-theoretic fuzzy clustering algorithm for image segmentation. *Computer Vision and Image Understanding*, 117(10), 1412-1420. doi: <http://dx.doi.org/10.1016/j.cviu.2013.05.001>

- WebMD. (2008) Webster's New World Medical Dictionary. John Wiley & Sons.
- Weinkauff, R., Estey, E. H., Starostik, P., Hayes, K., Huh, Y. O., Hirsch-Ginsberg, C., . . . Freireich, E. J. (1999). Use of peripheral blood blasts vs bone marrow blasts for diagnosis of acute leukemia. *American journal of clinical pathology*, 111(6), 733-740.
- Weszka, J. S., Dyer, C. R., & Rosenfeld, A. (1976). A Comparative Study of Texture Measures for Terrain Classification. *IEEE Transactions on, SMC- Systems, Man and Cybernetics*, 6(4), 269-285. doi: 10.1109/tsmc.1976.5408777
- Wilkinson, M. H., & Schut, F. (1998). Digital image analysis of microbes: imaging, morphometry, fluorometry and motility techniques and applications: John Wiley & Sons.
- Wilson, L., & Moore, J. (2010). *Design Matters: Creating Powerful Imagery for Worship*: Abingdon Press.
- Wittekind, D. (1979). On the nature of Romanowsky dyes and the Romanowsky-Giemsa effect. *Clinical & Laboratory Haematology*, 1(4), 247-262.
- Won, C. S., Nam, J. Y., & Choe, Y. (2005, March 08, 2005). *Extraction of leukocyte in a cell image with touching red blood cells*. Paper presented at the Proc. SPIE 5672, Image Processing: Algorithms and Systems IV, 399. doi:10.1117/12.593335; <http://dx.doi.org/10.1117/12.593335>
- Wu, K., Gauthier, D., & Levine, M. D. (1995). Live cell image segmentation. *IEEE Transactions on Biomedical Engineering*, 42(1), 1-12.
- Wu., Q., Merchant., F., & Castleman., K. (2010). *Microscope Image Processing*: Academic Press.
- Yang, X., Tridandapani, S., Beitler, J. J., David, S. Y., Yoshida, E. J., Curran, W. J., & Liu, T. (2012). Ultrasound GLCM texture analysis of radiation-induced parotid-gland injury in head-and-neck cancer radiotherapy: an in vivo study of late toxicity. *Medical physics*, 39(9), 5732-5739.
- Yao, Y., Frasconi, P., & Pontil, M. (2001). *Fingerprint classification with combinations of support vector machines*. Paper presented at the Audio-and Video-Based Biometric Person Authentication.
- Yuan, B., & Liu, W. (2012). A Measure Oriented Training Scheme for Imbalanced Classification Problems. In L. Cao, J. Huang, J. Bailey, Y. Koh & J. Luo (Eds.), *New Frontiers in Applied Data Mining* (Vol. 7104, pp. 293-303): Springer Berlin Heidelberg.
- Zamani, F., & Safabakhsh, R. (2006). *An unsupervised GVF snake approach for white blood cell segmentation based on nucleus*. Paper presented at the Signal Processing, 2006 8th International Conference on.

- Zephyris, R. W. (2007). Microscope And Digital Camera. Retrieved 17/5/2008, from http://commons.wikimedia.org/wiki/File:Microscope_And_Digital_Camera.JPG
- Zhang, D., & Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37(1), 1-19. doi: <http://dx.doi.org/10.1016/j.patcog.2003.07.008>
- Zhang, X.-W., Song, J.-Q., Lyu, M. R., & Cai, S.-J. (2004). Extraction of karyocytes and their components from microscopic bone marrow images based on regional color features. *Pattern Recognition*, 37(2), 351-361. doi: <http://dx.doi.org/10.1016/j.patcog.2003.07.007>
- Zhang, H., Fritts, J. E., & Goldman, S. A. (2008). Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2), 260-280.
- Zhang, Y. J. (1996). A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8), 1335-1346.
- Zhen-zhen, X., & Su-yu, Z. (2012). *A Non-linear Approximation of the Sigmoid Function Based FPGA*. Paper presented at the Proceedings of the 2011, International Conference on Informatics, Cybernetics, and Computer Engineering (ICCE2011) November 19–20, 2011, Melbourne, Australia.
- Zhuang, L., & Dai, H. (2006). Parameter optimization of kernel-based one-class classifier on imbalance learning. *Journal of computers*, 1(7), 32-40.
- Zijenah, L. S., Kadzirange, G., Madzime, S., Borok, M., Mudiwa, C., Tobaiwa, O., . . . Katzenstein, D. A. (2006). Affordable flow cytometry for enumeration of absolute CD4+ T-lymphocytes to identify subtype C HIV-1 infected adults requiring antiretroviral therapy (ART) and monitoring response to ART in a resource-limited setting. *Journal of translational medicine*, 4(1), 33.

List of Publications and Papers Presented

- 1- Madhloom, H., Kareem, S., & Ariffin, H. (2012). An Image Processing Application for the Localization and Segmentation of Lymphoblast Cell Using Peripheral Blood Images. *Journal of Medical Systems*, 36(4), 2149-2158. doi: 10.1007/s10916-011-9679-0
- 2- Madhloom, H. T., Kareem, S. A., & Ariffin, H. (2012). *A Robust Feature Extraction and Selection Method for the Recognition of Lymphocytes versus Acute Lymphoblastic Leukemia*. Paper presented at the Advanced Computer Science Applications and Technologies (ACSAT), 2012 International Conference on. Kuala Lumpur, Malaysia. doi: [10.1109/ACSAT.2012.62](https://doi.org/10.1109/ACSAT.2012.62)
- 3- Madhloom, H. T., Kareem, S. A., Ariffin, H., Zaidan, A. A., Alanazi, H. O., & Zaidan, B. B. (2010). An Automated White Blood Cell Nucleus Localization and Segmentation using Image Arithmetic and Automatic Threshold. *Journal of Applied Sciences*, 10(11), 959-966.