AN INTELLIGENT DECISION SUPPORT SYSTEM FOR ACUTE LYMPHOBLASTIC LEUKAEMIA DETECTION

WORAWUT SRISUKKHAM

PhD

2017

AN INTELLIGENT DECISION SUPPORT SYSTEM FOR ACUTE LYMPHOBLASTIC LEUKAEMIA DETECTION

WORAWUT SRISUKKHAM

A thesis submitted in partial fulfilment
of the requirements of the
University of Northumbria at Newcastle
for the degree of
Doctor of Philosophy

Research undertaken in the Faculty of
Engineering and Environment

March 2017

Abstract

The morphological analysis of blood smear slides by haematologists or haematopathologists is one of the diagnostic procedures available to evaluate the presence of acute leukaemia. This operation is a complex and costly process, and often lacks standardized accuracy owing to a variety of factors, including insufficient expertise and operator fatigue.

This research proposes an intelligent decision support system for automatic detection of acute lymphoblastic leukaemia (ALL) using microscopic blood smear images to overcome the above barrier.

The work has four main key stages. (1) Firstly, a modified marker-controlled watershed algorithm integrated with the morphological operations is proposed for the segmentation of the membrane of the lymphocyte and lymphoblast cell images. The aim of this stage is to isolate a lymphocyte/lymphoblast cell membrane from touching and overlapping of red blood cells, platelets and artefacts of the microscopic peripheral blood smear sub-images. (2) Secondly, a novel clustering algorithm with stimulating discriminant measure (SDM) of both within- and between-cluster scatter variances is proposed to produce robust segmentation of the nucleus and cytoplasm of lymphocytic cell membranes. The SDM measures are used in conjunction with Genetic Algorithm for the clustering of nucleus, cytoplasm, and background regions. (3) Thirdly, a total of eighty features consisting of shape, texture, and colour information from the nucleus and cytoplasm of the identified lymphocyte/lymphoblast images are extracted. (4) Finally, the proposed feature optimisation algorithm, namely a variant of Bare-Bones Particle Swarm Optimisation (BBPSO), is presented to identify the most significant discriminative characteristics of the nucleus and cytoplasm segmented by the SDM-based clustering algorithm. The proposed BBPSO variant algorithm incorporates Cuckoo Search, Dragonfly Algorithm, BBPSO, and local and global random walk operations of uniform combination, and Lévy flights to diversify the search and mitigate the premature convergence problem of the conventional BBPSO. In addition, it also employs subswarm concepts, self-adaptive parameters, and convergence degree monitoring mechanisms to enable fast convergence. The optimal feature subsets identified by the proposed algorithm are subsequently used for ALL detection and classification. The proposed system achieves the highest classification accuracy of 96.04% and significantly outperforms related meta-heuristic search methods and related research for ALL detection.

List of Contents

List of Publications

Publications

Srisukkham, W., Zhang, L., Neoh, S. C., Todryk, S., & Lim, C. P. (2017). Intelligent leukaemia diagnosis with bare-bones PSO based feature optimization. *Applied Soft Computing*, *56*, 405-419. https://doi.org/10.1016/j.asoc.2017.03.024

Al-Mamun, M., Ravenhill, L., **Srisukkham, W.**, Hossain, A., Fall, C., Ellis, V., & Bass, R. (2016). Effects of Noninhibitory Serpin Maspin on the Actin Cytoskeleton: A Quantitative Image Modeling Approach. *Microscopy and Microanalysis : The Official Journal of Microscopy Society of America, Microbeam Analysis Society, Microscopical Society of Canada*, *22*(2), 394–409. http://doi.org/10.1017/S1431927616000520

Neoh, S. C., **Srisukkham, W.**, Zhang, L., Todryk, S., Greystoke, B., Peng Lim, C., Hossain, A., & Aslam, N. (2015). An Intelligent Decision Support System for Leukaemia Diagnosis using Microscopic Blood Images. *Scientific Reports*, *5*, 14938. http://doi.org/10.1038/srep14938

Al-mamun, M. A., **Srisukkham, W.**, Fall, C., Bass, R., & Hossain, A. (2014). A Cellular Automaton Model for Hypoxia Effects on Tumour Growth Dynamics. In *The 8th International Conference in Software, Knowledge, Information Management and Applications (SKIMA2014)* (pp. 1–8). http://doi.org/10.1109/SKIMA.2014.7083562

Lepcha, P., **Srisukkham, W.**, Zhang, L., & Hossain, A. (2014). Red Blood based Disease Screening using Marker Controlled Watershed Segmentation and Post-Processing. In *The 8th International Conference in Software, Knowledge, Information Management and Applications (SKIMA2014)* (pp. 1–7). Dhaka, Bangladesh. http://doi.org/10.1109/SKIMA.2014.7083556

Srisukkham, W., Lepcha, P., Hossain, A., Zhang, L., Jiang, R., & Lim, H. N. (2013). A mobile enable intelligent scheme to identify blood cancer for remote areas- cell membrane segmentation using marker controlled watershed segmentation phase. In *The 7th International Conference on Software, Knowledge, Intelligent Management and Applications (SKIMA2013)* (pp. 104–114). Chiang Mai, Thailand.

Poster

A Mobile Enable Intelligent Scheme to Identify Blood Cancer for Remote Areas- Cell Membrane Segmentation using Marker Controlled Watershed Segmentation Phase, Northumbria Research Conference, 2014.

List of Tables

List of Figures

List of Abbreviations

Abbreviations	Descriptions
ALL	Acute Lymphoblastic Leukaemia
AML	Acute Myeloid Leukaemia
BBPSO	Bare-Bones Particle Swarm Optimisation Algorithm
C	Cytoplasm
CIELAB	CIE L*a*b* colour space
CLL	Chronic Lymphoid Leukaemia
CML	Chronic Myeloid Leukaemia
CS	Cuckoo Search Algorithm
CT	Computed Tomography
DA	Dragonfly Algorithm
DNA	Deoxyribonucleic acid
DP	Decision Profile
DT	Decision Trees classifier
ELPSO	Enhanced Leader Particle Swarm Optimisation Algorithm
FAB	French-American-British cooperative group
FCM	Fuzzy-c-means
FCS	Fuzzy Compactness and Separation
FCS1	Fuzzy Compactness and Separation variant proposed by (Li, Kuo, & Lin, 2011)
FCS2	Fuzzy Compactness and Separation variant proposed by (Wu, Yu, & Yang, 2005)
FDR	Fisher’s discrimination ratio
FF-NN	Feed-forward Neural Networks classifier

GA	Genetic Algorithm
GLCM	Gray Level Co-occurrence Matrix
HD	Hausdorff Dimension
HIS	Hue-Saturation-Intensity colour space
HSV	Hue-Saturation-Values colour space
kNN	k-Nearest Neighbour classifier
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
MI	Mutual Information
MLP	Multi-Layers Perceptron classifier
mRMR	Minimum-redundancy-maximum-relevance
N	Nucleus
NWFE	Nonparametric weighted feature extraction
PCA	Principal component analysis
PSO	Particle Swarm Optimisation
RBCs	Red Blood Cells
RBF	Radial Basis Function Kernel
RBFN	Radial Basis Function Neural Networks classifier
RNA	Ribonucleic acid
SCM	Shadow C-mean
SDM	Stimulating Discriminant Measures clustering algorithm
SE	Structuring Element
SVM	Support Vector Machine classifier
WBCs	White Blood Cells
WHO	World Health Organization

Acknowledgements

First of all, I am extremely grateful to King Bhumibol Adulyadej, who provides great guidance (The Guidance of His Majesty The King) for the Thai people and myself. One of the great sayings that stays in my mind and reminds me during my PhD journey to achieve this goal is "When confronted with difficult tasks, remember never giveup (in Thai language: เมื่อเผชิญหน้ากับงานหนัก คิดเสมอว่า เป็นไปไม่ได้ที่จะล้มเหลว)". The body of His Majesty King Bhumibol Adulyadej, has passed away. His Loving Soul is still with the Thai people and myself forever.

I am very grateful and owe a huge debt to my principal supervisor, Assoc. Prof. Dr. Li Zhang. Without her excellent support, guidance, discussions, enthusiasm and patience, I would have never completed this thesis. She also inspired me to work hard and stretch beyond my limits.

I would like to extend my heartfelt gratitude to my second supervisor, Prof. Stephen Todryk, who introduced me to the world of biology, immunology and haematology. He really inspired me and opportunity to bring my supervision team and me to meet haematologists in The Royal Victoria Infirmary (RVI), Newcastle Upon Tyne, United Kingdom. Without his support, guidance, discussions and enthusiasm, this PhD research would not have been possible.

I would like to thank Assoc. Prof. Dr. Neoh Siew Chin (UCSI University, Malaysia) for her support, guidance and discussions during her days as a post-doctoral fellow at Northumbria University. She gave me the invaluable guidance ("To make my hand dirty, when I have to learn something new"). It is still in my mind forever.

I am exceedingly grateful to The Royal Thai Government Scholarship, The Ministry of Science and Technology of Thailand for granting me full financial support to complete this PhD thesis. I also very grateful to Office of Education Affair (OEAUK), The Royal Thai Embassy, London, United Kingdom and The Student Scholarship Division of The Ministry of Science and Technology, Thailand for their wonderful support to achieve my goal.

I would like to extend my sincere thanks to my workplace, The Department of Computer Science, Faculty of Science, Chiang Mai University, Thailand for the permission to study for my PhD abroad. Moreover, thanks should be extended to all of my colleagues in the Department of Computer Science, who kindly help me to look after my students and took over my workload during my PhD study at Northumbria University.

I am extremely indebted to Assoc. Prof. Dr. Fabio Scotti (Department of Computer Science, University of Milan, Italy), who provided the high quality microscopic blood smear images: ALL-IDB database for my PhD research. Also, thanks must go to Dr. Brigit Greystoke and Mr. John Lambert (The RVI Hospital, Newcastle Upon Tyne, United Kingdom) for their

helpful consultation and discussion regarding the clinical diagnosis of an acute lymphoblastic leukaemia and for providing annotations for all of the microscopic blood smear images in this project. Moreover, I would like to thank Clin. Prof. Niwes Nantachit, M.D., and Asst. Prof. Adisak Tantiworawit, M.D. (Maharaj Nakorn Chiang Mai Hospital, Faculty of Medicine, Chiang Mai University, Thailand) for their helpful, support and consultation regarding the clinical diagnosis of acute lymphoblastic leukaemia as well.

I would like to thank Prof. Alamgir Hossian (Anglia Ruskin University), who was my principal supervisor, for his helpful support and guidance me during his time at Northumbria University. He was the first person, who opened the door to my PhD journey.

I also would like to extend my sincere thanks to Assoc. Prof. Dr. Rosemary Bass, Dr. Gillian Brooks and Ms. Sasirassamee Bouvirat for their helpful support to my research.

I am very grateful to my previous teachers, lecturers and mentors, who taught and motivated me to achieve this thesis. I also would like to thank Dr. Mamun, Dr. Phoebe, Dr. Sirichai, Dr. Suwitcha, Dr. Tawunrat, Mr. Puttipong, Mr. Francesco, Ms. Pooja, my colleagues in the Computational Intelligence Research Group and all of my friends in our big home at the Pandon Building, F7 laboratory for their help in my PhD life in Newcastle Upon Tyne. I have enjoyed my life spent with you all. Also, I am grateful to Mr. Andrew Lawson and Dr. Kushwanth Koya for their help in proofreading my thesis.

Last but not least, I would like to express my love thank to my father, Mr. Chalad Srisukkham, my mother, Mrs. Sriwan Srisukkham, and my sister, Mrs. Wanwarang Kannakulsoontorn for their pure love, teaching and unconditional support to make me strong and to help friends as if they were my family members. Moreover, I also grateful to all of my lovely cousins in Thailand, who look after my parents during my study at Northumbria University. Especially, my thanks go to my grandmother, Mrs. Junsom Jantakarak, and my aunt, Mr. Jarun Jantakarak, who passed away before I completed my thesis. This thesis is for all of them.

Worawut Srisukkham

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the Faculty Ethics Committee on 17 June 2013.

I declare that the Word Count of this Thesis is 42,608 words

Name: Worawut Srisukkham

Signature: Worawut Srisukkham

Date: 23 March 2017

Chapter 1: Introduction

1.1 Background

Cancer is a disease in which malignant tumours (neoplasms), which are characterised by uncontrolled growth of abnormal cells, destroy and reduce healthy cells, block body functions, take nutrients away from the body tissues and spread to other parts of the body (metastasis) (Scott & Fong, 2014). Globally, cancer is the second leading cause of death (the first being heart disease) and 12.7 million people were diagnosed with cancer, causing an estimated 7.6 million deaths, in 2008 (Jemal, Bray, & Ferlay, 2011). In 2012 the World Health Organization (WHO) reported an estimated 14 million people with new cancer cases and 8.2 million deaths from cancer (American Cancer Society, 2015; Stewart & Wild, 2014).

According to American Cancer Society reports, in the United States, a total of around 1,685,210 new cancer cases and 595,690 cancer deaths are projected to occur in 2016 (Siegel, Miller, & Jemal, 2016). Cancer Research UK states that, in the UK, new cases of cancer were around 352,197 in 2013, with 161,823 deaths from cancer in 2012 (CancerResearchUK, 2016a). As an example of developing countries, the National Cancer Institute of Thailand reported that there were about 112,392 new cases of cancer and 63,272 deaths from cancer in 2012 (National Cancer Institute Thailand, 2015; Thairath, 2014). Additionally, the international agency for research on cancer, the GLOBOCAN project, estimates that, in India, about 1,000,000 new cancer cases and nearly 700,000 deaths occurred in 2012 (Mallath et al., 2014). The incidences and mortalities of cancers mentioned above reveal that cancer is a severe disease and that both the number of new cancer cases and deaths worldwide are still increasing.

In general, the diagnostic rate of cancer occurrence in developed countries is higher than in developing countries (Jemal et al., 2011). Scientific evidence reports that most of the cancers are caused by smoking, consuming excess alcohol over a long period of time, infectious agents and obesity (CancerResearchUK, 2016b). However, screening test programmes for early diagnosis of cancers can offer the opportunity of a complete cure or recovery at an early stage in a variety of cancers.

Leukaemia is a type of blood cancer, which affects the white blood cells (WBCs), an important part of the immune system that fights infection in human bodies. People with leukaemia produce abnormal and malignant WBCs that accumulate in the bone marrow and enter the blood stream. The malignant WBCs prevent the production of other important blood cells, i.e. mature WBCs, red blood cells (RBCs), and platelets (Campbell, 2011; Turgeon, 2012).

Leukaemia can affect people at any age. In December 2015, Cancer Research UK reported that "Leukaemia is the ninth most common cause of cancer death in the UK". In the year 2012, around 4,807 people died from leukaemia in the UK. Worldwide, more than 265,000 people died from leukaemia in 2012, with a variation in mortality rate across the world (CancerResearchUK, 2015a). The American Cancer Society estimates that, in 2016, there will be around 24,400 deaths from leukaemia in the US (Siegel et al., 2016). Additionally, in Thailand, new leukaemia cases were estimated at around 43,868 in 2012 (National Cancer Institute Thailand, 2015). Also, in India, about 32,632 new cases of leukaemia were predicted to occur in 2012 (Mallath et al., 2014). In particular, there are two types of acute leukaemia, acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). The severity of the acute leukaemia is that, without treatment, people diagnosed with acute leukaemia tend to die within a few weeks. The risk of a child under the age of 15 being diagnosed with acute lymphoblastic leukaemia by an incidence rate per 100,000 population in the United Kingdom during the year 2012 - 2014 is 22.1 or about 1 in 4,525 (CancerResearchUK, 2017). Over 80% of children survive for at least five years after receiving treatment and therapy following a diagnosis of the most common type of childhood leukaemia - ALL (CancerResearchUK, 2008). Moreover, Shah et al. (2008) showed that the proportion of children cured of leukaemia has increased dramatically because they received a diagnosis at an early stage and received appropriate treatment according to the state of the disease; however, the period of excess mortality associated with acute lymphoblastic leukaemia has increased because of late relapse, secondary malignancy and toxicity from treatment.

Screening or early stage diagnosis of cancer is a crucial process which can lead to a decreasing number in the risk of mortality or the development and spread of the disease into other parts in the body (Chamberlain & Moss, 1996). In December 2015, Cancer Research UK reported that, currently, they do not have a screening test which is reliable enough to test for ALL and AML. Thus, there is no UK screening programme for acute leukaemia cancer (CancerResearchUK, 2015b, 2015c). However, early diagnosis of acute leukaemia is essential for patient recovery and cure from the disease, especially for children (Putzu, Caocci, & Di Ruberto, 2014). Moreover, the screening or early stage detection of acute leukaemia can identify individuals suspected of having acute leukaemia and then carry out a further full investigation with specific accurate equipment so as to identify sub-types of acute leukaemia, after which appropriate treatment can be given to the patient. Therefore, the individual will have a high possibility of being cured.

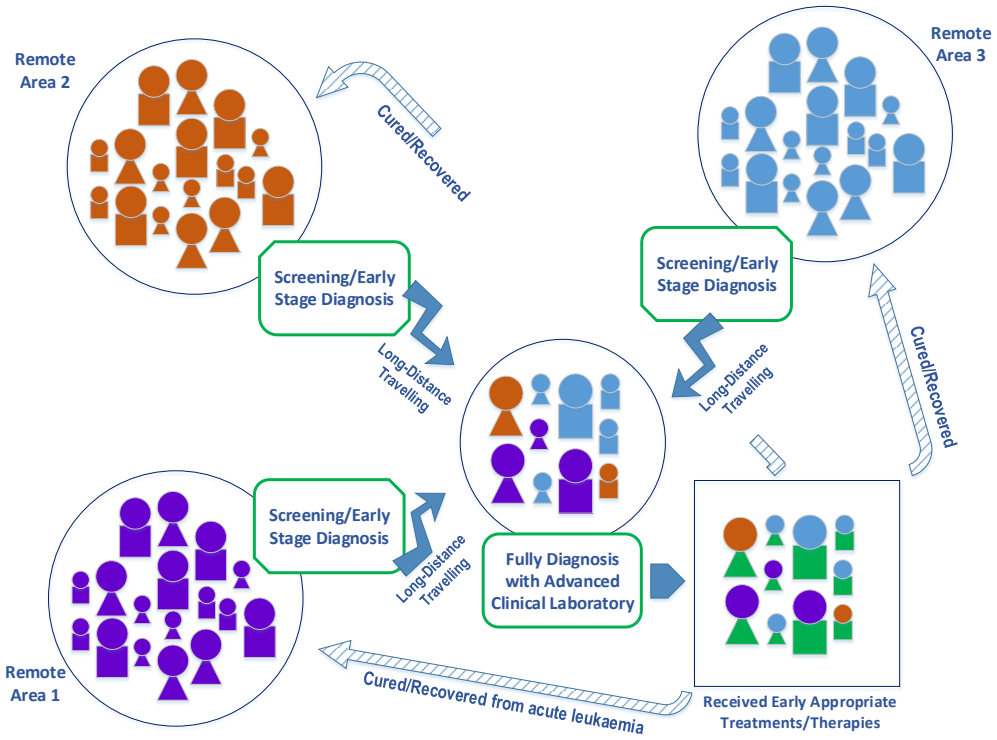


Figure 1.1 The screening or early stage diagnosis of acute leukaemia of individuals from remote areas in order to receive full diagnosis at an advanced clinical laboratory for accurate diagnosis and appropriate treatments and therapies

In fact, it may be very difficult or there may be major barriers for people, particularly in the resource-limited regions, to travel to main centres to receive a full investigation or specialised tests for an acute leukaemia diagnosis (Yeoh et al., 2013). As shown in Figure 1.1, a sustainable and possible approach that people or a community can adopt is go to the local health care service and get a screening or early stage diagnosis from the physicians. After the screening, if the result shows that the individual has suspected acute leukaemia, the individual then has the opportunity to go to a facilitated health care service for a full investigation by an advanced clinical laboratory to confirm whether they have acute leukaemia or not. Then, if the individual is diagnosed with acute leukaemia, a schedule for early appropriate treatments or therapies is applied. Therefore, the chance for an individual diagnosed with acute leukaemia to be cured or recovered is high. In particular, people in remote or rural areas that are resource-poor of health care facilities and expertise, can benefit from the early detection of acute leukaemia owing to the barriers to receiving good healthcare at a hospital in the main city, such as the long distances, limited transportation services and the expense of travelling.

With healthcare costs increasing throughout the world, there is a pressing need to reduce the cost and complexity of biomedical devices (Smith et al., 2011). Unfortunately, much of the high power light microscopy, especially fluorescent imaging and the opportunity for remote consultation and electronic record-keeping, remains inaccessible in rural and developing areas owing to the high price of medical equipment and training expense (Breslauer, Maamari, Switz, Lam, & Fletcher, 2009). The most reliable test for blood diagnosis in a good facilities hospital is the flow cytometer test. This test could make morphological analysis under microscopic examination obsolete, but the very high costs of the flow cytometer also means that morphological analysis is still required. Especially as, in developing countries or in remote and rural areas, most hospitals do not have flow cytometers (Escalante et al., 2012).

Therefore, a sustainable and cost-effective process of screening or early stage detection of acute leukaemia across the globe requires morphological analysis of blood test sample slides incorporated with an efficient digital diagnosis system by using the quantitative microscopic analysis techniques.

1.2 Research Problems and Motivation

Morphological analysis of blood smear slides by haematologists or haematopathologists is one of the diagnostic procedures available to evaluate the presence of acute leukaemia. This operation involves a lengthy processing time to produce the results and is a complex and costly process. Furthermore, the present results lack standardised accuracy owing to a variety of factors, including insufficient expertise or fatigue or imperfection of the samples (Mohapatra & Patra, 2010; Piuri & Scotti, 2004; Scotti, 2005, 2006). Additionally, the variability of report results by human manual diagnosis is possible due to the heterogeneous morphology of cells and poor-quality or dirty, stained blood smear slides (Turgeon, 2012). To limit manual operation problems, a digital diagnosis system is required to analyse microscopic blood smear images for disease detection and assist experts in the diagnosis of acute leukaemia.

In recent years, many researchers have developed digital diagnosis systems to analyse microscopic blood images for acute leukaemia detection (Buavirat & Srisa-an, 2008; Nasir, Mashor, & Hassan, 2013; Piuri & Scotti, 2004) and some researchers have developed automated systems for both blood count (Ongun & Halici, 2001; Sinha & Ramakrishnan, 2003) and acute leukaemia detection (Agaian, Madhukar, & Chronopoulos, 2014; Escalante et al., 2012; Khashman & Abbas, 2013; Madhukar, Agaian, & Chronopoulos, 2012; Mohapatra, Patra, & Satpathy, 2014; Putzu et al., 2014; Scotti, 2005). Further investigation is

still needed for robust and efficient acute leukaemia detection systems. However, many researchers have contributed various techniques in the segmentation of WBCs. The majority of segmentation techniques focus on the nucleus or nuclei of WBCs (Huang & Hung, 2012; Madhloom, Kareem, & Ariffin, 2012a; Meera & Mathew, 2014; Mohapatra & Patra, 2010; Nasir et al., 2013; Singhal & Singh, 2014) and use subsequent steps such as feature extraction and feature classification to identify acute leukaemia. Rarely have researchers contributed segmentation techniques for white blood cell membranes which have both nucleus and cytoplasm in each membrane (Mohapatra, Patra, Kumar, & Satpathi, 2012; Mohapatra et al., 2014; Mohapatra, Patra, & Satpathy, 2012; Putzu et al., 2014). This segmentation is a more difficult and complex process than segmentation of only the nucleus or nuclei from the WBCs. Also, accurate clinical diagnosis of acute leukaemia needs both parts of the WBC membrane to examine the abnormality of the blood smear samples.

Subsequently, in the analysis of each segmented WBC membrane image, the segmentation between nucleus and cytoplasm also needs a robust and reliable technique to separate them. This part is also a challenging task to separate the cell nucleus with regular or irregular shapes and with similar colours to the colours of cytoplasm from the cell cytoplasm of the WBC membrane. Some research applications provide good techniques and are able to achieve good reliable results for nucleus and cytoplasm separation (Jiang, Liao, & Dai, 2003; Mohapatra, Patra, Kumar, et al., 2012; Mohapatra et al., 2014; Mohapatra, Patra, & Satpathy, 2012). Hence, the robustness of existing works is compromised owing to the limitation of the separation algorithms (Kuo & Landgrebe, 2004; Li, Kuo, & Lin, 2011). Therefore, nucleus and cytoplasm separation still requires further investigation for a robust and reliable method.

Most of the acute leukaemia detection applications performed the feature extraction task to extract the discriminative characteristics from the segmented cell images. There are four common groups of features, including shape-based, texture-based, statistical-based and colour-based features. Many researchers have proposed a variety of feature sets that they have extracted and used in their work and then used the extracted features for the recognition of normal and abnormal acute leukaemia systems (Agaian et al., 2014; Madhukar et al., 2012; Piuri & Scotti, 2004; Putzu et al., 2014). Some researchers have used the process of feature selection, which is needed to reduce the redundancy of the non-significant features and increase the efficiency of the recognition system with the significant features (Escalante et al., 2012; Madhloom et al., 2012a; Mohapatra et al., 2014). The feature selection part is also important, and a more challenging task to select the significant discriminative characteristics from the raw feature subsets and then use the selected feature subsets to support the recognition for acute leukaemia detection with greater accuracy and robustness.

The challenges observed by the aforementioned researchers need to be explored. This research study focuses on acute lymphoblastic leukaemia, which is the most common in childhood, owing to the chance of children who have screening tests for early detection of this malignant blood disease, and to be cured, with high survival rates from the appropriate treatments, as previously reported. This has motivated us to develop an intelligent decision support system for acute lymphoblastic leukaemia detection using microscopic blood smear images.

This research is to develop a whole decision support system for early detection of acute lymphoblastic leukaemia disease. In terms of computerised diagnosis, it does quantitative morphological features on healthy and blast lymphocyte cells samples to differentiate among them. Therefore, the quantitative measurements of the lymphocytic cell samples can enable a robust and efficient early-computerised diagnosis of ALL. Overall, this research comprises of the following key stages for the robust automatic detection of ALL, i.e. (1) segmentation method for lymphocyte and lymphoblast cells membranes segmentation, (2) clustering algorithm based cell nucleus and cell cytoplasm separation, (3) feature extraction, (4) evolutionary algorithm based feature selection, and (5) ALL classification (Figure 3.1 of Chapter 3 for further details).

1.3 Research Aims and Objectives

In this research, the main aim is to develop an intelligent decision support system for acute lymphoblastic leukaemia detection using microscopic blood smear images. Moreover, we also aim at the robust and efficient computerised early stage diagnosis of ALL.

In order to achieve this goal, the following objectives are established:

i. Investigate the existing methods, tools, and techniques in microscopic blood images analysis for acute lymphoblastic leukaemia detection.
ii. Design an effective model of an intelligent decision support system for acute lymphoblastic leukaemia detection.
iii. Develop lymphocyte and lymphoblast cells membranes segmentation method for microscopic blood smear images.
iv. Devise improved clustering approach for cell nucleus and cell cytoplasm separation.
v. Extract shape-based, texture-based, colour-based features in microscopic blood smear images and utilise these feature sets to differentiate healthy (mature) and unhealthy (blast) lymphocyte cells images.

vi. Develop an evolutionary optimisation method for feature selection and robust ALL classification

vii. Test and validate the implemented experiments.

viii. Evaluate the experiments with the publicly available ALL-IDB2 database.

1.4 Research Contributions

The contributions to the knowledge of this PhD research include:

i. White blood cell membranes segmentation using a modified marker-controlled watershed method and morphological operations (Chapter 3)

   a. White blood cell membranes segmentation for microscopic blood smear sub-images, particularly lymphocyte (healthy lymphocyte cell) and lymphoblast (unhealthy lymphocyte cell) sub-images, using integration of the modified marker controlled watershed method and morphological operations, is presented. This method can segment and identify a WBC membrane from a noisy background sub-image, which is touching and overlapping with red blood cells, and retrieve the original RGB pixels colour of the identified cell membrane in the white background sub-image.

ii. The separation of nucleus and cytoplasm of the identified lymphocyte and lymphoblast cell membrane using a novel stimulating discriminant measures (SDM)-based clustering technique and the feature extraction from the separated nucleus and cytoplasm (Chapter 4)

   a. The novel clustering technique to separate nucleus and cytoplasm of lymphocytic (lymphocyte and lymphoblast) cell membrane images, namely SDM-based clustering, takes both within- and between-cluster scatter variants into consideration, and overcomes the limitation of the objective function of conventional Fuzzy C-mean (FCM) clustering, which focuses on only within-cluster scatter variance. It also outperforms other clustering methods, including Linear Discriminant Analysis (LDA) and Fuzzy Compactness and Separation (FCS) (Wu, Yu, & Yang, 2005) for robust identification of cell nucleus and cell cytoplasm. This clustering technique can also produce robust results of the separation nucleus and cytoplasm of the cell images.

b. A total of 80 features, which include shape-based features, texture-based Gray Level Co-occurrence Matrix (GLCM) features, colour-based CIELAB colour space features, and the statistical measurement of these feature sets are used to discriminate between healthy and unhealthy lymphocyte cells, as well as used for acute lymphoblastic leukaemia screening or an early detection system with image processing and artificial intelligent machine learning techniques.

c. Diverse single and ensemble classifiers are used in the experimental study for lymphocyte and lymphoblast detection. In this research study, Dempster-Shafer ensemble achieves the highest accuracy of 96.72% for bootstrap validation, whereas SVM with Gaussian Radial Basis Kernel (RBF) achieves an accuracy of 96.67% for 10-fold cross validation.

iii. The identification of the most significant discriminative characteristics of lymphocyte and lymphoblast cells to enable efficient ALL recognition using a proposed evolutionary Bare-Bones Particle Swarm Optimisation (BBPSO) variant algorithm (Chapter 5)

a. The proposed BBPSO variant algorithm incorporates two meta-heuristic search algorithms, i.e. cuckoo search (CS) and dragonfly algorithm (DA), and the following schemes such as convergence speed monitoring mechanisms, self-adaptive parameter setting and a subswarm concept to reduce premature convergence of the conventional BBPSO.

b. The proposed BBPSO variant algorithm combines the multiple search strategies, crossover and mutation techniques, and local and global random walk operations and enables them to work in a co-operative manner to balance between exploration and exploitation to overcome the local optima.

c. In comparison with advanced and classic nature-inspired and meta-heuristic algorithms, e.g. Enhanced Leader Particle Swarm Optimisation (ELPSO), PSO, BBPSO, Genetic Algorithm, CS and DA, the proposed BBPSO-based feature optimisation algorithm has efficient discriminative capabilities in which the significant discriminating feature subsets for lymphocytes and lymphoblasts are revealed.

1.5 Thesis Layout

The remainder of this research study is structured as follows:

Chapter 2: Literature Review. This chapter describes the extensive literature review of the biological background of leukaemia, acute lymphoblastic leukaemia, laboratory diagnosis of acute lymphoblastic leukaemia and the limitations of traditional methods. Additionally, an image analysis on blood smear samples using computer technology and image processing is provided to indicate the benefit of using a quantitative microscopic image analysis to reduce human operation error and assist the experts in diagnosis of the acute lymphoblastic leukaemia. Moreover, this chapter also explains the state-of-the-art of development for ALL detection, by organising the related literature review under five sequential processes, including image segmentation, image separation of nucleus and cytoplasm of the cell membrane images, feature extraction of cell nucleus and cell cytoplasm of the segmented cells, feature selection of the extracted descriptors to reduce the redundancy of the non-significant features, and acute lymphoblastic leukaemia identification. Finally, the scope of this research study is also provided.

Chapter 3: White Blood Cells Membrane Segmentation Using Marker-Controlled Watershed Method and Morphological Operations. This chapter introduces the segmentation of white blood cells (leukocytes), particularly lymphocyte and lymphoblast cells membrane using the integration of a modified marker-controlled watershed algorithm, and morphological operations. The microscopic sub-images from ALL-IDB2 database is applied in this research. The overall system architecture of this PhD research and details of materials used in the experiments and evaluations are explained. Subsequently, the identification of the WBCs membrane and retrieval of the segmented cells' membrane on the white background sub-image are described. Finally, the simulation and evaluation results are also revealed.

Chapter 4: The Separation of Nucleus and Cytoplasm Using Stimulating Discriminant Measures (SDM). In this chapter, we present the separation of cell nucleus and cell cytoplasm of the identified WBCs membrane, in particular the lymphocytic (lymphocyte and lymphoblast) cell membranes, using the novel SDM clustering technique. Additionally, to overcome the limitation of the conventional FCM algorithm, the motivation of the proposed SDM algorithm is presented. Moreover, this chapter reveals the simulation and evaluation results of the SDM clustering compared with the state-of-the-art clustering techniques using the identified 180 lymphocytic cell membrane images. The details of lymphocytic cell features chosen from the literature review and

consultation with haematologists are also described. The extraction of 80 features from nucleus and cytoplasm of the identified lymphocytic images is described. Finally, the classification and evaluation of normal/lymphocyte and abnormal/lymphoblast cases for the ALL detection using the extracted features with single and ensemble classifiers are presented.

Chapter 5: The Proposed BBPSO Variant for Feature Optimisation. This chapter introduces the proposed feature subsets selection method using an optimisation algorithm, namely the BBPSO variant algorithm, to select the significant discriminative characteristics of the nucleus and cytoplasm of the identified lymphocytic cell images. The proposed BBPSO variant algorithm with the two objective functions for fitness evaluation is introduced. Moreover, the details of the proposed different search strategies which are combined in the BBPSO variant algorithm are explained. This chapter also reveals the simulation and evaluation results of the proposed BBPSO variant algorithm compared to the state-of-the-art nature-inspired and meta-heuristic algorithms reported in the literature. Finally, the recognition results of normal/lymphocyte and abnormal/lymphoblast for the ALL detection using the selected feature subsets from the BBPSO variant algorithm and other baseline meta-heuristic algorithms are compared and presented in this chapter.

Chapter 6: Conclusion and Future Work. This chapter provides the concluding remarks of this research study. The contribution to knowledge of this research in the field of acute lymphoblastic leukaemia detection is presented. Finally, this chapter concludes with future directions which contain recommendations to overcome potential deficiencies of this research and the other application of this work.

Chapter 2: Literature Review

2.1 Introduction

This chapter presents previous studies that have been conducted in multidisciplinary areas, including biomedical engineering, haematology and computer science. The structure of this chapter is as follows: Section 2.2 describes the biological background of human blood, blood-related diseases in humans (such as acute lymphoblastic leukaemia, i.e. the main target disease of this research study), laboratory diagnosis of acute lymphoblastic leukaemia and the limitations of the traditional methods. Section 2.3 explains the image analysis on blood smear samples using computerised technology and image processing techniques. Moreover, state-of-the-art related work regarding the development for acute leukaemia detection is also discussed as the foundation for this thesis. Finally, the scope of this PhD research is presented.

2.2 Biology of Leukaemia and Computer-Aided Diagnosis for Blood Smear Samples

2.2.1. Human Blood

In the human body, on average, only five of seventy litres of human body fluid are blood (Uthman, 2016). It is the fluid, which flows through the heart, blood vessels and tissues. It conveys oxygen and nutrients to the tissues and unwanted products to the lungs, liver and kidneys, where they can be removed from the body. Blood cells are composed of various types of cells suspended in plasma, which is a transparent and pale yellow coloured fluid (Bain, 2004). The blood cells are mainly of three types, as shown in Figure 2.1, including Red Blood Cells (RBCs), or Erythrocytes, which transport oxygen from the lungs to the tissues of the body: White Blood Cells (WBCs), or Leukocytes, which are responsible for protecting the body or producing antibodies against infections, e.g. viruses, bacteria and fungi, and destroying parasites; and Platelets, or Thrombocytes, which are important in the clotting of blood and prevent blood loss at locations of injury (Bain, 2004; Hough, 2015a; Moor, Gary, Blann & Knight, 2010). Comparing by respective size, the biggest blood cell type is WBC, then RBC and the smallest blood cell type is a platelet.

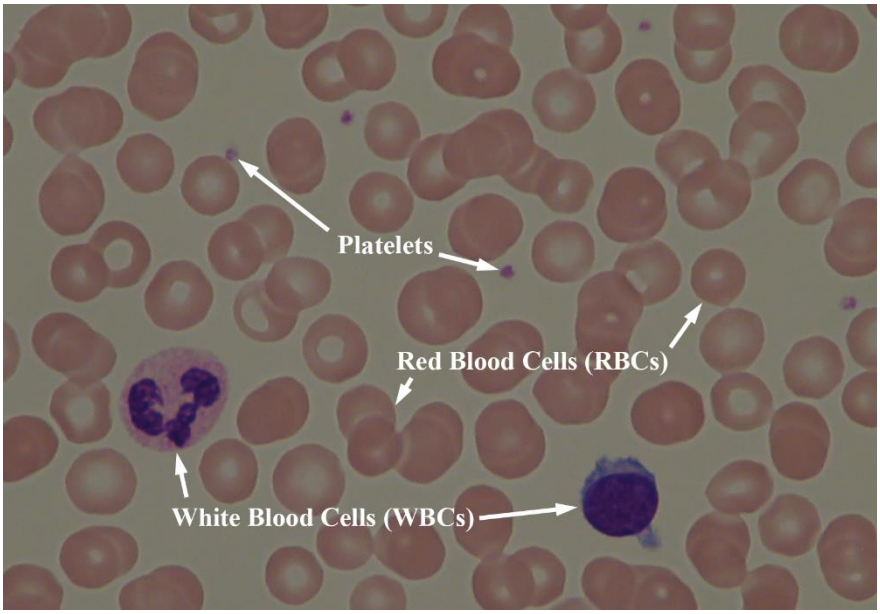


Figure 2.1. The three types of blood cells including RBCs, WBCs and Platelets (Labati, Piuri, & Scotti, 2011a)

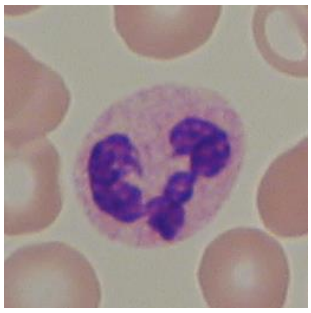
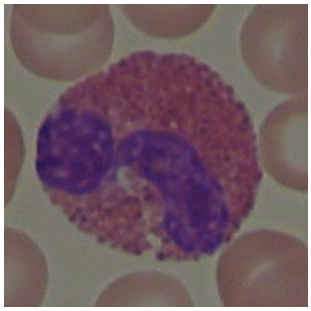
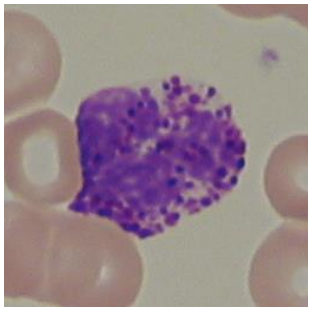
All types of blood cell are generated from stem cells in the bone marrow, which is a sponge-like tissue in the middle of bones. The first stage of the blood cells' development is from stem cells. They have several development stages to form each type of blood cell and then they enter into the peripheral blood stream which circulates in the body (Hough, 2015a).

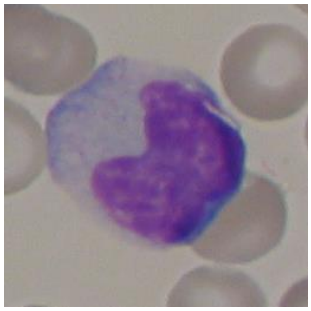
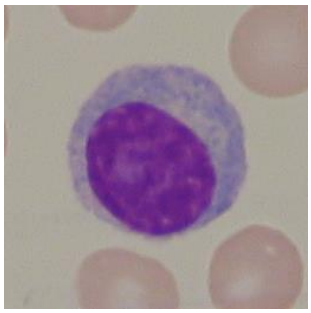
Leukocytes, or WBCs, work like soldiers in the body. They are responsible for defending the human body from attack by microorganisms, e.g. bacteria, viruses and parasites. When the body is attacked by microbes, the number of WBCs increase dramatically to destroy and absorb the attackers. Increased number of WBCs may be present in a number of conditions, such as after surgery, during fever and in cancer (Gary et al., 2010).The peripheral blood stream contains healthy or mature RBCs, platelets and WBCs.

In particular, healthy WBCs have nuclei and cytoplasm. Each WBC membrane contains a nucleus and cytoplasm. WBCs can be classified by the size and shape of nucleus, and by the presence and absence of granules in the cytoplasm, giving them the names of polymorphonuclear leukocytes (granulocytes) and mononuclear leukocytes (agranulocytes), respectively. The nucleus contains chromatin, which is the material of the chromosomes of organisms consisting of protein, ribonucleic acid (RNA) and deoxyribonucleic acid (DNA). The chromatin is one of the characteristics of WBC used to differentiate between healthy (mature) and unhealthy (immature) cells. Generally, WBCs can be categorised in five different types: Neutrophil, Eosinophil, Basophil, Monocyte and Lymphocyte. Their functional

characteristics, morphological features and percentage of WBC are depicted as a sample in Table 2.1 (Gary et al., 2010; Turgeon, 2012).

Table 2.1. The five different types of white blood cells and their characteristics

The type of WBCs	**Functional characteristics, morphological features, and percentages in human body**
 Neutrophil	This type has the functional characteristic of participation in inflammation, absorbing bacteria and yeast into the cell body, searching for usable nutrients and removing discarded waste. It generally has nucleus with two to five lobes, and light purple granules in the cytoplasm when stained with Wright-Giemsa. This type represents the majority of white blood cells in range of forty to seventy per cent.
 Eosinophil	Eosinophil's function is to defend against infections of parasites, participate in allergic reactions, and the release of histamine to the body. It has two large lobes of nucleus and with coarse red-brown granules in the cytoplasm when stained with Wright-Giemsa. The percentage of this type is between one and five.
 Basophil	Basophil take part in hypersensitivity responses and release of histamine and heparin. The nucleus of this type has two lobes and, specifically, the granules of cytoplasm are dark purple or black coloured, which often cover and make the nucleus not clear for visualisation. This type is less than one per cent in a healthy body.

 Monocyte	This cell type has functionality similar to Neutrophil. In addition, it acts to release cytokines, participate in haemostasis and cooperates with Lymphocyte cells in producing antibodies. The Monocyte is the largest of peripheral blood WBCs. The nucleus of Monocyte is not balanced in shape (irregular) and occupies areas around seventy to eighty per cent of the cell. Moreover, it has no granules in cytoplasm, but, in rare cases, may have granules occurring in cytoplasm. The percentage of Monocyte is in the range of two to seven.
 Lymphocyte	Lymphocyte's function is to collaborate in and generate antibodies, which are very important in the human immune system and destroy and absorb the cells infected with viruses. This cell nucleus has a round and balanced shape (regular) and resides in areas around ninety-five per cent of the cell. In addition, the cytoplasm is blue-grey coloured when stained with Wright-Giemsa and shows no presence of granules. Lymphocyte is often smaller than Neutrophil and slightly larger than RBCs. This cell type has between twenty and forty per cent in humans.

The normal healthy or mature leukocytes play a crucial role in defending against infections by microbes and maintain the immune system in the human body. In addition, the immature leukocytes also occur in bone marrow as well as contributing to the generation of the blood cells. However, if high numbers of immature WBCs escape from bone marrow and enter the bloodstream, it may indicate something anomalous occurring in the body, such as blood disorder or blood disease. The next section discusses blood diseases in humans.

2.2.2. Blood Disease in Humans

Blood disease is a disease or disorder of the blood. Haematology is the study of the development and diseases of blood (Turgeon, 2012). A medical doctor who has expertise in the diagnosis of blood, particularly the disorders of blood, and treats patients with blood conditions, is known as a haematologist (Bain, 2004). Neoplastic disorders of the blood mean there is an uncontrolled growth of abnormal blood cells, which can be classified as benign or malignant disorders. Benign blood disorders mean that the blood disease can resolve

completely with treatment and does not affect patients' whole life. Examples of benign disorders are Anaemia, which is insufficient numbers of RBCs and amount of haemoglobin in the blood, and Thrombocytopenia, which is low numbers of platelet count. On the other hand, malignant blood disorders mean that the blood disease can cause severe problems and threaten a patient's life. Examples of malignant disorders are Leukaemia, which is the number of immature WBCs greatly increasing abnormally in the bloodstream; and Multiple Myeloma, which is the abnormal multiplying of plasma cells in the bone marrow (Cleveland Clinic, 2016; Scott & Fong, 2014; St Luke's Cancer Center, 2016). Additionally, malignant disorders of WBCs are termed blood cancers. Leukaemia is a blood cancer and is the main focus for this research study and details of which are presented in the next section.

2.2.3. Leukaemia, Clinical Signs and Symptoms

Leukaemia is a blood cancer, whereby high numbers of abnormal white blood cells are formed in the bone marrow and hence in the blood. It causes the uncontrolled production (overproduction) of immature and mature white blood cells in bone marrow. When vast numbers of immature white blood cells are increased in bone marrow, they hinder other mature cells by replacing red blood cells and platelets (Leonard, 1993; Scott & Fong, 2014). An inadequate number of RBCs, platelets and mature WBCs can affect patients with symptoms of leukaemia, including anaemia, frequent fever, fatigue, night sweats, and easy bruising and bleeding. Leukaemia is classified based on the haematopoietic cell of origin (line production of blood cells) as lymphoid (lymphocytic) and myeloid (myelocytic). Moreover, based on clinical symptoms, leukaemia is classified as either 'acute', which is a rapidly progressing disease with high numbers of immature (blastic or blast) WBCs and can threaten patient death within a few months if treatment is not received; or 'chronic', which represents a slowly progressing disease with increased numbers of more mature WBCs. Chronic leukaemia can affect patients over a longer period than acute leukaemia and may not cause the patient's death (American Cancer Society, 2016b; Leonard, 1993; Scott & Fong, 2014; Turgeon, 2012). Generally, there are four types of leukaemia, as below:

- Acute Lymphoblastic Leukaemia (ALL), which is the most common type of childhood blood cancer and can also affect older people. The survival rates of ALL in children over the past 50 years have increased dramatically and around ninety per cent of children with ALL have successful treatment (Abbott, 2008; American Cancer Society, 2016a; Campbell, 2011; Hough, 2015a).
- Acute Myeloid Leukaemia (AML), which denotes the common type of blood cancer in elderly people and is rare in children. In the last two or three years, better treatment

can increase the survival rates of AML patients. However, the types of treatment will be different depending on individual investigation (Campbell, 2011; Hough, 2015b).

- Chronic Lymphocytic Leukaemia (CLL), which represents the common type of blood malignancy in people over 70 years old, but does not occur in children (Abbott, 2008; Agrawal & Deardean, 2014; American Cancer Society, 2016a).
- Chronic Myeloid Leukaemia (CML), which is the common type of blood cancer in elderly people averaging around 60 years old rather than in children (American Cancer Society, 2016a; Apperley, 2015).

In this research, we focus on acute leukaemia, specifically ALL, which is the most common in childhood, because of the opportunity of people having screening tests for early detection of this malignant blood disease, and to be cured, with high survival rates from the appropriate treatments, as reported in many public health organisations (Campbell, 2011; Hough, 2015a). More information about ALL and its characteristics are presented in the next section.

2.2.4. Acute Lymphoblastic Leukaemia, Its Characteristics and Its Classification Systems

ALL is a malignant blood disease in which the immature lymphocyte white blood cells form in the bone marrow; the source of blood cells. The immature lymphocytes, known as lymphoblasts, are over produced in bone marrow and replace the mature RBCs and platelets. Moreover, they move out from the bone marrow and then spread into the peripheral blood stream with an immaturity to defend against any infections. This causes human body weakness with common symptoms such as fatigue, weakness, fever, bone or joint pain, and easy bleeding and bruising. Body weakness occurs due to insufficient RBCs to carry oxygen and nutrients to the body tissues and the low number of platelets for blood clotting can cause death within a few months in ALL patients if untreated (American Cancer Society, 2016b; Leonard, 1993; Scott & Fong, 2014).

The characteristics of ALL in biological and clinical knowledge can be used to differentiate the maturity (immature and mature) of lymphocyte cells, which are both produced from the same lymphoid pathway. The predominant characteristics of ALL are used in supporting the diagnosis process and referring to the treatment solutions according to the diagnostic results. The next paragraph shows the two common classification systems for ALL, which are widely used in laboratories and hospitals, including the French-American-British (FAB) cooperative group and the World Health Organization (WHO) classifications. The experienced experts, such as haematologists, oncologists, haematopathologists and specially trained general pathologists, use the cell's origin to identify and differentiate ALL from other leukocyte cells.

The traditional classification of ALL cells, by their morphological characteristics and cytochemical studies, was first proposed by the French, American and British (FAB) cooperative group in 1976 (Bennett et al., 1976). The FAB system provides clear guidelines for uniformity and consistency of diagnosis for haematologists, oncologists and haematopathologists worldwide for classification of ALL (Abdul-Hamid, 2011; Bain, 2010). The criteria based on the FAB system to classify ALL, particularly the blast or lymphoblast cells, includes three subtypes (L1 to L3), which are depicted in Table 2.2 (Abdul-Hamid, 2011; Rodak & Carr, 2012).

Table 2.2. FAB morphological classification of ALL

FAB Morphological Classification	
Subtype	**Criteria or features of lymphoblast cells**
L1	Small cells predominant, nuclear shape is regular; round with rare cleft. Nuclear contents, such as chromatin, nuclear shape and nucleoli, rarely occur. Cytoplasm is small to moderately pale with blue colour (basophilia), when stained with Wright or Wright-Giemsa staining technique.
L2	Large and heterogeneous cells with an irregular nuclear shape, usually found cleft in the nucleus. One or more large nucleoli are visible. Cytoplasm area varies in colour and the nuclear membrane is also irregular shape.
L3	Cells are large and homogeneous in size. Nuclear shape is round or oval. In the nucleus area, there are one to three nucleoli and sometimes up to five. Cytoplasm is deep blue (basophilia) with vacuoles often clearly visible. The deep-blue colour in cytoplasm is visible in every cell, with bubbles or vacuoles occurring inside.

In addition, all morphological criteria from the three subtypes can be used to distinguish ALL or lymphoblast cells from mature lymphocyte cells. In other words, we can use the integration of the three subtypes criteria based on the FAB system to identify unhealthy lymphocyte or lymphoblast cells.

The second standard for classification of ALL was proposed by WHO in 1997 (Harris et al., 2000), which added another criteria rather than just morphology to evaluate the lymphoblast cells. The WHO standard needs more information of the blast cells from other equipment, including flow cytometric immunophenotyping, cytogenetics and molecular analysis, to provide more accurate classification of ALL and refers to the specific treatments and therapies of ALL patients with high survival rates (Albitar, Giles, & Kantarjian, 2008; Bain, 2010; Craig & Foon, 2008). The performance of the flow cytometry machine can provide the specific type of leukaemia and its subtypes with high accuracy (Craig & Foon, 2008). In addition, the flow cytometer test produces reliable results and can replace the morphological test process. However, the high cost of the machine is the main obstacle to hospitals and the medical facilities in developing countries and resource-poor regions do not have them. Therefore, the morphology and cytochemistry of the blood cell samples are still needed to analyse the condition of ALL disease (Escalante et al., 2012).

Therefore, this research study used the criteria based on the FAB standard by integrating the morphological criteria of the three subtypes as the conditions to identify the ALL blast cells and differentiate between the healthy (lymphocyte) and unhealthy (lymphoblast) cells for early detection, early diagnosis or screening tests for ALL in resource-poor medical facility regions and developing countries.

2.2.5. Diagnosis of Acute Lymphoblastic Leukaemia

Diagnosis of ALL is the process of identifying or determining the characteristics and causes of the disease based on information from a clinical history of the suspected patients and their family, a physical examination, and clinical laboratory studies. Sufficient information from all sources of examination is essential to interpret the correct outcome of diagnosis and lead to the appropriate treatment programme for the suspected patients. It usually begins with clinical suspicion (Bain, 2010). In addition, the personal clinical characteristics of suspected patients lead to further investigation and include the presence of symptoms, e.g. fever, bone pain, and fatigue, as well as family history, medical history and social history. A suspicion of ALL leads the physician to find more incidences by a physical examination, such as unnatural lack of colour in the skin (pallor), enlarged liver, palpable spleen and bruising. Furthermore, if incidences of suspected patients from their clinical history and physical examination reveal a relation to abnormality, a laboratory examination will take place by performing blood count and blood film of bone marrow aspiration (Bain, 2010). Moreover, a light microscope is then usually used for the process of ALL diagnosis by examination of the peripheral blood smear samples and bone marrow aspiration smears. In modern laboratories, the diagnosis of ALL

with accurate results is based on the morphology, cytochemistry, immunophenotyping, cytogenetic and molecular nature of the suspected patients' blood and bone marrow samples (Albitar et al., 2008; Bain, 2010; Kebriaei, Anastasi, & Larson, 2002). However, immunophenotypic and genetic analysis, which uses specific machines such as flow cytometry, and flow cytogenetics, are high cost (Andrews, Holm, & Myers, 2005; Logan et al., 2010; Muslimani et al., 2010) and not available for all medical facilities in developing countries and resource-poor regions. Therefore, the need for microscopic examination of stained blood smear and bone marrow slides still remains (Zini et al., 2010) and is used as a standard method for ALL diagnosis with the FAB classification across the third world and resource-poor countries (Bain, 2010). Moreover, the preliminary screening and early detection of ALL patients using a human microscopic examination of blood slide samples are the most applicable and suitable ways for those regions to screen and then refer suspected patients with ALL to receive a full investigation in better medical facilities, with the outcome of a suitable treatment plan to cure them of the disease. The importance of preliminary screening and early diagnosis is explained in the next section.

A. Why is Screening or Early Diagnosis Important?

There are two terms related to early diagnosis of disease: screening or early detection; and diagnosis. The former means the process to detect disease in the preclinical phase, when the individual or suspected patient has the disease, but doesn't know it, or during the presence of disease before the occurrence of clinical symptoms. In addition, early detection may allow the beginning of therapy before the disease increases to an invasive level and becomes uncontrolled, which can compromise the effectiveness of chemotherapy (El Rassi, Little, Holloway, Roberts, & Khoury, 2012). Whereas, the latter is also the process of checking suspected patients who have existing symptoms or show positive results in screening tests, with medical and laboratory examination of patients who have specific indication of the disease, to determine and confirm whether or not they have the disease and then provide them with an appropriate treatment plan for the next stage to cure them of the disease (Lewis, Sheringham, Lopez Bernal, & Crayford, 2014).

For ALL, a report by the American Cancer Society (American Cancer Society, 2016a, 2016b) indicated that nowadays there are no special tests recommended for detecting the disease in the early stage. Moreover, it suggested that the best way to find the disease early is to report any presence of signs and physical symptoms of the disease to the physician as quickly as possible.

In this research study, we define the meaning of early diagnosis as the process of determining suspected patients who have the presence of clinical symptoms related to the physical symptoms of ALL and then go through the process of clinical examination, such as complete blood count and microscopic examination of blood or bone marrow slide samples, to indicate whether the suspected patients have the disease or not in the earliest possible time. Therefore further investigation with advanced techniques and modern equipment is needed to confirm the disease and design a special treatment plan for each of the ALL patients.

Therefore, the importance of early diagnosis is that it can detect the presence of ALL in suspected patients at an early stage, so that the disease can be managed with appropriate treatments before the serious symptoms occur. Furthermore, with the realistic clinical examination methods, e.g. microscopic blood or bone marrow samples examination, in the medical facilities of resource-poor countries, it can support early diagnosis of ALL patients and refer them to receive better medical facilities with the right treatments and therapies.

B. Peripheral Blood Smear

A peripheral blood or blood film smear is a thin layer of blood smeared or spread on a glass microscope slide. In fact, blood cells are transparent and cannot be recognised when they are examined under a microscope. Therefore, a peripheral blood smear is stained with a mixture of several methods or dyes (Bain, 2004). The purpose of staining the blood smear is to identify blood cells and recognise the morphology of the individual cells under microscopic examination (Bain, 2004; Rodak & Carr, 2012). Moreover, a properly prepared blood smear is crucial to the accurate examination of blood cell morphology using a microscope. The staining technique enhances the colour of RBCs, WBCs and platelets with a variety of colours, which enable the detailed structure of the cells to be recognised (Bain, 2004). The Wright or Wright-Giemsa stain is the most widely used staining method for peripheral blood and bone marrow smears. It contains both eosin and methylene blue and is, therefore, termed a polychrome stain. The colours of stained peripheral blood smears vary slightly from laboratory to laboratory, depending on the method of staining the smears (Rodak & Carr, 2012). The characteristics of properly stained blood smear are as follows (Rodak & Carr, 2012; Turgeon, 2012):

- The colour of RBCs should be pink to salmon.
- The colour of nucleus or nuclei is deep blue to purple.
- The cytoplasm of mature lymphocyte should be blue to grey colour and that of immature lymphocyte or blast is deep blue or purple, termed basophilic colour.
- The colour of cytoplasmic granules of neutrophil is lavender to lilac.

- The cytoplasmic granules of basophils are deep blue to black colour.
- The colour of cytoplasmic granules of eosinophils is red to orange.
- The area between the cells should be transparent, clean, and free of artefact stain.

A properly prepared peripheral blood or bone marrow smear is crucial to the accurate diagnosis of ALL under the light microscope. Thus, the next section presents the laboratory diagnosis of ALL using microscopic examination.

2.2.6. Laboratory Diagnosis Using Microscope and Classification of ALL

A peripheral blood smear is the most helpful addition to the history and physical findings in diagnosis of paediatric haematological disorders (Abbott, 2008). The microscopic examination of stained peripheral blood or bone marrow smear slides remains a standard method for ALL diagnosis, particularly for resource-poor countries and regions around the world (Bain, 2010; Zini et al., 2010). The counting of blast cells leads to diagnosis of ALL. Patients with leukaemia present with decreased RBCs and elevated WBCs count in 60% to 70% of cases (Turgeon, 2012). The presence of more than 30% blasts count in a peripheral blood smear should be considered as acute leukaemia (Cason et al., 1989; Kebriaei et al., 2002). Furthermore, in WHO classification criterion, the occurrence of more than 20% blasts in the bone marrow or peripheral blood smear is also crucial for the indication of acute leukaemia (Kebriaei et al., 2002). Microscopic diagnosis of ALL disease is done by counting the blast cells, which distinguish between healthy or mature lymphocyte cells based on the morphology of both nucleus and cytoplasm of cells on the blood smear slides. A light microscope can assist the haematologists and hematopathologists to examine and differentiate the morphology of the immature lymphocyte or lymphoblastic cells and the mature lymphocyte cells corresponding to the standard FAB classification.

However, the standard WHO classification is proposed and produces more accurate classification of results compared with the FAB system, owing to it requiring more information, including immunophenotyping, cytogenetic and molecular analysis (Abdul-Hamid, 2011; Albitar et al., 2008; Kebriaei et al., 2002), which are supplemented by flow cytometry and flow cytogenetics, in the evaluation of the lymphoblastic cells to confirm the subtype of the blast cells and then transfer to the appropriate treatment plans. However, the very high costs of advanced equipment, as mentioned above, is an obstacle for medical healthcare in developing or resource-poor countries. Therefore, the microscopic examination

of stained peripheral blood or bone marrow smear slides is still a common method for screening and identifying of ALL.

2.2.7. Limitations of Diagnosis of Blood Diseases with The Traditional Method

The microscopic examination of stained peripheral blood smear slides enables experts, such as haematologists and haematopathologists, to investigate the characteristics of healthy and unhealthy lymphocyte cells for the diagnosis of ALL. It also provides both manifestations and presents visual images of morphological components of blood cells under microscopic examination. Moreover, it can assist the experts in the diagnosis process by magnifying the morphological and textural content of lymphocyte or lymphoblast cells' components, including nucleus and cytoplasm regions, and then interpreting them to indicate the condition of the patients. Therefore, the aforementioned are the benefits of using microscope examination of stained blood slides.

An examination of the peripheral blood smears using a light microscope requires skilled and experienced haematologists or haematopathologists. The experts produce and interpret the examination results based on their clinical experiences by distinguishing between the healthy (mature) and unhealthy (immature) lymphocyte cells corresponding to the FAB standard. However, a variety of reports by human manual diagnosis may occur (Argyle, Benjamin, Lampkin, & Hammond, 1989; Elsheikh et al., 2008) in all types of haematological disorders including ALL cancer.

The causes of the variability of reports of manual diagnosis may include the heterogeneous morphology of cells, the poor-quality or dirty stained blood smear slides (Turgeon, 2012), the inconsistency of results, whereby the same slide samples are examined by the same experts more than once, known as intraobserver variability, and the various diagnosis outcomes, (whereby more than one expert determines the same stained slides) known as interobserver variability. According to Browman et al. (1986), who are pioneers regarding observer discrepancies in light microscopic examination based on manual diagnosis of ALL, the outcomes in evaluating intraobserver and interobserver concordance between two experts were 64.8% and 70.5% for the former and 63% and 72% for the latter studies, respectively.

To reduce the human errors of the aforementioned, quantitative microscopy techniques have been developed for microscopic examination in the process of haematology disorder diagnosis (Das, Chakraborty, Mitra, Maiti, & Ray, 2013). These use computerised technologies to assist human experts in the diagnosis of blood cancer and to reduce human intervention during the

diagnosis process. Moreover, it can repeat the investigation process with consistently reported results.

This becomes challenging for researchers to investigate the novel quantitative approaches in the field of biomedical engineering, haematology and computer science in order to overcome the difficulty of differentiating between mature lymphocyte and lymphoblast cells.

The image processing technique is a widely used computer algorithm which performs the analysis of digital images, particularly medical digital images. It can enable researchers to develop crucial quantitative methods, which aim at early detection and accurate diagnosis of ALL by integrating image processing techniques with the clinical procedures to analyse the disease. The next section illustrates the quantitative analysis of microscopic blood smear images for acute leukaemia diagnosis.

2.3. Image Analysis on Blood Smear Samples Using Computerised Technology and Image Processing Techniques

As previously stated, the aim of staining blood smears is to identify blood cells and recognise the morphology of the individual cells under a light microscope (Bain, 2004; Rodak & Carr, 2012). Moreover, the variety of colours of the cells that are enhanced by the chemical staining techniques, e.g. Wright-Gemsia, Leishman, enables the detailed structure of the blood cells to be recognised by haematologists or haematopathologists (Rodak & Carr, 2012).

The digital images of blood cells, which are taken by a microscope connected to a digital camera or a digital microscope, are valuable for the interpretation and evaluation of the experts, more than once, to confirm the examination and diagnosis results. It is also the initial point of the computerised image analysis of microscopic blood smear images.

The history of medical imaging started in 1895 when German physicist Wilhelm Roentgen invented the X-ray to image the bony interior of his wife's right hand, which was captured on an X-ray film (Okada & Blankstein, 2009). In the early 1970s, the first X-ray computed tomography (CT) scanner, known as a CT scanner, was developed and used a computer to record the scanned data (Dougherty, 2009). Since then, computers have become an important element of many medical imaging modalities, including ultrasound, CT, radionuclide imaging and magnetic resonance imaging (Okada & Blankstein, 2009). Nevertheless, without using X-rays as the main device for medical imaging, in the late 1950s, a microscopic image was first applied using computing technology in an attempt to automate screening for gynaecological cancer (Cooper et al., 2012). Also, in the early 1950s, the first evidence of using a computer in a medical laboratory for processing data was published (Park et al., 2013).

Image processing is a method to perform some operations, i.e. computer operations or mathematical operations, on either an image or series of images in order to achieve an enhanced image or to segment/isolate the interesting object from the background or to extract some useful information, i.e. a set of characteristics and parameters related to the image (Gonzalez, Woods, & Eddins, 2004). Moreover, with the advantage of using computer hardware incorporated with image processing, techniques in various medical applications have been developed to imitate expert diagnosis procedures for a variety of blood disorders/blood cancers, such as an analysis of nuclear stained cells (Held & Banks, 2013), an image processing application for the localisation and segmentation of lymphoblastic cells using peripheral blood images (Madhloom, Kareem, & Ariffin, 2012b). In the analysis procedures of blood cancer diagnostic applications, the stained blood smear slides are placed under the digital microscope or light microscope attached to a digital camera for scanning the target cells and, then, the microscopic images or digital images of the field of view targeting cells are obtained. Consequently, the digital images or scanned microscopic images are passed through the process of quantitative analysis by the processes of a digital image processing-based system for blood disease detection, generally including four stages: image pre-processing and enhancement, image segmentation, image feature extraction, and feature classification/detection.

The aim of the image pre-processing and enhancement stage is to remove artefacts and noises from the input image and to adjust image quality to benefit the segmentation stage. Next, the image segmentation stage is processed to isolate the interesting objects for analysis, such as WBCs or RBCs, from the noisy background with other objects in the image. It also is a crucial and difficult stage in the image analysis owing to the segmented targeting cells influencing the accuracy of diagnostic results. An image feature extraction stage is processed to extract/measure the interesting characteristics or features of the segmented cells into quantitative data representation, such as textures, colours or intensities, and the morphological shape of each segmented cell image. Finally, an image feature classification/detection is computed to recognise the cell features of both normal and abnormal conditions of the sample cell images. In addition, the aim of feature detection is to distinguish between normal/healthy and abnormal/unhealthy lymphocyte cells in the new unseen or testing cell samples.

Since image analysis and computer technology have become essential in a digital diagnosis system, the next section illustrates the state-of-the-art developments for a digital acute leukaemia diagnosis system using microscopic blood images. This research study reviews and categorises the literature review of the state-of-the-art developments for a digital acute leukaemia diagnosis system into five stages, including image segmentation for leucocytes and

separation of nucleus and cytoplasm techniques for the identified cell membrane images, image feature extraction, image feature selection and image feature detection/classification, as follows.

2.4. Image Segmentation for Leukocytes and Image Separation of Nucleus and Cytoplasm Techniques for the Identified Cell Membrane Images

This section indicates the literature review of leukocytes, known as WBCs, image segmentation and image separation of cell nucleus and cell cytoplasm techniques for the identified cell membrane images. There have been active researches in the field of blood cell segmentation and various methods have been proposed. They are usually not only based on a single image processing technique, but also the integration of as many of them as can contribute benefits for an accurate blood cell segmentation and separation of its cell elements, i.e. nucleus and cytoplasm, into the acute leukaemia detection system.

From the review of literature, the group of common techniques adopted for segmentation of blood cells elements, such as only nucleus or nucleoli and both nucleus and cytoplasm or cell membrane, from microscopic blood cell images, include threshold-based, region-based, edge-based, clustering-based, and morphology-based approaches (Fatma, 2014; Gautam, Bhadauria, & Singh, 2014). The information of each group of common segmentation techniques is described as follows.

2.4.1. Threshold-based Segmentation Techniques

The threshold method is widely used and is claimed to have a fast performance in the segmentation of microscopic blood images. The underlying principle of this method is that the cells or objects and their background are at different intensity levels. Some researchers have chosen to keep a fixed threshold value while others employed adaptive values. For examples, Otsu's global thresholding method was employed by Scotti (2005) to separate nucleus from cytoplasm for an automated classification of normal and abnormal lymphocytes in greyscale blood smear images. Abbas and Mohamad (2014), Gautam et al. (2014) and Kulkarni et al. (2014) also applied Otsu's thresholding method for nucleus or nuclei segmentation of leucocytes. The binary image of the segmented nucleus, which is a result of the Otsu segmentation method, is then further processed with the morphological operations to clean and remove artefacts, which are not the WBCs, in the image. Liao and Deng (2002) proposed a WBC image segmentation using both a simple thresholding approach combined with mathematical morphology operations and contour identification. The algorithm is based on prior information and assumption that the blood cells are round boundaries. Dorini et al.

(2007) introduced an algorithm based on size distribution information of RBCs to segment a cytoplasm of WBC using Otsu's thresholding method and morphological operation. Moreover, the nucleus of the WBC was segmented using the watershed transform base on the image forest transform. Furthermore, Khasman and Abbas (2013) proposed a fast and cost-effective method for ALL identification, which applied Otsu's thresholding, Canny edge detection and pattern averaging kernel are used in order to achieve the boundary of single white blood cell images for their classification system. However, the threshold method is not able to perform well when the segmentation clusters have very small variances.

2.4.2. Region-based Segmentation Techniques

Region-based segmentation methods are the process of finding connected regions of objects or cells based on similar properties, e.g. brightness, colour, texture of pixels and then combining them together, as the same region, to increase their connected region, if they have similar properties corresponding to the defined criteria (Dougherty, 2009; Marques, 2011). For example, Halim et al. (2011) used the S-component of the Hue-Saturation-Intensity (HSI) colour space and set the fixed threshold value as 100 to segment the nucleus of the lymphocyte cell images followed by median filter and region growing techniques to obtain the region area of nucleus in pixels. Owing to the variations that could arise from different medical microscopic imaging databases, the parameter setting of their proposed approach might fail to provide a consistent performance for images across databases.

2.4.3. Edge-based Segmentation Techniques

Edge-based segmentation or deformable models or boundary-based segmentation methods are the process of finding pixel differences along the closed boundaries of foreground objects or cells, known as an inside/internal boundary, and the background, known as an outside/external boundary (Dougherty, 2009). Moreover, in deformable models, e.g. active contours, or snakes, the finding of boundaries of the interesting objects or cells is processed by evolving the contours or surfaces that are guided by internal and external energy to fit the object boundaries, which are satisfied by minimising the energy of the contours as a summation of internal and external energy (Kass, Witkin, & Terzopoulos, 1988). For example, Kumar et al. (2002) introduced a Teager energy operator for edge detection in nuclei segmentation of leucocyte sub-images, whereas Piuri and Scotti (2004) employed Canny's edge detection technique along with morphological operations to segment leucocyte cell membrane. Although image edges could provide rich information for recognition of image characteristics, edge detection methods tend to be sensitive to the image quality and noise (Lakshmi & Sankaranarayanan, 2010). As such, a good background and foreground contrast is important for enhancing the

detection performance (Joshi, Karode, & Suralkar, 2013). Furthermore, Ongun et al. (2001) used an active contours method with balloons algorithms to segment leucocyte cell membranes. In contrast, the contours are difficult to initialise around the region of interest of objects. Sadeghian et al. (2009) used the gradient vector flow algorithm, which is an extension of the active contours method that converge to concavities and also does not need to be initialised close to the boundaries of objects, to segment nucleus, employing Zack thresholding to segment cytoplasm of lymphoblast cell images.

2.4.4. Morphological-based Segmentation Techniques

Morphological-based segmentation is the process of finding the object's region by employing mathematical morphological operations, e.g. erosion, dilation, opening, closing, etc., and morphological tools, e.g. watershed transformation, morphological gradient, distance function, etc., in the process of segmentation (Beucher & Meyer, 1992; Soille, 2004). Watershed transform is one of the main morphological tools (Beucher & Meyer, 1992) to segment objects in greyscale image into the region of interest, which is indicated by its label number, and is adaptable to apply to different types of image objects, and is also capable of distinguishing extremely complex objects (Gonzalez & Ballarin, 2009). It has been used in many fields, such as medicine, biomedicine, industry, computer vision, remote sensing, computer-aided design, video coding and more (Pan, Zheng, & Wang, 2003; Sun & Luo, 2009). There are some active researches employing watershed transform for leucocyte segmentation. For example, Madhloom, Kareem and Ariffin (2012b) proposed an algorithm based on morphological reconstruction to localise and segment the whole cell of WBCs from the microscopic blood smear image for the diagnosis of acute lymphoblastic leukaemia. Srisukkham et al. (2013) introduced a method to segment the WBC membranes using integration of modified marker-controlled watershed transform with morphological operations. Additionally, this method can segment WBC membrane using microscopic blood smear sub-image and then isolates and places it on a white background. Moreover, Pan et al. (2003) proposed the robust Hue-Saturation-Values (HSV) based on colour image segmentation of leukocyte cells employing mean shift procedure and marker-controlled watershed algorithm. Jiang et al. (2003) applied a scale-space filtering technique to extract nucleus region from WBC sub-images, which was followed by watershed clustering to extract the cytoplasm region for the segmentation of WBC sub-images. However, although watershed segmentation was able to produce boundaries with closed and connected regions, over-segmentation could occur (Amoda & Kulkarni, 2013; Pan et al., 2003), as a common problem in the traditional watershed segmentation, if image information and a priori knowledge are not fully utilised for the process of watershed operation (Pan et al., 2003).

2.4.5. Clustering-based Segmentation Techniques

Clustering-based methods, as the name suggest, are the unsupervised classification of image pixels or feature vectors into different groups, known as clusters or classes (Jain, Murty, & Flynn, 2000). It is categorised into two main types: hard clustering, known as exclusive clustering, wherein one data sample belongs to only one cluster, and soft clustering, known as fuzzy clustering, wherein one data sample belongs to one or more than one cluster (Jain et al., 2000). K-means algorithm is one of the hard clustering methods and clusters data samples into *k* clusters based on similarity, by measuring the distance between the data sample and the cluster centres (Jain, et al., 2000; Nilima, Dhanesh, & Anjali, 2013). Fuzzy c-means (FCM) is one of the soft clustering algorithms and assigns/clusters a membership to each data sample. In addition, a data sample can belong to multiple clusters (Bezdex, 1981; Jain et al., 2000; Naz, Majeed, & Irshad, 2010). There are active researchers who have applied clustering methods in microscopic blood images segmentation. For example, several clustering techniques have been investigated by Mohapatra and his fellow researchers (Mohapatra, Patra, & Kumar, 2012; Mohapatra, Patra, & Satpathi, 2010; Mohapatra & Patra, 2010; Mohapatra et al., 2014). As an example, Mohapatra et al. (2012) employed hard clustering techniques, including K-means, K-Medoid, and fuzzy clustering methods, such as FCM, Gustavson Kessel and Fuzzy Possibilistic C-means, for locating the nucleus of WBC images in ALL detection. Kernel Induced Rough C-means clustering (Mohapatra, Patra, Kumar, et al., 2012) and shadowed C-means (Mohapatra et al., 2014) were applied to segmentation of nucleus and cytoplasm, and identification of lymphocyte cell images. In addition, Nasir et al. (2011) employed K-means clustering on the H and S components of the HSI colour space for segmentation of nucleus and membrane of WBCs. Pronab et al. (2014) proposed the WBC segmentation employing FCM algorithm followed by morphological operations, i.e. erosion and dilation, as post-processing, to identify white blood cells from the blood images. Since clustering techniques rely heavily on the principles of intra-class similarity and inter-class separability to perform grouping, these similarity and separability measures play significant roles in determining the resulting cell segmentation quality (Patil & Deore, 2013).

In addition to the separation of nucleus and cytoplasm of the WBC membrane, pastel blue and non-granular cytoplasm with closed and clumped nucleus chromatin are usually observed in mature lymphocytes (Rodak & Carr, 2012; Turgeon, 2012). For the blasts or unhealthy lymphocyte cells, variations in terms of nucleus to cytoplasm ratio, existence of nucleoli and vacuoles, nucleus and cytoplasm colour as well as chromatin patterns are observed. Therefore, discrimination of cell nucleus from cell cytoplasm and the characteristics of nucleus and cytoplasm play significant roles in accurate diagnosis of normal and abnormal lymphocytes.

Moreover, according to Rezatofighi and Soltanian-Zadeh (2011), improvement of nucleus and cytoplasm segmentation is the most challenging step that consumes most research efforts. In one of the main tasks of this research study, we have studied and focused on the robust separation of nucleus and cytoplasm of the white blood cell membrane, particularly the healthy and unhealthy lymphocyte cells (Chapter 4 for further details), for a robust ALL detection system.

2.5. Image Feature Extraction

Feature extraction is the process that measures certain properties, known as features, or converts the segmented images/objects into quantitative data representation, known as alphanumeric data, which include either letters or numerals (Dougherty, 2009; Marques, 2011). Moreover, feature extraction is another major step in contributing to accurate recognition of normal and blast lymphocyte cells. Features or descriptors, which are commonly extracted/measured from the segmented microscopic white blood cell image, include shape, colour, texture and statistical-based information. In general, shape-based features are related to geometric information, such as area, compactness, centroid, form factor, major and minor axis lengths, orientation, perimeter, elongation, and eccentricity, while colour-based features refer to the type of colour space information, such as RGB, CIE L*a*b* (CIELAB) and HSI. For texture-based features, Gray Level Co-occurrence Matrix (GLCM) provides information such as homogeneity, contrast, energy, correlation, cluster shade, cluster prominence and entropy. As for statistical-based features, information such as mean and standard deviation is often used, particularly for calculation of mean and standard deviation in colour and texture based features (Amnis Corporation, 2010; Nixon & Aguado, 2008).

There are active researches that have extracted features from segmented cell images, i.e. WBCs, and employed them in their recognition/classification systems. For example, Ongun et al. (2001) adopted affine invariants, the CIELAB colour space, colour histogram and shape-based features, from the heuristic reasoning of a haematologist to form a total of 57 features for classification of 12 types of blood cells, such as monocyte, neutrophil, myelocyte, plasma, etc. Putzu et al. (2014) focused on the detection of abnormality for lymphocyte images. A total of 30 shapes, 21 colours, and 80 GLCM-based texture descriptors were extracted from the normal and abnormal lymphocyte cell sub-images. Moreover, Rawat et al. (2015) introduced a computer aided diagnostic system to differentiate lymphoblast (abnormal lymphocyte) cells from normal lymphocyte images employing 26 GLCM texture features of nucleus and cytoplasm and 11 shape-based features of nucleus for their recognition system. Besides the GLCM textural and shape-based features, some researchers employed different methods to

interpret textural information from the segmented cell images. As an example, Local Binary Pattern (LBP) textural extraction was proposed by Singhal and Singh (2014) for detection of lymphocytes and lymphoblasts, while Rezatofighi and Soltanian-Zadeh (2011) employed LBP features of the segmented WBC images for an automated WBCs recognition. In addition, Hausdorff Dimension (HD) was adopted by Mohapatra et al. (2014) to extract roughness of the nucleus boundary pertaining to lymphocytes and lymphoblasts for an automated recognition of the lymphoblasts system. Meanwhile, Madhukar et al. (2012) proposed a decision support system for ALL classification and employed shape-based, texture-based and HD features, which extracted the segmented nuclei images, to distinguish normal and blast cell images.

In this research study, we form a set of 80 descriptors that are utilised in the subsequent steps, i.e. feature selection and ALL detection. The details of all features for this thesis are described in Chapter 4, Section 4.4.

2.6 Image Feature Selection

In the field of acute leukaemia detection, several researches have undertaken the process of feature selection that is needed to reduce the redundancy of the non-significant features and increase the efficiency of the recognition system with the significant features (Escalante et al., 2012; Madhloom et al., 2012a; Mohapatra et al., 2014). A feature selection task is also a crucial and more challenging task that selects the significant discriminative characteristics from the raw features and then employs the selected ones to support the recognition process for a highly accurate and robust acute leukaemia detection system.

The available data for an analysis task may comprise more numbers of irrelevant or redundant features. A relevant feature can influence the learning task to achieve high classification accuracy. Alternatively, a redundant feature is highly correlated with other features and can reduce the performance of the learning task to perform low recognition accuracy. Therefore, a good feature subset comprises features, that are highly relevant to the learning task, i.e. highly correlated to a decision variable and uncorrelated to other features (de la Iglesia, 2013). Moreover, the selection of feature subsets influences the performance of classification results. In terms of techniques of feature subset selection, there are three major techniques: filter, wrapper and embedded techniques, and their details are as follows. The filter techniques implement the feature selection by working on the general characteristics of training data without interaction with a classifier, which affords this method low computational cost. The wrapper techniques employ a classifier in the feature selection process to achieve the optimal feature subsets; however, wrapper methods are computationally expensive owing to they have

to interact with a classifier many times to achieve the quality of selected feature subsets. Moreover, the embedded techniques conduct the feature selection in the process of training, which interacts with a classifier and its learning algorithm, and these methods have computational cost in-between the filter and the wrapper methods (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013; Tang, Alelyani, & Liu, 2014).

As mentioned above, some researchers in the field of acute leukaemia recognition and detection have employed feature selection methods in their works. For example, Escalate et al. (2012) applied PSO algorithm, which is one of the evolutionary computation methods, to guide the search population and select the classification models to build ensemble classifiers automatically for acute leukaemia types or subtypes classification. Their research used Relief technique, which is categorised in filter methods, for the feature selection process. Furthermore, Madhloom et al. (2012a) employed Fisher's discrimination ratio (FDR), which is one of the filter methods, for their feature selection. The FDR inputs thirty features into account and uses cross correlation among all features to rank and select the top seven features for the recognition process in order to differentiate between normal and abnormal lymphocyte cell images. On the other hand, Mohapatra et al. (2014) conducted an independent-sample *t* test, which is one of the filter methods, to evaluate the raw input of 44 features, and then employed the selected 32 features, which had statistically significant value, for the ALL early diagnosis. Furthermore, Huang and Hung (2012) proposed principal component analysis (PCA) to reduce the dimensions of features from 85 to seven in leucocyte recognition. PCA is an approach for dimension (feature) reduction, that searches for k of n-dimensional orthogonal vectors that can best be employed to represent the data or have high variance among the data, where $k \leq n$ (Han, Kamber, & Pei, 2012). In addition, Rezatofighi and Soltanian-Zadeh (2011) further proposed sequential forward selection along with FDR for the recognition of five types of WBCs. Despite the popularity of the filter-based feature selection approach, Osowski et al. (2009) proposed an embedded approach for recognition of 11 types of blood cells (e.g. basophilic erythroblast, neutrophilic myelocyte, lymphocyte, etc.) with the integration of the GA to fine tune the feature subsets corresponding to the SVM performance during the training stage.

Moreover, there are many feature selection and dimension reduction techniques, such as mutual information (MI), minimum-redundancy-maximum-relevance criterion (mRMR), gain ratio, and evolutionary computation methods, such as GA, genetic programming, and PSO, etc. Furthermore, the meta-heuristic optimisation algorithms, for instance, DA, CS, and those related to PSO variants algorithms, e.g. enhanced leader particle swarm optimisation (ELPSO) and BBPSO, have become popular and useful for solving a variety of optimisation problems

and for feature subset optimisation as well. Details of the aforementioned algorithms, which are usually used for the feature selection and dimension reduction, are as follows:

MI is an information-based feature selection algorithm that measures how much information the presence/absence of a selected feature contributes to making a correct classification decision on the target groups/classes (Manning, Raghavan, & Schütze, 2008). In addition, it is also able to maximise information in a group/class (Zhang, Zhang, & Hossain, 2015c).

The mRMR, as proposed by Peng et al. (2005), aims to minimise the mutual information, (i.e. a redundancy) among the selected feature subset and to maximise the mutual information, (i.e. a relevance) between the selected features and the targeted output (Zhang, Zhang, & Hossain, 2015c).

Gain ratio is an extension to information gain or the modification version of information gain, which attempts to reduce its bias. When choosing an attribute/feature, gain ratio takes number and size of branches into consideration. It modifies information gain by taking the intrinsic information of a split information into account. Intrinsic information is an entropy of distribution of instances into branches, i.e. how much information do we need to tell which branch an instance belongs to. Gain ratio value of an attribute/feature decreases as intrinsic information gets larger (Han et al., 2012; Priyadarsini, Valarmanthi, & Sivakumari, 2011).

GA is the most classic and widely used evolutionary algorithm. It draws inspiration from the Darwinian evolution theory, in survival of the fittest in human or animal societies. In the process of GA, each individual has its information chain, which is a fixed-length binary array or binary string, e.g. '101101001101', as its genotype. Then, the fitness of each individual is calculated for the selection process. The algorithm then processes to select parents for one-point crossover to produce offspring individuals, which subsequently undergo mutation operations. The offspring individuals become the population in the next generation. A termination of the GA occurs when the fittest reaches satisfaction or the maximum number of generations is achieved (Bäck & Schwefel, 1993; Wong, 2016; Zhang et al., 2015d).

PSO was inspired by the behaviour of bird flocking. It was devised by Kennedy and Eberhart (1995) and has been widely used to solve optimisation problems in many fields. In the process of PSO, it uses two main variables, which are personal best position, known as *pbest*, and global best position, known as *gbest*, in its search mechanism, to move the position of particles (solutions) in search spaces towards the optimal solution(s). Moreover, the PSO needs to adjust many parameters, such as velocity and weight inertia, to initialise the algorithm and

incorporate with *pbest* and *gbest* to move the particles in search spaces (Section 5.3.1 in Chapter 5 for further details).

BBPSO algorithm is one of the PSO variants and is a compact and parameter-free algorithm. It was invented by Kennedy (2003), who was one of the PSO inventors. The BBPSO employs Gaussian distribution instead of velocity in the PSO and also incorporates with *pbest* and *gbest* to move the particles in search spaces towards the optimal solution(s) as well (Section 5.3.1 in Chapter 5 for further details).

DA is a nature-inspired meta-heuristic optimisation algorithm and was proposed by Mirjalili (2015). In the process of DA, the dragonflies move/swarm for only two specific purposes, including hunting, known as static (feeding) move, and migrating, known as dynamic (migratory) move, towards the best food source(s) or optimal solution(s). These static and dynamic moving behaviours cause the DA to be different from the PSO and other meta-heuristic optimisation algorithms (Section 5.3.3 in Chapter 5 for further details).

CS is a meta-heuristic searching algorithm for continuous optimisation, which was proposed by Yang and Deb (2009). This algorithm is also a nature-inspired optimisation algorithm. The search mechanism of this algorithm is based on an interesting reproduction strategy, such as the brood parasitism of cuckoo birds (Yang & Deb, 2009). In particular, it integrates with Lévy flight behaviours and has been applied to optimisation and optimal search with promising results in the field of science and engineering. Furthermore, research reveals the performance of CS is far more efficient and can outperform other meta-heuristic algorithms, i.e. PSO, GA, for many optimisation problems (Ljouad, Amine, & Rziza, 2014; Yang & Deb, 2010) (Section 5.3.2 in Chapter 5 for further details).

Jordehi (2015) proposed an ELPSO algorithm, which employs successive mutation strategies, such as Gaussian, Cauchy, opposition-based and differential evolution-based mutation, to further enhance the swarm leader to search in search spaces towards the optimal solution(s). Evaluation results indicate its efficiency in terms of accuracy and scalability.

Zhang et al. (2015b) proposed a binary BBPSO-based feature selection algorithm. Their work used a reinforced memory strategy for personal best updating of each particle to retain particle diversity. In addition, it also used a uniform combination to diversify the swarm, when stagnation[1] occurred. The effects of uniform combination were strengthened along with the

[1] Stagnation is the situation wherein the search algorithm finds no improvement of fitness value of the global best solution from iteration to iteration.

increase of stagnant iterations. The binary BBPSO showed competitive performance in terms of classification accuracy and convergence rate.

2.7 Feature Detection/Classification

In an acute leukaemia recognition and classification system, one of the main tasks is the recognition of normal and abnormal lymphocytes that employs a classifier for learning knowledge from microscopic blood image samples. In this section, we describe the classifier techniques employed for the acute leukaemia detection. Details of the related research are as follows.

2.7.1 Multi-layers Perceptron (MLP)

MLP is one of the most popular supervised neural networks modelling techniques and has been widely used in pattern recognition, computer vision, bioinformatics and control systems (Howard & Mark, 1998; Mohapatra et al., 2014). It has also been applied to WBC identification and classification systems in related works. For example, Piuri and Scotti (2004) proposed the leucocytes classification system using the morphology of the microscopic blood cell images as a set of features for the recognition system to classify five types of white blood cells. In addition, three types of classifiers, i.e. k-Nearest Neighbour (kNN), feed-forward neural networks (FF-NN) and radial basis function neural networks (RBFN), are employed in the classification process and the parallel of five FF-NNs provides the highest recognition accuracy in their research. Mohapatra, Patra and Stpathy (2014) introduced a multiple classifiers system for early ALL detection using microscopic blood images. They employed the combination of classifiers, i.e. kNN, MLP and SVM, to the classification system and performs experimental results with higher accuracy compared with the results of the single classifiers, which are produced from a single classifier of kNN and MLP, except a single SVM classifier. Moreover, Teera-Umpon and Dhompongsa (2007) presented an automatic WBCs classification employing morphological granulometric features of only the cell nucleus of microscopic bone marrow images. Their classification results show that the neural networks classifier performs with the highest classification accuracy against Bayes and Decision Trees (DT) classifiers. In addition, Khasman and Abbas (2013) proposed a fast and cost-effective method for ALL identification that employed the extracted pixels of the boundaries of each of segmented white blood cell images as the inputs for classification using MLP. Three learning strategies with different ratios of training and testing sets (a percentage of training/a percentage of testing), i.e. 75%:25%, 50%:50% and 25%:75%, respectively, were employed to evaluate their system performance. The ratio of 75%:25% learning strategy produced the

highest classification with accuracy of 90%, as compared with those from other schemes in their work.

2.7.2 Support Vector Machine (SVM)

SVM is a kernel-based classification technique. The basic idea of SVM is to compute a linear function in a higher dimensional feature space, where the lower dimensional input data are mapped using a kernel function (Basak, Pal, & Patranabis, 2007). It possesses strong regularisation properties that are able to produce generalised models for any new datasets (Agaian et al., 2014). In SVM, a hyperplane is constructed in a high-dimensional feature space to classify data based on a set of support vectors that are members of the training samples. The hyperplane with the largest functional margin to the nearest training sample of any class usually produces lower generalisation error and gives better separation between classes (Jain, Duin, & Mao, 2000). Moreover, SVM has been adopted to leukocytes/WBCs recognition and classification applications. For example, Ongun et al. (2001) proposed an automated WBCs counter for a differential blood count system, which used extracted features for the recognition process with variety of classifiers, i.e. kNN, linear vector quantisation, MLP and SVM. Their evaluation results showed that the SVM performs with the highest classification accuracy compared to other baseline classifiers. Furthermore, Putzu, Caocci and Di Ruberto (2014) proposed an automated WBCs classification and identification for ALL detection using microscopic images. The SVM, Naïve Bayes and DT classifiers were employed for the recognition process to identify lymphocyte and lymphoblast cell images. In evaluation results of their work, a single SVM with Gaussian radial basis function (RBF) kernel achieved the highest recognition accuracy against the other classifiers. Madhukar, Agaian and Chronopoulos (2012) introduced a new decision support tool for ALL classification employing microscopic blood smear images. They used a SVM classifier for learning about the microscopic blood image samples, that each image contained multiple nuclei, to identify ALL.

2.7.3 Ensemble Classifier

Ensemble classifier is the integration of classifiers with the combination rules, which aims to improve classification performance (Rokach, 2010). There are nine weighting strategies, known as combination rules, which are usually used in the combination outputs of classifiers, known as classifiers fusion or ensemble classifiers, to obtain the final classification result, including majority voting, minimum and maximum probability, distribution summation, average of probabilities, product of probabilities, Bayesian combination, decision templates

and Dempster-Shafer. Brief information regarding the aforementioned combination rules is as follows:

A. Majority Voting

This combination rule uses the output labels of classifiers to process the final classification or prediction result by either all classifiers predicting with the same class, known as unanimous voting, or at least one more than half of the number of classifiers predicting the same class, known as a simple majority vote, or the highest number of classifiers predicting for the class (the most frequent class), known as the plurality vote, which can process even when the highest number of votes is less than half the number of classifiers (Kuncheva, 2014; Polikar, 2006; Rokach, 2010).

B. Minimum, Maximum, Average and Product Probabilities

These combination rules mainly use the continuous output values in range [0,1] of the individual classifiers, which are estimates or predictions of the posterior probability by the classifier as the degree of support for a given input sample to each class in the system, to obtain the final classification for each class by using algebraic functions, i.e. minimum, maximum, average or mean, and product, with the supported results from the individual classifiers, which are stored in the decision profile matrix (DP) (Kuncheva, 2014; Polikar, 2006). Kuncheva et al. (2004;2014) defined the DP, which represents the outputs degree of support given by the classifier to each class in the system, for the combination rules, which use continuous values of the posterior probability output from classifiers, to obtain the final decision output. The DP consists of two elements, including each row, which represents the output given by a single classifier to each of the classes, and each column, which denotes the estimated output received by a particular class from all classifiers (Polikar, 2006). To obtain the final classification result of these combination rules, for example, using minimum function as the algebraic combiners to classify one input sample, after creating the DP, each column of class finds one minimum posterior probability value from all classifiers and then the class that has a maximum of minimum posterior probability value of each class is the final assigned classification class for that input sample.

C. Distribution Summation

The distribution summation combination rule combines a posterior probability of the individual classifiers from the DP that supports the same class and then the class that has a maximum of total summation values is the final decision class to assign for each input sample (Rokach, 2010).

D. Bayesian Combination

The Bayesian combination rule adopts the method where a posterior probability classification results in range [0,1] of each individual classifier given the training sample dataset, known as weight, multiplied to the probability of support of that class given an input test sample. Then the class that has the maximum result is the final assigned class label for that input test sample (Rokach, 2010).

E. Decision Templates

Kuncheva et al. (2004) proposed the decision templates combination rule, which calculates the mean of the decision profiles (DP) of all members of each class from the training dataset, known as the most typical decision profile (Kuncheva, 2014), and then brings it to compare with the current decision profile of an input test sample using similarity measure techniques, i.e. Euclidean distance. Finally, the class that has the smallest distance, known as closest match, will assign the label of that class to the input sample (Polikar, 2006).

F. Dempster-Shafer

The Dempster-Shafer combination method is inspired by data fusion, which is a subject area of data analysis mainly involved in combining elements of evidence that are provided by different sources of data. Data fusion is based on the Dempster-Shafer theory of evidence, which employs belief function (instead of probability) to measure the quantity of the evidence (the output of a classifier) by the source that generated the training data (Kuncheva, 2004; Polikar, 2006). Once the belief values are obtained for each classifier, known as source, they can be fused or combined by Dempster's rule of combination, which basically states that the evidence (belief values) from each source (classifier) should be multiplied to obtain the final support for each class (Polikar, 2006). Finally, the class that has maximum support values (degree of belief) in range [0,1] is then assigned to the input sample (Kuncheva, 2004).

2.8 Scope of the Research

This research study focuses on acute lymphoblastic leukaemia, which is the most common in childhood, owing to the chance of children who have screening tests for early detection of this malignant blood disease, and to be cured, with high survival rates from the appropriate treatments, as reported in Section 1.1 of Chapter 1 and Section 2.2 of Chapter 2. This has motivated us to develop an intelligent decision support system for ALL detection using microscopic blood smear images.

2.9 Chapter Summary

This chapter has described the extensive literature review of the relevant background in the field of biomedical engineering and haematology, including human blood, neoplastic disorders of blood, particularly leukaemia, and the laboratory diagnosis of ALL, past and present, through to using computerised technology to assist the experts, e.g. haematologists and haematopathologists, in diagnosis of the disease. In addition, the related works of the state-of-the-art research in the field of computer science involved in the quantitative analysis of microscopic blood smear images for acute leukaemia detection/classification have been presented. Moreover, the existing related researches in the development for ALL detection/classification under five sequential stages, including image segmentation and separation of the white blood cell membrane images, feature extraction of the segmented/labelled cells, feature selection of the extracted descriptors to reduce the redundancy of the non-significant features, and ALL identification, have also been explained. As observed in the related works, there are still challenging tasks for improvement of these successive stages for the quantitative analysis of ALL detection/classification. The next chapter describes the first key stage of this PhD research for an intelligent decision support system for ALL detection, namely the segmentation of the white blood cell membranes images.

Chapter 3: White Blood Cells Membranes Segmentation Using Marker-Controlled Watershed Method and Morphological Operations

3.1. Introduction

The first key stage of this research study for acute lymphoblastic leukaemia detection is the segmentation of the white blood cell membranes, specifically for the lymphocyte segmentation, which is the major white blood cell type for acute lymphoblastic leukaemia diagnosis. The aim of this stage is to isolate the lymphocyte/lymphoblast cell membrane from touching and overlapping of the red blood cells, platelets and artefacts of the microscopic peripheral blood smear sub-images and then further analyse the segmented cell membrane with the proposed quantitative image analysis methods, as presented in Chapter 4 and Chapter 5, to identify whether the segmented cell is healthy or unhealthy.

The watershed segmentation is the powerful segmentation technique, which is employed successfully in many domains, such as medical image analysis, computer vision etc. This technique can isolate the target objects in the complex image close to the boundary of these objects. However, the traditional watershed segmentation still has the problem of over-segmentation. The challenge to overcome the aforementioned problem is about how to create the good markers for the watershed transform to segment objects in a specific domain, i.e. white blood cell membrane segmentation, to achieve the high accurate segmentation results. This chapter presents the segmentation of white blood cell (leukocyte) membranes, particularly lymphocyte and lymphoblast cell types, based on the extension of our previous study (Srisukkham et al., 2013), using the microscopic sub-images of ALL-IDB2 database (Labati, Piuri, & Scotti, 2011a). Overall, it proposes a modified marker-controlled watershed algorithm integrated with the morphological operations, as shown in the top-right green rectangle dashed-line of Figure 3.1, for the segmentation of the membrane of lymphocyte and lymphoblast cell images. The structure of this chapter is as follows. Section 3.2 describes the overall system architecture of this research study, while Section 3.3 presents the details of the microscopic blood smear image database and the consultation with the haematologists to provide information about the materials and the ground truths of the microscopic blood image dataset for the experiments and evaluations in this research. Furthermore, Section 3.4 explains the proposed algorithm, i.e. the modified marker-controlled watershed algorithm, for the

lymphocytic membranes segmentation. Finally, the evaluation and discussion of the proposed segmentation method are illustrated in Section 3.5.

3.2. The Overall System Architecture of This PhD Research

The system architecture of this research study is given in Figure 3.1. It comprises five main stages: (1) the lymphocytic cell membranes segmentation and identification from microscopic blood smear sub-images; (2) the separation of nucleus and cytoplasm of each lymphocytic cell membrane; (3) the feature extraction; (4) the feature selection; and (5) the healthy/lymphocyte and unhealthy/lymphoblast cell detection.

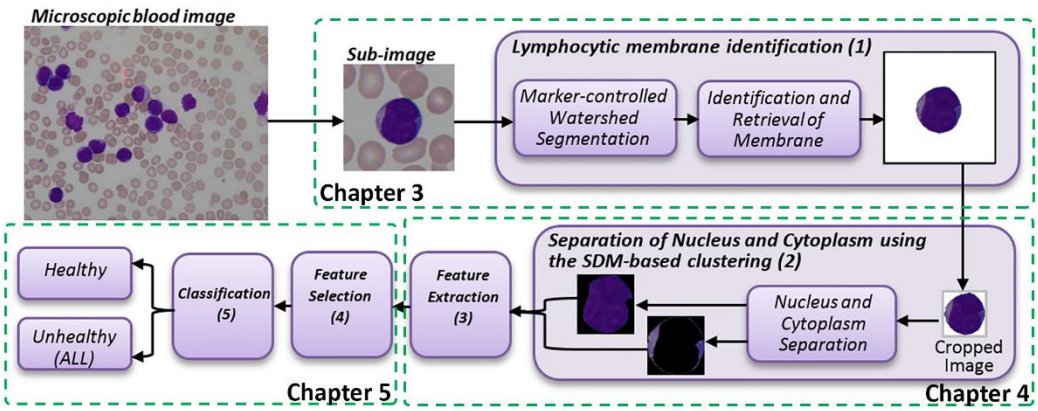


Figure 3.1 System architecture of this research study

Firstly, we used 180 sub-images of lymphocytic (healthy and blast lymphocyte) cells from ALL-IDB2 database (Labati et al., 2011a), including 120 unhealthy (blast) and 60 healthy lymphocyte cell sub-images, which were annotated by haematologists, to the lymphocytic membrane segmentation and identification as stage 1 and as presented in this chapter. This stage aims to segment the lymphocytic white blood cell membrane of input sub-images using the integration of modified marker-controlled watershed algorithm and morphological operations, and then identify each lymphocytic cell membrane with the retrieval of the identified one placed on the white background sub-image. Then, the separation of nucleus and cytoplasm of each identified lymphocytic membrane, as stage 2, is processed, as described in Chapter 4. The cropped images of the identified cell membranes are used as the input for the separation of the nucleus and the cytoplasm of each cell membrane using the proposed stimulating discriminant measure (SDM) clustering algorithm (Neoh et al., 2015), which is robust in terms of discriminating cell nucleus from cell cytoplasm of the identified lymphocytic membrane images with diverse irregular morphology. The results of this stage are the separated cell nucleus and cell cytoplasm images of each identified lymphocytic membrane of 180 sub-images. Subsequently, the feature extraction as stage 3 is also employed

to extract the significant discriminant characteristics of the separated nucleus and cytoplasm images of each identified lymphocytic membrane. Furthermore, the shape-based descriptors, texture-based GLCM descriptors, colour-based CIELAB colour space descriptors, and the statistical calculation of those descriptors are extracted according to the consultation with the haematologists and the state-of-the-art existing researches. Hence, we achieved 80 features for all identified 180 lymphocytic cells from this stage. Subsequently, the feature selection, as stage 4, as presented in Chapter 5, is used to select the significant discriminant characteristics of the extracted features for the robust and efficient lymphocyte and lymphoblast cell identification. We also proposed a novel feature selection algorithm, namely the BBPSO-based feature optimisation. This BBPSO variant algorithm uses the 80 extracted descriptors as the input features and then computes the results, which are the most significant feature subsets. Finally, the classifiers, i.e. SVM, MLP and ensemble of classifiers, use the feature subsets for the robust and efficient recognition process for the healthy and unhealthy (blast) lymphocyte cell detection.

3.3. Microscopic Blood Images from ALL-IDB Database and the Consultation with the Haematologists

This research study uses the public microscopic stained blood smear images database, namely the acute lymphoblastic leukaemia image database, for image processing: ALL-IDB (Labati et al., 2011a). The database was published by the department of Information Technology, UniversitàdegliStudi di Milano, Italy (Labati, Piuri, & Scotti, 2011b). The ALL-IDB database has high quality microscopic peripheral blood sample images of healthy volunteers and ALL patients. The image dataset has been collected by experts of M. Tettananti Research Center for childhood leukaemia and haematological diseases, Monza, Italy, and was captured with optical laboratory microscopes coupled with both Canon PowerShort G5 and Olympus C2500L digital cameras. Additionally, the various magnifications, ranging from 300 to 500 times, are applied when the images are taken in Red, Green, Blue (RGB) colour. In addition, the database provides the ground truth of healthy and unhealthy (suffering from ALL or blast) with a label annotation on each image as named by expert oncologists. The ALL-IDB has two distinct versions of dataset, ALL-IDB1 and ALL-IDB2. The ALL-IDB1 dataset is composed of 108 whole blood sample images. Furthermore, the ALL-IDB2 dataset is a group of 260 sub-images which have one white blood cell in each sub-image, which is cropped area of interest of normal and blast cells from the ALL-IDB1 dataset. In this research study, the dataset of microscopic blood smear sample images is collected from ALL-IDB2 dataset, the 180 sub-images, which comprise of 60 lymphocyte cells and 120 lymphoblast cells images,

are taken into consideration. The sub-images are selected according to the consultation with the haematologists.

A meeting was conducted in consultation with the haematologists in the Royal Victoria Infirmary (RVI Hospital at Newcastle-Upon-Tyne, United Kingdom) to identify the criteria for clinical diagnosis of ALL. The haematologists categorised the criteria for diagnosing ALL in three groups including: (i) outside information criteria, i.e. low numbers of haemoglobin (Anaemia), low numbers of platelets (Thrombocytopenia), enlarged liver, fever, loss of weight, childhood age and bone pain; (ii) general information from the blood film criteria, i.e. the shape of RBCs not round as in doughnut-shaped (Anaemia), the presence of RBC teardrop poikilocytes, and the shape and size of platelets larger than normal, which are similarly the size of small lymphocyte cells (Thrombocytopenia); and (iii) specific features of lymphocytes (healthy cell) criteria, i.e. the size larger than normal lymphocyte, high ratios between nucleus and cytoplasm of each lymphocyte cell membrane, deeply basophilic cytoplasm, and open or fine chromatin. However, as the presence of only one criterion of each group is not enough to identify ALL, all criteria of all the groups will be integrated together for the diagnosis of ALL.

According to the consultation with the haematologists and with the clinical diagnosis criteria based on haematologists' experiences, most consultation information is similar to the clinical diagnosis of ALL, as described in Chapter 2, Section 2.2. Moreover, the valuable information from clinical diagnosis criteria and the significant descriptors of the lymphocytic cells for the quantitative image analysis from the state-of-the-art researches, as aforementioned in Chapter 2, Section 2.6, is selected to use in the feature extraction stage of this research. Additionally, the sub-image samples of the lymphocytic cells with the ground truths and annotations from the haematologists that are employed in this research are depicted in Figure 3.2.

The samples of ground truths and annotations of the lymphocytic sub-images, as depicted in Figure 3.2, are used in the evaluation step of this chapter, as presented in Section 3.5. The ground truths of lymphocytic membrane are used in the comparison with the segmented cell membranes results of the modified marker-controlled watershed and the traditional marker-controlled watershed algorithms by employing the two-dimensional correlation coefficient as the comparison technique.

3.4. The Proposed Modified Marker-Controlled Watershed Algorithm for the Lymphocytic Membranes Segmentation

In this section, we present proposed segmentation algorithm, which is the integration of the modified marker-controlled watershed transform and the morphological operations, to

segment the lymphocytic cell membrane images. As aforementioned in Chapter 2, Section 2.4, watershed transform is a powerful morphological tool that is adaptable to apply to different types of image objects and is also capable of distinguishing extremely complex objects (Gonzalez & Ballarin, 2009). It has been used in many fields, such as medicine,

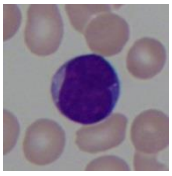
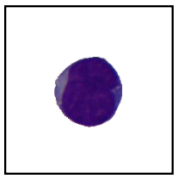
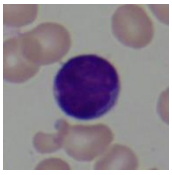
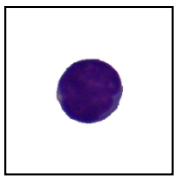
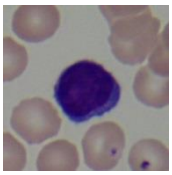
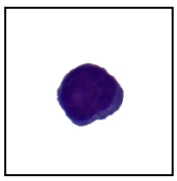
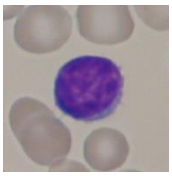
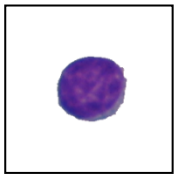
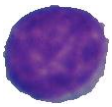
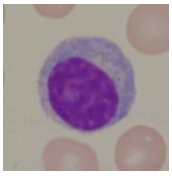
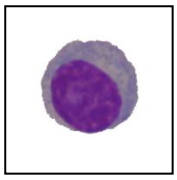
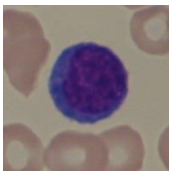
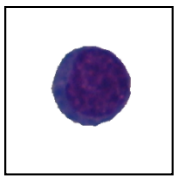
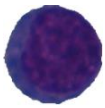
Lymphocytic cells file name and condition	**Original WBC sub-images**	**Ground Truths**	
		WBC membranes	**Cropped membranes**
Unhealthy (lymphoblast) Im004_1			
Unhealthy (lymphoblast) Im015_1			
Unhealthy (lymphoblast) Im024_1			
Healthy (lymphocyte) Im156_0			
Healthy (lymphocyte) Im196_0			
Healthy (lymphocyte) Im204_0			

Figure 3.2 The sub-image microscopic blood samples of the lymphocytic cells with ground truths and annotations from the haematologists

biomedicine, industry, computer vision, remote sensing, computer-aided design, video coding and more (Pan, Zheng, & Wang, 2003; Sun & Luo, 2009). However, the drawback to the watershed transform is over-segmentation, which may occur as a common problem in the traditional watershed segmentation if image information and a priori knowledge are not fully utilised for the process of watershed operation (Pan et al., 2003). Therefore, a marker for each object of interest in the image has to be created before being applied to the watershed transform to overcome or avoid the over-segmentation (Gonzalez & Ballarin, 2009). The attractive performance of the watershed transformation motivates us to employ it as the base segmentation method in this research study.

In the operation of watershed transform, we assume that an input grayscale image can be converted as a topographic of landscape surface with three different local minimum areas, known as basins, as shown in Figure 3.3 (b). We drill holes at each local minimum area, as the starting point of water coming through the basin, and then we immerse this landscape surface into a lake. The water entrance into the holes and flood the surface where the water coming from two or more different minima or basins would meet, when the water level increases from level *i* to *j*, as shown in Figure 3.3 (a), and over flooding is caused by overflow the ridge to the walls of another catchment basin. Thus, a dam is built on the points of the landscape surface to avoid the merging of different floods, as shown in Figure 3.3 (a).

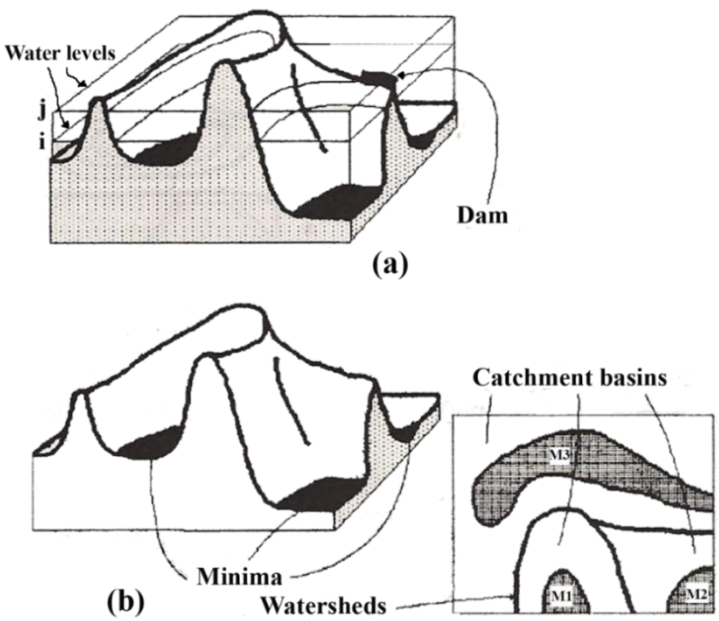


Figure 3.3 The watershed transform (a) Flooding of the surface, water levels: *i* and *j* and dam building; (b) Top view shows catchment basins, watershed lines and minimum areas: *M1*, *M2*, *M3* (Beucher & Meyer, 1992).

The process of flooding is stopped and only the dams become apparent when the water level reaches the highest point of the landscape surface. Therefore, the dams, or lines, known as watersheds or watershed lines, separate the landscape surface into various regions according to the different catchment basins, in which each one contains only one minimum, as shown in Figure 3.3 (b) (Beucher & Meyer, 1992; Vincent & Soille, 1991).

The marker-controlled watershed segmentation is the watershed segmentation algorithm that needs some additional information, called markers, wherein the minima correspond to the objects and to the background, to be identified before the flooding process. Otherwise, the process of flooding and the building of dams are the same as the previously mentioned watershed transform. In addition to the building of dams in marker-controlled watershed algorithm, a dam is built only for separating floods which originate from different holed minima (Beucher & Meyer, 1992). Therefore, the good markers, which depend on specific application, are important for the marker-controlled watershed algorithm to perform the accurate segmentation results.

In this research study, we proposed the integration of the modified marker-controlled watershed algorithm and the morphological operations to segment the lymphocytic white blood cell membranes using the microscopic blood smear sub-images from the ALL-IDB2 database. The procedure of this proposed method involves the following steps and the image results of each step are shown in Figure 3.4.

From Figure 3.4, first, conversion of an input original RGB lymphocytic image to grayscale is conducted. An image filtering technique is deployed to remove noise and then enhance the quality of images with contrast enhancement technique, as a pre-processing and image enhancement step. In order to avoid over-segmentation of the watershed transform, the integration of morphological operations with gradient magnitude, distance transform and assignment of infinity values to background is used to produce good seed markers, including foreground and background markers, for the watershed segmentation. In addition to segmenting complex image, the watershed transform also assigns a labelled indexing value to each segmented object. The most frequently labelled indexing value in the square area size of 10x10 pixels from the centre of the image is used to identify the lymphocytic membrane, and then only the binary pixels corresponding to the identified lymphocytic membrane are selected as a binary mask. The binary mask is subsequently used in the retrieval process of the original RGB pixels of the identified lymphocytic membrane. The details of each procedure according to the flow chart diagram of the proposed method are explained in subsequent steps as follows:

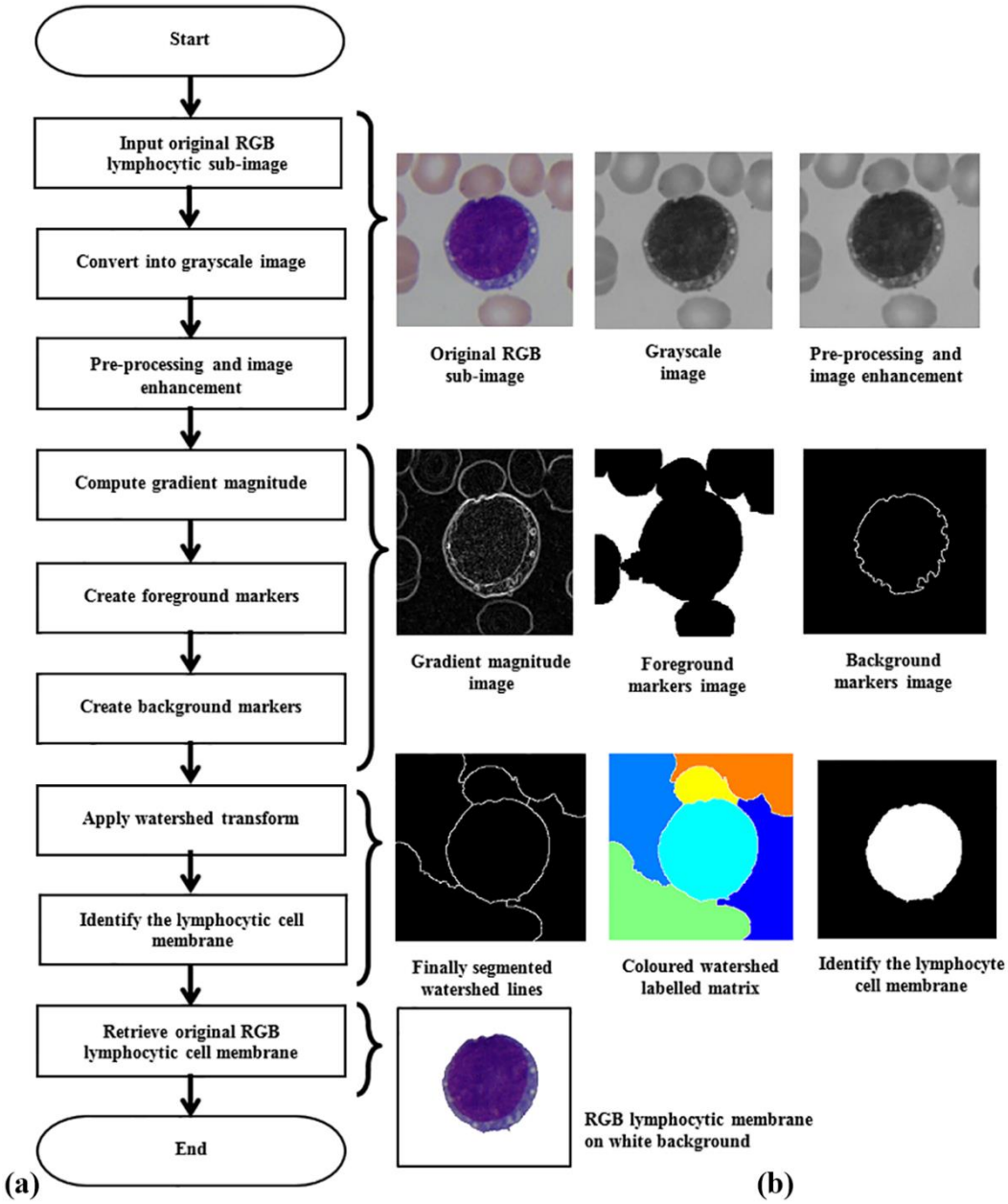
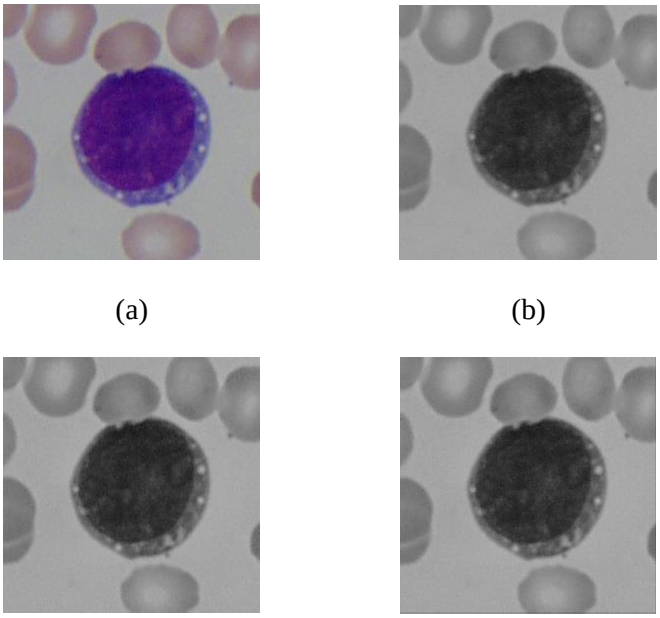


Figure 3.4 (a) Overview of the proposed method; (b) Image results of each step derived from the original RGB lymphocytic sub-image until the final result of RGB lymphocytic membrane on a white background.

3.4.1. Pre-processing with Filtering Technique and Image Enhancement

In this section, we introduce the pre-processing and image enhancement step applied to the input microscopic images for the segmentation process. The pre-processing step is a necessity for the segmentation task in this research study due to the presence of noise and acquisition of microscopic blood sample images under uneven lighting conditions, i.e. the variation of contrast in image. The watershed transform is sensitive to the noise in the image, as it can easily detect any noise to a catchment basin during the flooding process. Filtering helps to reduce noise in the image before applying the segmentation algorithm. In this research study, the noise in the input microscopic images is reduced by an operation of Gaussian low-pass filter technique, which was efficient and useful to reduce noise in images of our previous study (Srisukkham et al., 2013) and many other state-of-the-art researches (Scotti, 2006; Sun & Luo, 2009) to achieve promising segmentation results. For an uneven lighting condition of the input images, we employ the image contrast enhancement operation, i.e. the contrast limited adaptive histogram equalisation (CLAHE) technique (Zuiderveld, 1994), to resolve the variation of contrast and illumination of the input images, which usually occur in the image database. First, we convert the input microscopic image, which is in RGB colour space, into grayscale. The conversion of an RGB image to grayscale of this research study is conducted by using the method of Rec. 601 nonlinear luma component (Y') (MathWorks, 2016; Poynton, 1996). The conversion starts by multiplying coefficient matrix A, which is [0.299, 0.587, 0.114], with the colour components of RGB image or RGB matrix. Then, simplifying Y' = A(1) R' + A(2) G' + A(3) B', where Y' is the grayscale image matrix with R' as Red colour component matrix, G' as Green colour component matrix and B' as Blue colour component matrix. Then, the grayscale image is applied by the CLAHE technique and the Gaussian low pass filter size 5x5 with standard deviation (σ) of 0.5, respectively, as shown in Figure 3.5. This step can help to adjust the quality of input microscopic images before doing the image analysis tasks. The next section presents the process of creating the markers as the good seeds for the watershed segmentation.



(a) (b)

(c) (d)

Figure 3.5 The sample of images before and after being conducted with the pre-processing and image enhancement: (a) the original RGB image; (b) the result after conversion to grayscale image; (c) the result after applied the CLAHE; and (d) the result after employing Gaussian low-pass filter.

3.4.2. Marker Generation for the Watershed Algorithm

In this section, one of the important tasks for the lymphocytic white blood cell membrane segmentation with the marker-controlled watershed transform is presented. Before the explanation of the method of generating the markers for this research study, we introduce the mathematical morphological operations which are employed in the process of creating the good seed markers of this research.

The mathematical morphology is a powerful image analysis technique and a widely used tool in image processing for representing, describing and analysing shapes and form of objects or images (Marques, 2011; Soille, 2004). The concept of mathematical morphology is the extraction of geometrical and topological information from an image or object through transformations using operations, known as morphological operations, with another set of known shapes, termed the structuring element (SE). Moreover, the design of SEs, i.e. the shape and size of SE, in morphological image processing is essential to the success of the morphological operations that employ them (Marques, 2011; Soille, 2004). In this research

study, we utilise morphological operations, including SE, dilation, erosion, area opening, closing, opening and reconstruction, in the process of creating the good markers. These operations and descriptions are as follows.

SE is a shape, matrix of pixels of binary numbers, which determines the effect of morphological operations, i.e. dilation and erosion, in an image. The SE forms pattern relative to the origin. The origin can be any of its pixels, but can also be outside the SE. Examples of SE patterns, or shapes, are ‘disk’ and ‘diamond’, as shown in Figure 3.6.

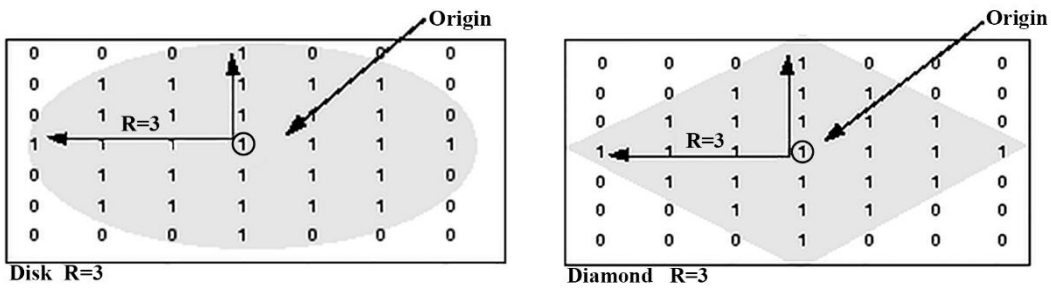
Dilation is a morphological operation that adds pixels to the object boundary or edge of an image and dilates it with respect to the shape and size of the structuring element. Dilation of an image A by structuring element S is written in Eq (3.1) as follows:

$$A \oplus S = \bigcup\{S + a : a \in A\} \tag{3.1}$$

Erosion is a morphological operation that removes pixels from the boundaries of an object and shrinks it with respect to the shape and size of the structuring element. If an image A is eroded by using structuring element S, it is denoted in Eq (3.2) as follows:

$$A \ominus S = \bigcap\{S - a : a \in A\} \tag{3.2}$$

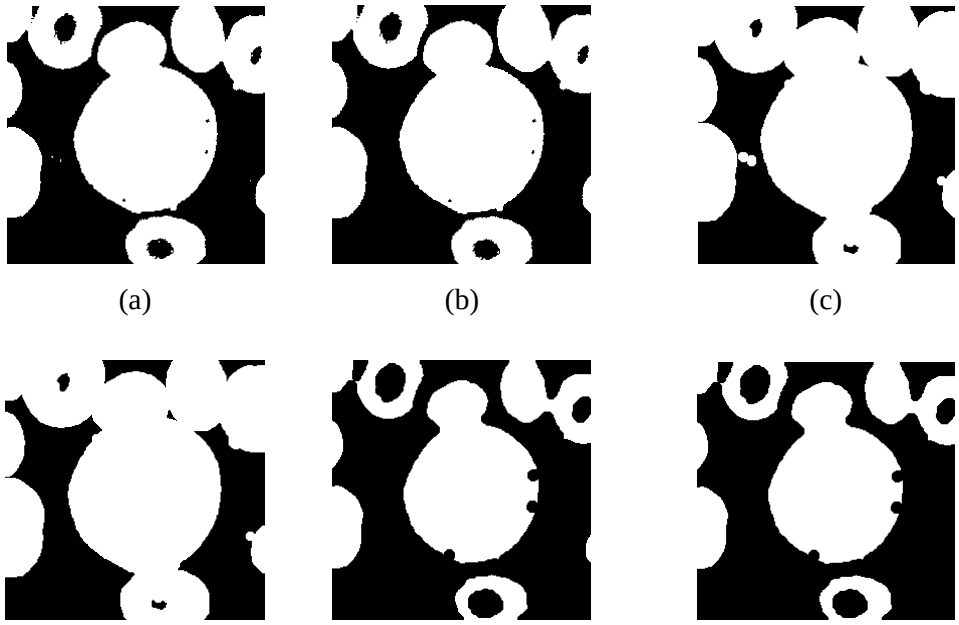
Where $\bigcup$ and $\bigcap$ is the set union and intersection, respectively, and a is pixel of image A.



(a) (b)

Figure 3.6 The example of SE shapes (a) Disk shape and (b) Diamond shape.

In addition, Figure 3.7 shows the examples of the variation in an image due to erosion, dilation and area opening.



(a) (b) (c)

(d) (e) (f)

Figure 3.7 The examples of the variation in the images due to morphological operations: (a) binary image; (b) area opening image with size 10 pixels; (c) dilation by SE 'Disk' shape r=5; (d) area opening image with size 200 pixels; (e) erosion by SE 'Disk' shape r=5; and (f) area opening image with size 300 pixels.

Area opening, or morphological binary open, is often referred to as size filter, and filters out foreground objects from a binary image that are smaller in area than the size provided (λ) in the Matlab bwareaopen function, i.e. bwareaopen (*binary_image,size*). It is an essential function to remove noises or artefacts from the binary image. If an image A is applied to area opening with connected component area size λ, $\lambda \in \mathbb{N}$, it is written in Eq (3.3) as follows:

$$AreaOpening(A, \lambda) = \{A - a | Area\big(C_a(A)\big) < \lambda,\ a \in A\} \qquad (3.3)$$

Where $Area(\cdot)$ is the union of the connected pixels wherein each pixel has intensity level more than zero, whereas $C_a(\cdot)$ is the connected pixels of object in image A.

The morphological closing and opening operations, derived from the dilation and erosion, are described as follows:

Closing is a dilation operation followed by an erosion operation. After applying the image with this operation, the objects in the image tend to remain their original size. The closing operation is useful to clean up images with object holes and other small particles or artefacts. If an image, A, is closing by using structuring element S, it is written in Eq (3.4) as follows:

$$A \bullet S = ((A \oplus S) \ominus S) \tag{3.4}$$

Opening is an erosion operation followed by dilation operation. After employing the image with this operation, the objects in the image also tend to remain their original size. The opening operation is useful to clean up images with noise and other small particles or artefacts. If an image, A, is opening by using structuring element S, it is written in Eq (3.5) as follows:

$$A \circ S = ((A \ominus S) \oplus S) \tag{3.5}$$

Reconstruction is a very useful image morphological operator used to extract meaningful information about objects, i.e. blood cells, in an image (Gonzalez, Woods, & Eddins, 2004). It can also be used to extract the marked objects, i.e. foreground objects, in the grayscale image of this research study. In addition to the process of the reconstruction operator, it makes iterating grayscale dilation of a marker image, J, in i times until the contour or edge of the marker image has stability under a mask image, I. It is written in Eq (3.6) as follows (Vincent, 1993):

$$Reconstruction\ (I, J) = R_I^{\delta}(J) = \vee_{i \geq 1} \delta_I^{(i)}(J) \tag{3.6}$$

where $\delta_I^{(1)}(J) = (J \oplus S) \wedge I$, $J \oplus S$ is the grayscale dilation of image J by structuring element S and $\wedge$ is the pointwise minimum.

For example, we employ the reconstruction operator to the grayscale image I as mask using the grayscale image J as marker. As a result, the new reconstructed image I has the maximum intensity levels extracted from the connected components of the mask image, I, which are marked by the marker image, J, as shown by its transformation in Figure 3.8.

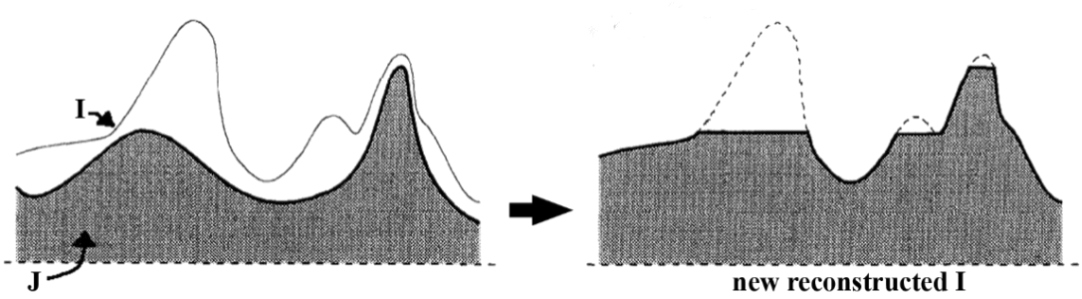
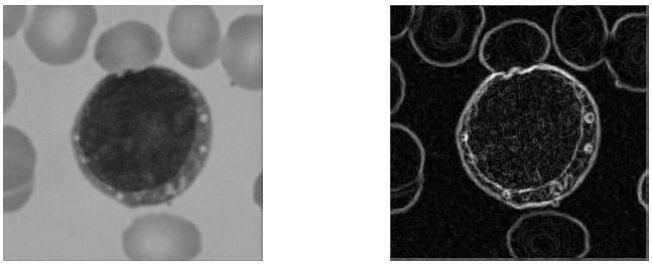


Figure 3.8 The grayscale reconstruction of the mask image, I, from the marker image, J (Vincent, 1993).

In creating markers for the proposed method, first, we compute the gradient magnitude of the grayscale image from the previous step. The gradient magnitude is used to mark the shape of objects in the grayscale image, which will be segmented, as high intensity at the borders or edges of the objects and low intensity inside the objects, as shown in Figure 3.9. After computing the gradient magnitude, we can see the contrast variation clearly between the objects' borders and the background; also, it appears many catchment basins have appeared in the image. However, we could not apply the watershed transform to segment the objects using purely the gradient magnitude image, owing to the image having considerable variation contrast, i.e. local minima, in the image. If we applied the watershed transform to it, over-segmentation would occur. Hence, the foreground and background markers are created in the following steps and will be combined with the result of gradient magnitude image as good seeds before being employed with the watershed transform. Next, the morphological operations are employed to create the foreground marker.

The foreground markers are the connected blobs of pixels inside each object of the foreground image, i.e. the blood cells, as local minimum area of each basin. This research study uses morphological operations to obtain the foreground markers for the watershed algorithm. Two sequence steps, erosion followed by reconstruction and dilation followed by reconstruction, are employed to make the grayscale image cleaned, particularly, the foreground objects, a smooth and fine intensity covering the object areas and also the image background, as shown in Figure 3.10 (a). Moreover, at this point, the edges or boundaries of the foreground marker blobs are cleaned and shrunk by the sequence of the morphological operations without changing the overall shape of the cells.



(a) (b)

Figure 3.9 (a) the input grayscale image before and (b) after computing gradient magnitude, respectively.

Next, we compute the regional maxima of the previous result, from the applied two sequence steps, to obtain the foreground object blobs as good foreground markers. As such, the results from applied regional maxima are the black foreground object blobs in the white background, as shown in Figure 3.10 (b). The important thing is that the boundaries of the foreground objects are preserved, especially the large object in the centre of image, which is the targeted lymphocytic cell membrane. Moreover, at this point, we can see the minima of each basin clearly. To complete creating the foreground markers, we clean up the good foreground object blobs from the previous step using a morphological closing and then remove the isolated small pixels or artefacts, usually present in the image, using an area opening operation, which is the bwareaopen function in Matlab, as shown in Figure 3.10 (c). Now, the binary foreground markers, as good seeds, are completed, as shown in Figure 3.10 (d).

The background markers in this research study are created to incorporate with foreground markers and gradient magnitude image, controlling the watershed transform to achieve the segmentation results more accurately. By observation and trial and error in our experiment, the a priori knowledge from the grayscale image is the nucleus of the lymphocytic cell membrane, which has darkest intensity in the cleaned grayscale image, from the resultant after applying two sequence steps in creating foreground markers, as shown in Figure 3.11 (a).

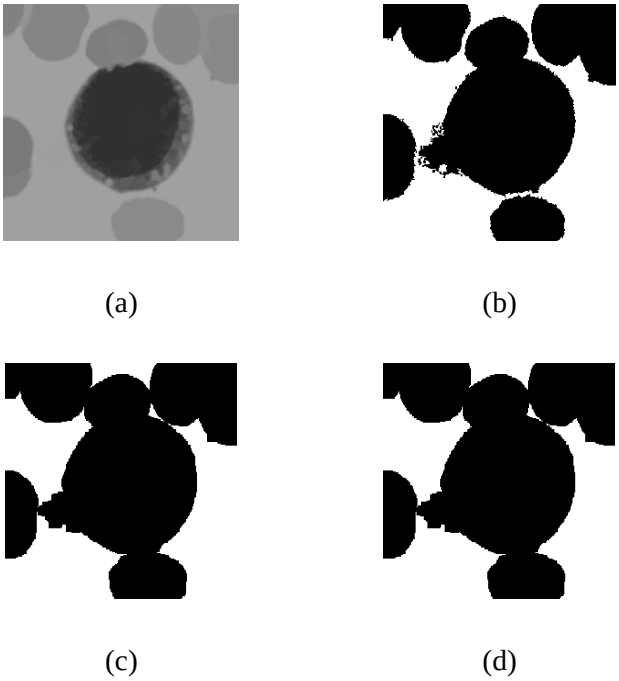
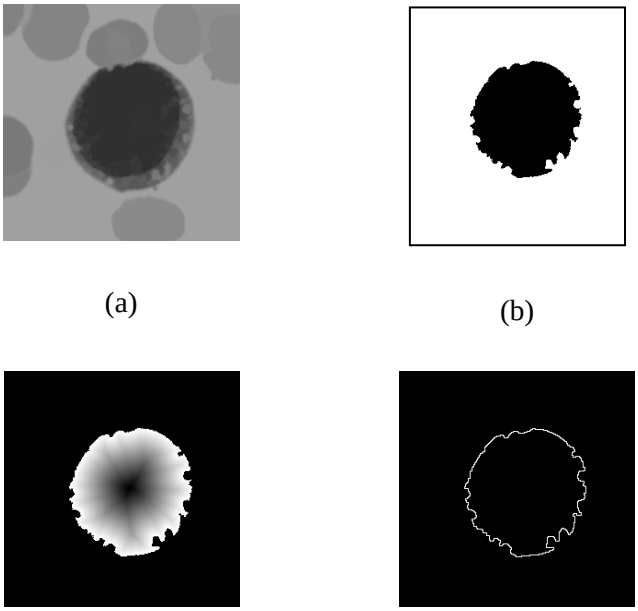


Figure 3.10 (a) result after applied two sequence steps; (b) after computed regional maxima; (c) after applied closing and area opening; and (d) the foreground markers.

However, the nucleus of the lymphocytic cell in the grayscale image is not an appropriate seed for the watershed transform to achieve the accurate segment result of the blood cell membrane, including nucleus and cytoplasm. Therefore, this research study uses this a priori knowledge to produce the background markers which have subsequence steps as follows. The cleaned grayscale image, from the resultant after applying two sequence steps in creating foreground markers, is used in its conversion to binary image, namely "*bw*", using global thresholding technique, Otsu's method, as shown in Figure 3.11 (b). Next, we create the drainage region in the area of the objects in the *bw* image as well as creating the catchment basins of the objects in the *bw* image using the distance transform applied to the *bw* image; the result of this step is "*D*". The result of distance transform, *D*, is then assigned to *D1* as *D1= -D*. The *D1* is superimposed with assigned negative infinity values to the pixel areas of *D1* corresponding to the white pixel areas, uninteresting areas, of the *bw* image as *D1(bw)* = -Inf. The result of this earlier operation is shown in Figure 3.11 (c). This step has been modified to give better segmentation results of watershed transform in this study. Then, we employ the watershed transform to the *D1* to obtain the watershed line of the *bw* image. Now, only the watershed line of the *bw* image is used as the background markers, as shown in Figure 3.11 (d).



(a) (b)

(c) (d)

Figure 3.11 (a) result after applying two sequence steps; (b) binary image after applied global thresholding to the grayscale image; (c) result after assigning the negative infinity values to the pixel areas of the distance transform of the *bw* image; (d) the background marker.

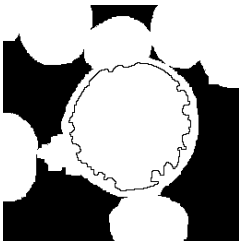


Figure 3.12 The modified gradient magnitude image, as good seeds, for marker-controlled watershed segmentation

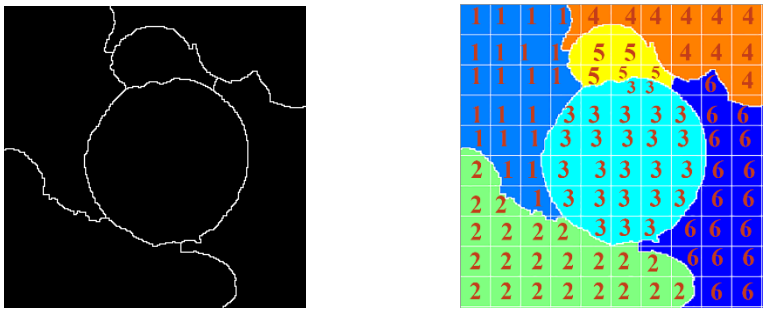
Before segmenting the input grayscale image with the watershed transform, the final process of generating the markers to be good seeds for the watershed algorithm is described as follows. We compute regional minima to the gradient magnitude image by using the combination of two markers, foreground and background, as a marker to the regional minima operation. As a result, the earlier gradient magnitude image has minimum areas, according to both markers, called modified gradient magnitude, as shown in Figure 3.12, which are ready to process with the watershed transform in the next section.

3.4.3. Segmentation with Watershed Transform

The modified gradient magnitude image, which has the markers embedded inside to mark the minima in the gradient magnitude image from the previous step, is ready to become good seeds for the watershed segmentation. In this section, we employ the watershed transform to compute the modified gradient magnitude image. The segmentation result from the watershed transform is kept in the labelled matrix and each region corresponding to the separated regions by the watershed lines, as shown in Figure 3.13 (a), has its own labelled indexing number, starting from the first region, as labelled number 1, to the last region. Figure 3.13 (b) shows the examples of the labelled matrix result of the watershed transform in colour, wherein each colour represents a different segmented region corresponding to its labelled number.

3.4.4. Lymphocytic Cells Membrane Identification and Retrieval

This section presents the identification of the targeted lymphocytic cell membrane from the labelled matrix of the watershed segmentation and the retrieval of the original RGB pixels of the identified lymphocytic membrane and places them on the white background corresponding to the same size, width and height, as the original image.



(a) (b)

Figure 3.13 The result of watershed segmentation: (a) the watershed lines and (b) the various colours represent each segmented region corresponding to the labelled matrix from the watershed segmentation.

In practice, the labelled matrix from the watershed transform is the same as the grayscale image matrix, wherein each pixel is represented with a label number corresponding to the segmented regions, which are assigned label indexing numbers from the watershed algorithm, instead of its gray level value. Moreover, in observation, the microscopic blood sub-images are usually present in the centre of the image as the interesting object in the image acquisition stage. Therefore, we identify the lymphocytic cell membrane by calculating from the centre of the labelled matrix and then finding the most frequently labelled indexing value with the square size of 10x10 pixels from the centre of the labelled matrix to isolate only the labelled indexing value of the identified lymphocytic membrane, as shown in Figure 3.14.

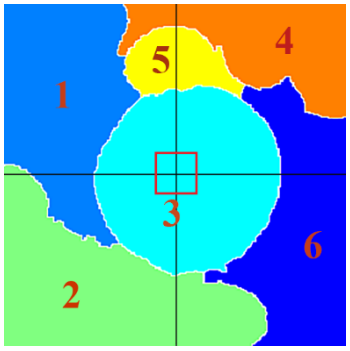


Figure 3.14 Most frequently labelled indexing value with the square size 10x10 pixels (red square box at the middle) is number 3, used to identify and select only labelled region number 3 (the lymphocytic cell membrane) from the labelled indexing matrix.

In order to retrieve the original RGB pixels of the identified lymphocytic cell membrane, a binary mask image of the identified cell membrane is created by setting the pixel to '1', or white colour, which has a labelled indexing number corresponding to the label of the identified cell membrane, otherwise set to '0', or black colour. Finally, the RGB pixels of the selected index value are retrieved on the white background with the same size as the original image. The procedure and the result of this stage are shown in Figure 3.15.

3.5. Evaluation and Discussion

In this section, we employ the 180 lymphocytic sub-images, from the ALL-IDB2 database, for the evaluation of this work. In order to test the performance of the proposed modified marker-controlled watershed segmentation, the experiments were carried out, including the proposed method and the traditional marker-controlled watershed transformation, with and without employing the Gaussian low-pass filter. In addition, the SE for the experiments is 'disk' shaped, as was used in our previous study (Srisukkham et al., 2013). Moreover, all experiments are implemented based on MATLAB software version 7.12 (R2011a) and using a CPU AMD Athlon II 3.0 GHz personal computer with 4 GB memory running on Microsoft Windows 7 Enterprise operating system.

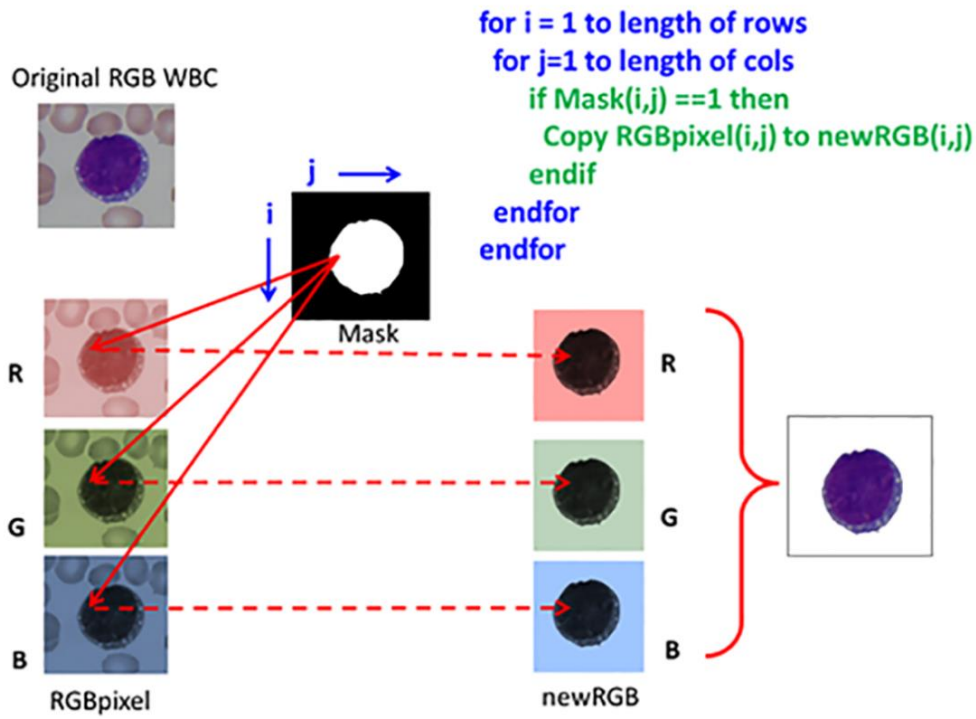


Figure 3.15 The procedures of retrieving the original RGB lymphocytic cell membrane placed on the white background by using binary image (Mask).

As aforementioned in Section 3.3, the ground truth of the 180 sub-images has been found based on the database annotations and in further consultation with the haematologists from the Royal Victoria Infirmary (RVI) Hospital at Newcastle-Upon-Tyne, United Kingdom, and examples are shown in Figure 3.16, column three from the left hand-side. The ground truths and annotations of the lymphocytic membrane manually segmented only the targeted cell membrane, lymphocyte and lymphoblast cells, and placed these on the white background corresponding to the same position and same size of the original images. The reason why we created the ground truth lymphocytic membranes with the size of the image, as well as the original image and with the white background, is that we will further conduct an automated comparison between the ground truth images and the identified lymphocytic membrane, from both the proposed methods and the compared methods, using the two-dimensions correlation coefficient, $Corr$, as the comparison technique depicted in Eq (3.7), without human intervention. The $Corr$ value varies from -1 to 1 which indicates that values closer to 1 have greater conformity with the segmented images of the proposed method, and vice versa.

$$Corr = \frac{\sum_r \sum_s (Y_{rs} - \bar{Y})(T_{rs} - \bar{T})}{\sqrt{(\sum_r \sum_s (Y_{rs} - \bar{Y})^2)(\sum_r \sum_s (T_{rs} - \bar{T})^2)}} \tag{3.7}$$

where r and s refer to the row and column pixels, while $\bar{Y}$ and $\bar{T}$ refer to the mean of matrix elements (pixels) in images Y and T, respectively.

In addition to the comparison using the correlation coefficient, we conduct the automated categorisation of the segmented lymphocytic cell membranes into four groups, including completely segmented: accurate (A), those that can be used for subsequence steps: usable (U), a few pixels missing: partial (P) and no segmentation (N), according to the qualitative manner by human visual inspection. Moreover, the criterion condition of each group is as follows: group A: $Corr \geq 0.9$, group U: $0.8 \geq Corr \geq 0.89$, group P: $0.7 \geq Corr \geq 0.79$, and group N: $Corr < 0.69$.

The experiments indicate that the proposed method, modified marker-controlled watershed segmentation, with and without using the Gaussian low-pass filter, achieves more promising results than the traditional marker-controlled watershed methods in the comparison, as illustrated in Figure 3.16. The proposed method provides better results in terms of the complete or accurate segmentation of the whole cell membrane, cell with nucleus and cytoplasm, of the lymphocytic cell image.

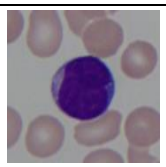
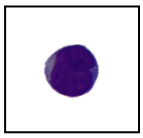
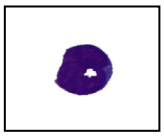
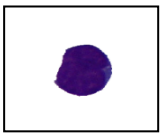
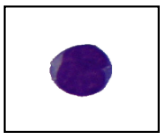
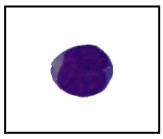
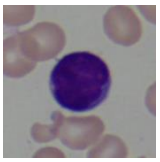
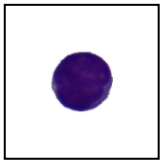
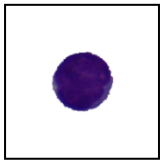
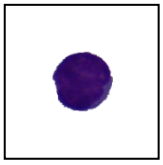
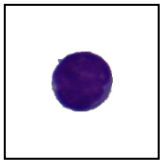
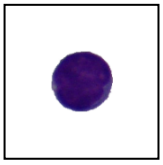
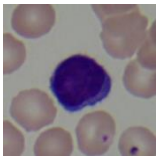
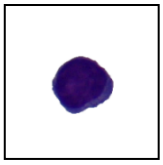
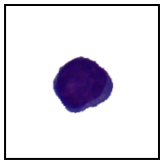
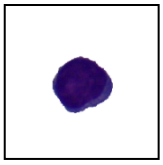
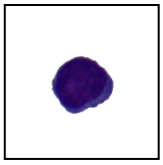
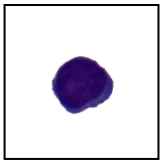
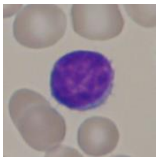
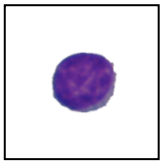
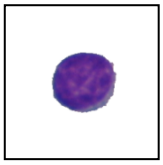
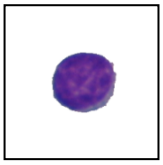
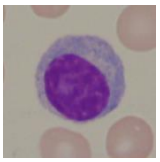
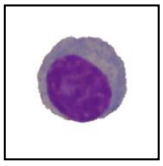
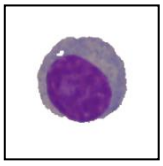
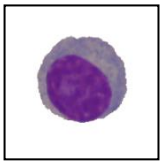
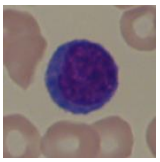
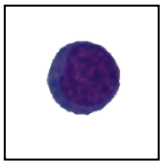
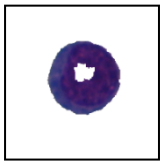
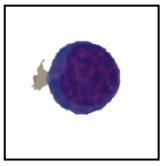
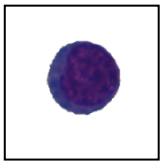
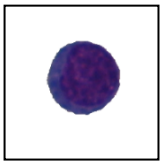
Original RGB lymphocytic cell sub-images with names and annotations	Ground truth lymphocytic membranes	Methods			
		Traditional WT method	Traditional WT method+ filter	Proposed method	Proposed method+filter
 Unhealthy, Im004_1					
 Unhealthy, Im015_1					
 Unhealthy, Im024_1					
 Healthy, Im156_0					
 Healthy, Im196_0					
 Healthy, Im204_0					

Figure 3.16 The comparison of the segmented lymphocytic cell membrane between the proposed method and the traditional marker-controlled watershed segmentation.

In order to validate the segmentation results of the lymphocytic cell membrane, a correlation coefficient, as shown in Eq (3.7), is employed to measure the degree of similarity compared with the manually-segmented lymphocytic cell membrane obtained in consultation with the haematologists. The average correlation coefficient of each compared method is revealed in Table 3.1. From the experimental results, the proposed modified marker-controlled watershed method with Gaussian low-pass filter performs the best, with highest correlation to human segmentation results corresponding to the ground truth lymphocytic cell membranes. Moreover, the proposed method with Gaussian low-pass filter performs better than the proposed method without the filter. This indicates that using the filter in the pre-processing stage before the subsequence steps in segmentation can help the segmentation process to achieve better accuracy in segmentation of the lymphocytic cell membranes. Also, the proposed method with the good seed markers incorporates with the watershed transform to achieve the best segmentation compared to the traditional marker-controlled watershed algorithm.

However, there are some lymphocytic cell images that the proposed modified marker-controlled watershed segmentation performs with inaccurate segmented results, in a usable group, as shown in Figure 3.17, owing to the cytoplasm in the images being very transparent and very close to the background colour of the image. The transparency colour of the cytoplasm of the cell membrane image is still challenging for the white blood cell membrane segmentation algorithm, which needs future research for investigation and improvement for an accurate or complete cell membrane segmentation. Overall, the proposed modified marker-controlled watershed segmentation is able to produce promising segmentation results of the whole lymphocytic cell membrane, including nucleus and cytoplasm, and is of benefit for the further step of an intelligent decision support system for acute lymphoblastic leukaemia detection.

Table 3.1 The correlation coefficient values (Corr) of the segmented cell membranes between the proposed method and the traditional marker-controlled watershed method in comparison to the ground truth manual segmented images for 180 sub-images.

Methods	**Average of *Corr***
Traditional WT method without filter	0.8556
Traditional WT method with filter	0.8753
Proposed method without filter	0.9153
Proposed method with filter	0.9374

Original WBC sub-images	Ground truth lymphocytic membranes	Proposed method + filter
		

Figure 3.17 A sample of lymphocyte cell images with inaccurate segmented result, in a usable group, from the proposed method with Gaussian low-pass filter.

3.6. Chapter Summary

This chapter has presented the proposed algorithm of the modified marker-controlled watershed segmentation for the segmentation of the lymphocytic cell membrane images. The unique contribution of this chapter is a novel combination of existing techniques, i.e. watershed transform and morphological operations, and the proposed method of generating the good markers for watershed transform to segment the lymphocytic cell membranes with promising results using microscopic blood smear sub-images. The proposed method with Gaussian low-pass filter performs segmentation of the lymphocytic cell membrane with highest correlation to the ground truth images compared with the traditional marker-controlled watershed algorithm. Moreover, this chapter has introduced the overall system architecture of this PhD research. The details of the microscopic stained peripheral blood smear image database used in this study are also explained. In addition, the ground truths and annotations of the lymphocytic cell images and clinical diagnosis criteria, according to the consultation with the haematologists, are revealed.

Overall, the proposed modified marker-controlled watershed segmentation is able to produce promising segmentation results of the whole lymphocytic cell membrane, including nucleus and cytoplasm. It is of benefit for the subsequent nucleus-cytoplasm separation using the proposed SDM-based clustering algorithm, which is presented in Chapter 4, for robust ALL detection with high accuracy.

Chapter 4: The Separation of Nucleus and Cytoplasm Using Stimulating Discriminant Measures (SDM)

4.1. Introduction

In this chapter, a novel clustering algorithm with stimulating discriminant measure (SDM) of both within- and between-cluster scatter variances is proposed to produce robust segmentation of nucleus and cytoplasm of lymphocytes/lymphoblasts. Specifically, the proposed between-cluster evaluation is formulated based on the trade-off of several between-cluster measures of well-known feature extraction methods. The SDM measures are used in conjunction with GA for the clustering of nucleus, cytoplasm, and background regions. Overall, the key steps of this study as shown in the green rectangle dashed-line in Figure 4.1 include: SDM-based nucleus-cytoplasm separation, features extraction of the separated nucleus and cytoplasm and the classification of the healthy and unhealthy (blast) lymphocyte cell images. The structure of this chapter is as follows: Section 4.2 explains the reason why the separation of nucleus and cytoplasm of the identified lymphocytic cell membrane images is required. Next, the proposed SDM-based clustering method for the separation of nucleus and cytoplasm of the cell membranes is introduced in Section 4.3. The feature extraction of the separated nucleus and cytoplasm sub-images as a raw feature subset for this research is illustrated in Section 4.4.

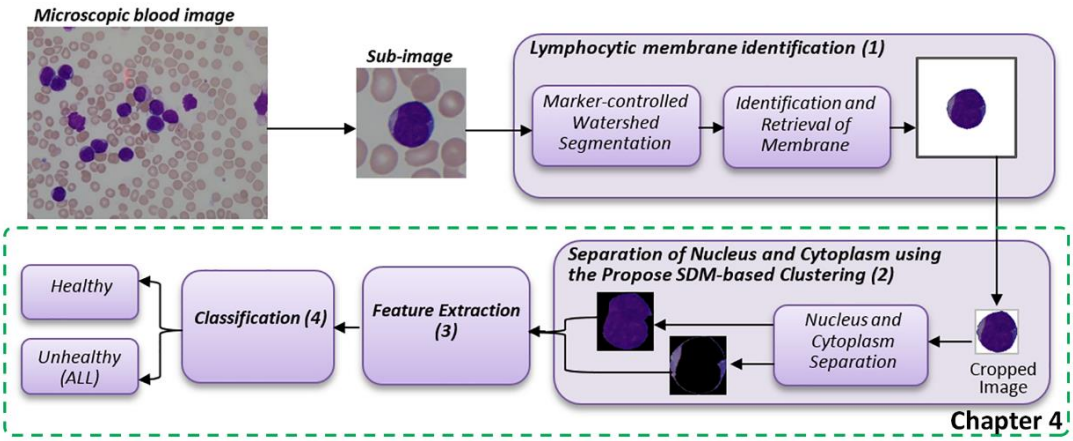


Figure 4.1 The proposed SDM-based clustering algorithm for robust ALL detection.

The ALL detection and classification is explained in Section 4.5. Finally, the evaluation and discussion of the proposed SDM-based clustering method in comparison with the results of the state-of-the-art algorithms in the literature are shown in Section 4.6.

4.2. Why the Separation of Nucleus and Cytoplasm of the Identified Lymphocytic Cell Membrane Images is required?

In the clinical diagnosis of ALL with the morphology of blood smear slides under the light microscopic examination, the components of a white blood cell membrane include nucleus and cytoplasm. Both parts have the information and characteristics and play an important role in the diagnostic abnormality of the lymphocyte cells for a haematologist or a haematopathologist to identify ALL more accurately. Some digital diagnosis systems were developed to analyse microscopic peripheral blood smear images for ALL detection. Conversely, they suffered from a number of limitations, in particular an accurate diagnosis of ALL requires discrimination of one individual cell type from another, and of cell nucleus from cell cytoplasm (Abdul-Hamid, 2011). More particularly, the separation of lymphocytic cell nucleus with diverse complex irregular morphology from cell cytoplasm is a challenging task. In addition, only a few existing clustering methods are able to achieve good adaptively processes for reliable separation of nucleus and cytoplasm (Mohapatra, Patra & Kumar, 2012; Mohapatra et al., 2012; Mohapatra et al., 2014). Therefore, the robustness of the existing approaches is compromised owing to the limitation of the existing clustering algorithms (Kuo & Landgrebe, 2004; Li et al., 2011). In this research study, we aim to overcome the previously mentioned challenges, and to develop an intelligent decision support system for ALL detection using microscopic peripheral blood smear images. Next section presents the SDM-based clustering algorithm, as follows.

4.3. The Separation of Nucleus and Cytoplasm with Stimulating Discriminant Measure (SDM) Technique

In this section, we discuss classical and state-of-the-art clustering algorithms and introduce the proposed SDM-based clustering with the consideration of both within- and between-cluster assessments for the separation of nucleus and cytoplasm in detail.

4.3.1. Clustering, Discriminant Analysis and Their Limitations

Clustering analysis is widely used to assess the hidden patterns of datasets and organise samples into different categories according to the quantitative measurement of distinctiveness (Naz, Majeed & Irshad, 2010). There are two types of clustering: hard and soft clustering. Hard clustering is normally applicable when there are significant differences between clusters

and each object in the dataset belongs to exactly one cluster. K-means is a popular example of hard clustering algorithms that find the centre of each cluster based on the minimization of J_{KM}, the sum of the square of the distances between sample points in each cluster and their centre:

$$J_{KM} = \sum_{j=1}^{C} \sum_{x_i \in Cj} \left\| x_i^{(j)} - c_j \right\|^2, C > 1 \tag{4.1}$$

where c_j indicates the centre of the cluster j where $j=1, 2, C$ and $x_i^{(j)}$ refers to the data point, i, in cluster j. Even though K-means clustering is simple to implement for a large dataset, it is highly sensitive to the initial clustering centres (Nilima, Dhanesh & Anjali, 2013).

In contrast to hard clustering like K-means, FCM is a soft clustering algorithm that assigns a membership to each data sample. The key difference here is that a data sample can belong to multiple clusters, and the minimisation function employed by FCM is as follows:

$$J_{FCM} = \sum_{j=1}^{C} \sum_{i=1}^{N} \left(\mu_{ij}\right)^m \left\| x_i - c_j \right\|^2, C > 1, m \in (1, \infty) \tag{4.2}$$

where

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \tag{4.3}$$

μ_{ij} represents the membership degree of data sample i with respect to cluster j, whereas m is a real value weighting component, which is greater or equals to 1. Notice that μ_{ij} is inversely related to the distance between the data sample and the cluster centre.

Even though the soft partitioning of FCM through μ_{ij} is sometimes more practical for segmenting objects that do not have significant boundaries in an image, FCM is not suitable for non-convex shapes, i.e. noisy data, such as very large and very small values that could skew the mean (Wang, 2010).

Apart from clustering algorithms, data classification techniques such as LDA are generally applied to classify data samples. In LDA, classification is conducted based on two discriminant measures: within-class scatter matrix, SW_{LDA} (Eq (4.4)) and between-class scatter matrix, SB_{LDA} (Eq (4.5)).

$$SW_{LDA} = \sum_{j=1}^{C} \sum_{i=1}^{N_j} \frac{1}{N} \left(x_i^{(j)} - c_j \right) \left(x_i^{(j)} - c_j \right)^T \tag{4.4}$$

$$SB_{LDA} = \sum_{j=1}^{C} \frac{N_j}{N} \left(c_j - c \right) \left(c_j - c \right)^T \tag{4.5}$$

where $c = \frac{1}{N}\sum_{j=1}^{C}\sum_{i=1}^{N_j} x_i^{(j)}$ and N is the number of samples with N_j representing the number of training samples in cluster j.

As explained in Theodoridis and Koutroumbas (2006) and Li et al. (2011), the criterion of J_{FCM} is similar to the trace of the fuzzy within-cluster scatter matrix, SW_{fcm} shown in Eq (4.6).

$$SW_{fcm} = \sum_{j=1}^{C}\sum_{i=1}^{N_j}\left(\mu_{ij}\right)^m\left(x_i - c_j\right)\left(x_i - c_j\right)^T \tag{4.6}$$

The equation of SW_{fcm} is closely related to the within-cluster scatter matrix of LDA shown in Eq (4.4). As a result, FCM is claimed to consider only the within-class similarity measure (Li et al., 2011). In other words, the exclusion of between-class discriminant measure reveals the limitation of conventional FCM. In addition, the same issue applies to K-means clustering in that the between-cluster criterion is not taken into consideration in the discriminant measure.

Motivated by the between-class discriminant measure, FCS was proposed by Wu et al. (2005) to minimise the within-cluster compactness and maximise the between-cluster separation. As explained in Li et al. (2011), fuzzy between-cluster matrix (SB_{FCS}), shown in Eq (4.7), and within-cluster scatter matrix (SW_{FCS}), shown in Eq (4.8), are defined as follows:

$$SB_{FCS} = \sum_{j=1}^{C}\sum_{i=1}^{N}\eta_j\left(\mu_{ij_{FCS}}\right)^m\left(x_i - c\right)\left(x_i - c\right)^T \tag{4.7}$$

$$SW_{FCS} = \sum_{j=1}^{C}\sum_{i=1}^{N}\left(\mu_{ij_{FCS}}\right)^m\left(x_i - c_j\right)\left(x_i - c_j\right)^T \tag{4.8}$$

where j=1,2,…C represents the jth cluster, and $x_i \in X_j$ with X_j as a set of data samples in the jth cluster that consists of N samples. Note that η_j is a weighting parameter as follows:

$$\eta_j = \frac{\left(\beta/_4\right)min_{j'\neq j}\|c_j - c_{j'}\|^2}{max_k\|c_k - c\|^2},\ 0 \leq \beta \leq 1.0 \tag{4.9}$$

With SB_{FCS} and SW_{FCS}, the objective function of FCS (J_{FCS}) or called FCS1 in this research study is defined as the difference between the trace[2] of the matrices SW_{FCS} and SB_{FCS} reported in Li et al. (2011). It is derived as:

$$J_{FCS} = \mathrm{tr}(SW_{FCS}) - \mathrm{tr}(SB_{FCS}) \tag{4.10}$$

$$J_{FCS} = \sum_{j=1}^{C}\sum_{i=1}^{N}\left(\mu_{ij_{FCS}}\right)^m\left\|x_i - c_j\right\|^2 - \sum_{j=1}^{C}\sum_{i=1}^{N}\left(\eta_j\right)\left(\mu_{ij_{FCS}}\right)^m\|x_i - c\|^2 \tag{4.11}$$

[2] The trace of a square matrix is defined as the summation of its diagonal elements.

Additionally, c_j and c indicate the centre of the cluster j and the mutual centre of all clusters, respectively, while $\mu_{ij_{FCS}}$ refers to the membership function of FCS (Li et al., 2011). Furthermore, the equations of c_j, c, and $\mu_{ij_{FCS}}$ are defined in Eq (4.12), Eq (4.13) and Eq (4.14), respectively, as follows:

$$c_j = \frac{\sum_{i=1}^{N}\left(\mu_{ij_{FCS}}\right)^m x_i - \eta_j \sum_{i=1}^{N}\left(\mu_{ij_{FCS}}\right)^m c}{\sum_{i=1}^{N}\left(\mu_{ij_{FCS}}\right)^m - \eta_j \sum_{i=1}^{N}\left(\mu_{ij_{FCS}}\right)^m} \tag{4.12}$$

$$c = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.13}$$

$$\mu_{ij_{FCS}} = \frac{\left(\|x_i - c_j\|^2 - \eta_j \|c_j - c\|^2\right)^{-1/(m-1)}}{\sum_{k=1}^{C}\left(\|x_i - c_k\|^2 - \eta_k \|c_k - c\|^2\right)^{-1/(m-1)}} \tag{4.14}$$

In reference to Wu et al. (2005), fuzzy within-cluster scatter matrix (SW_{FCS_Wu}) fuzzy between-cluster scatter matrix (SB_{FCS_Wu}), developed on the basis of the fuzzy sample mean, a_j, is given as follows:

$$SW_{FCS\_Wu} = \sum_{j=1}^{C}\sum_{i=1}^{n}\left(\mu_{ij_{FCS}}\right)^m \left(x_j - a_j\right)\left(x_j - a_j\right)^T \tag{4.15}$$

$$SB_{FCS\_Wu} = \sum_{j=1}^{C}\sum_{i=1}^{n}\left(\mu_{ij_{FCS}}\right)^m \left(a_j - c\right)\left(a_j - c\right)^T \tag{4.16}$$

where $a_j = \frac{\sum_{i=1}^{N} \mu_{ij_{FCS}}{}^m x_i}{\sum_{i=1}^{N} \mu_{ij_{FCS}}{}^m}$.

Hence, the proposed objective function of Wu et al. (2005),J_{FCS_Wu}, termed FCS2 in this research, is defined as the difference between the trace of the matrices SW_{FCS_Wu} and SB_{FCS_Wu}. It is derived as follows:

$$J_{FCS\_Wu} = \mathrm{tr}\left(SW_{FCS\_Wu}\right) - \mathrm{tr}\left(SB_{FCS\_Wu}\right) \tag{4.17}$$

$$J_{FCS\_Wu} = \sum_{j=1}^{C}\sum_{i=1}^{N}\left(\mu_{ij_{FCS}}\right)^m \left\|x_i - a_j\right\|^2 - \sum_{j=1}^{C}\sum_{i=1}^{N}\left(\eta_j\right)\left(\mu_{ij_{FCS}}\right)^m \left\|a_j - c\right\|^2 \tag{4.18}$$

with some slight modifications on a_j, as follows:

$$a_j = \frac{\sum_{i=1}^{N} \mu_{ij_{FCS}}{}^m x_i - \eta_j \sum_{i=1}^{N} \mu_{ij_{FCS}}{}^m c}{\sum_{i=1}^{N} \mu_{ij_{FCS}}{}^m - \eta_j \sum_{i=1}^{N} \mu_{ij_{FCS}}{}^m} \tag{4.19}$$

From Eq (4.18), when $\eta_j = 0$, J_{FCS_Wu} will be equivalent to J_{FCM}. Conversely, when $\eta_j = 1$, J_{FCS_Wu} will be equivalent to the validity index of Fukuyama-Sugeno index (Fukuyama & Sugeno, 1989).

Although between-cluster variations have been embedded into FCS, it is essential to note that the membership function, $\mu_{ij_{FCS}}$ (Eq 4.14), can be negative when $\|x_i - c_j\|^2 \leq \eta_j \|c_j - c\|^2$. A negative membership value poses an issue to determine the ownership of a data sample in a particular cluster. Wu et al. (2005) made a restriction for tackling this issue by proposing $\mu_{ij_{FCS}} = 1$, and $\mu_{ij'_{FCS}} = 0$, for all $j' \neq j$, when a negative value is obtained. The assumption is made such that the data sample belongs to cluster j completely with $\mu_{ij_{FCS}} = 1$, when $\|x_i - c_j\|^2 \leq \eta_j \|c_j - c\|^2$. However, such an assumption may not be always correct because data samples at the boundary of one cluster can easily be misclassified into another cluster, especially when the distribution of data samples along the boundaries of two clusters is close to one another. Figure 4.2 shows an example of such a condition where two clusters are compact, but not well separated. In this case, the distance between the red-coloured point of interest and the centre of cluster 1(c_1), which is indicated by D_2, is smaller than the distance (D_1) between c_1 and the mutual centre of the two clusters (c). According to Wu et al. (2005), this data sample should belong to cluster 1. However, the ground truth indicates that it belongs to cluster 2.

In this research, such condition can be observed during the segmentation of nucleus and cytoplasm of the identified lymphocyte/lymphoblast cell membrane images. When the colour and pixel intensity of the nucleus and cytoplasm are close to each other, as shown in Figure 4.3, the data samples in the boundary of the nucleus cluster get very close to the data samples in the boundary of the cytoplasm cluster. Therefore, the assumption of Wu et al. (2005) can mislead the separation of cytoplasm and nucleus where similar cluster distribution situations, as shown in Figure 4.2, occur.

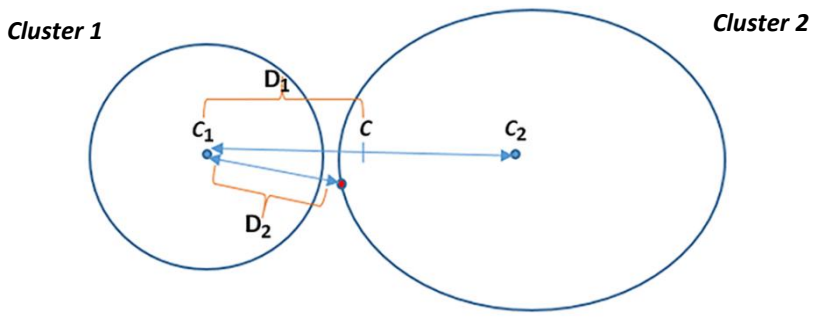


Figure 4.2 Compact, but not well separated clusters (Left: Cluster 1, Right: Cluster 2).

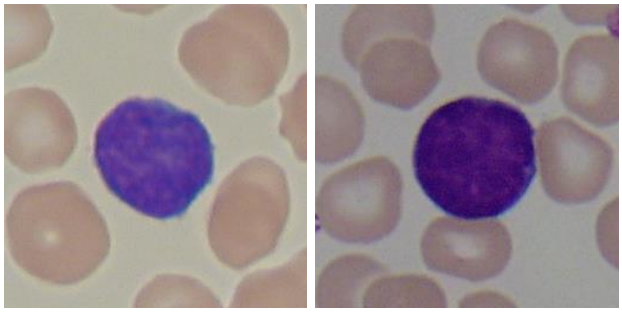


Figure 4.3 Example of lymphocyte sub-images with very similar colour and pixel intensity in both nucleus and cytoplasm.

Thus, FCS proposed by Wu et al. (2005) sometimes has comparatively less robustness and adaptivity for the segmentation of nucleus and cytoplasm with very close cluster scatter measures.

In practice, the distribution of data samples (with very large or very small values) in a cluster affects the position of the cluster centre; therefore, the distances between a data sample to the centre of different clusters may be varied. In addition, a cluster can consist of multiple dimensional data that may lead to larger variances to shift the position of the cluster centre. Furthermore, the setting of parameter η_j in Eq (4.14) plays a significant role in separating the clusters. However, the optimised setting of η_j is subject to different application domains and may be computationally costly to obtain.

In this research, a novel discriminant measure, namely SDM, is proposed to stimulate robust clustering and segmentation of cell nucleus and cell cytoplasm of the identified lymphocyte/lymphoblast cell membrane images. Considering the lack of between-cluster evaluation in FCM, both within- and between-cluster assessments are taken into account for the development of SDM. As an example, an improved clustering process that integrates SDM, FCM and the GA is introduced to obtain the optimum threshold for the separation of nucleus, cytoplasm and the background of each identified cell image. The developed SDM-based clustering is compared with FCM, LDA and FCS. In the next section, we introduce the design of SDM-based clustering.

4.3.2. Stimulating Discriminant Measures (SDM)

In this section, the novel SDM with both within-cluster and between-cluster assessments are introduced. As observed in Eq (4.4), Eq (4.6) and Eq (4.8), i.e. the within-cluster evaluation from LDA, FCM and FCS is dependent on the summation of $(x_i - c_j)$ from all data samples in each cluster. Since the centre of each cluster is calculated based on the mean of all samples

in the cluster, those that are not normally distributed skew the value of the within-cluster evaluation. For example, when a cluster contains ninety-five data samples with very small values of $(x_i - c_j)$ and only five with significantly large values of $(x_i - c_j)$, the use of summation will tend to bias towards smaller within-cluster variations. In fact, the data sample with the largest value of $(x_i - c_j)$ indicates the largest variation from the mean of samples, which indicates that there are no other data samples within the cluster that exceed such a limit. Therefore, in this research study, the argument with the maximum value of $(x_i - c_j)$, q_j, shown in Eq (4.20), is used to indicate the maximum variation per cluster and the total within-cluster scatter matrix, SW_{SDM} shown in Eq (4.21), is defined as follows:

$$q_j = \underset{x_i \in X_j}{\arg\max} \left\| x_i - c_j \right\| \tag{4.20}$$

$$SW_{SDM} = \sum_{j=1}^{C} (q_j - c_j)(q_j - c_j)^T \tag{4.21}$$

where $c_j = \frac{1}{N_j} \sum_{x_i \in X_j} x_i$.

As for the between-cluster evaluation shown in Eq (4.5), Eq (4.7) and Eq (4.16), LDA and FCS take the distance between the centre of a particular cluster and the mutual centre of all clusters, $(c_j - c)$, into consideration. Even though the cluster centres are normally used to give a global view of a specific cluster location with respect to another cluster of interest, the separation between two clusters relies more on the boundary data samples in both clusters. Figure 4.4 shows the relationship of two non-compact clusters.

In Figure 4.4, it is possible that the mutual centre of clusters A and B falls at the location where the two clusters have the larger separation. In this case, the distances $(c_A - c)$ and $(c_B - c)$ do not provide enough information pertaining to the closest separation between these two clusters, which is highlighted in the yellow circle dashed-line. In fact, the closest separation is the most accurate indicator of the separability between the clusters. As a result,

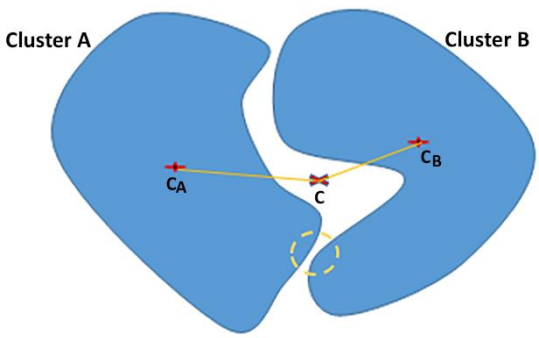


Figure 4.4 The two non-compact clusters.

the boundary of one cluster that is closer to that of the other cluster reveals more information about the separation between both clusters, A and B. In addition, Kuo and Landgrebe (2004) pointed out the importance of using boundary points for evaluating the scatter matrix in their nonparametric weighted feature extraction (NWFE) method. The between-cluster scatter matrix, SB_{NWFE} , defined by Kuo and Landgrebe (2004), is as follows:

$$SB_{NWFE} = \sum_{j=1}^{C} P_j \sum_{l=1}^{C} \sum_{i=1}^{N_j} \frac{\lambda_i^{(j,l)}}{N_j} \cdot \left(x_i^{(j)} - M_l\left(x_i^{(j)}\right)\right)\left(x_i^{(j)} - M_l\left(x_i^{(j)}\right)\right)^T, j \neq l \quad (4.22)$$

where

$$M_l\left(x_i^{(j)}\right) = \sum_{k=1}^{N_l} w_{ik}^{(j,l)} x_k^{(l)} \quad (4.23)$$

$$\lambda_i^{(j,l)} = \frac{\mathrm{dist}\left(x_i^{(j)}, M_l\left(x_i^{(j)}\right)\right)^{-1}}{\sum_{z=1}^{N_j} \mathrm{dist}\left(x_z^{(j)}, M_l\left(x_z^{(j)}\right)\right)^{-1}} \quad (4.24)$$

$$w_{ik}^{(j,l)} = \frac{\mathrm{dist}\left(x_i^{(j)}, x_k^{(l)}\right)^{-1}}{\sum_{q=1}^{N_l} \mathrm{dist}\left(x_q^{(j)}, x_k^{(l)}\right)^{-1}} \quad (4.25)$$

Let *j* and *l* indicate different clusters. N_j is data sample size of cluster *j*, where P_j is the prior probability of cluster *j*. The $w_{ik}^{(j,l)}$ (ranging from 0 to 1), is a weight value given according to the euclidean distance between data samples $x_i^{(j)}$ and $x_k^{(l)}$ where the smaller the distance, the closer is the weight to 1. The summation of the multiplication of $w_{ik}^{(j,l)}$ and $x_k^{(l)}$ gives the weighted mean of $x_i^{(j)}$ in cluster *j*, which is denoted as $M_l\left(x_i^{(j)}\right)$. Then, based on the euclidean distance between the weighted mean and the data samples in cluster *j*, a second weighting function, $\lambda_i^{(j,l)}$ (ranging from 0 to 1), is introduced to the data sample, $x_i^{(j)}$ in cluster *j*. In this case (Eq (4.24)), the smaller the distance to the weighted mean, the closer the value of $\lambda_i^{(j,l)}$ to 1, which indicates the data sample $x_i^{(j)}$ is closer to the boundary of cluster *l*. As a result, SB_{NWFE} emphasises the cluster boundaries rather than the mutual centre for the evaluation of cluster separation based on complicated point-to-point distance weighting assignment.

Although SB_{NWFE} shows a significantly more thorough measurement of cluster separation based on the weighted mean distance, the requirement for twice weighting assignments ($w_{ik}^{(j,l)}$ and $\lambda_i^{(j,l)}$) and the necessity to check distance from each data sample of cluster *j* to each data sample of cluster *l* can be computationally heavy when a large size of data samples is involved in both clusters. For example, if thousands of pixels in an identified lymphocyte/lymphoblast

image were to be represented as the data samples during the between-cluster evaluation of hundreds of possible separations of nucleus and cytoplasm, the computational complexity is significantly high.

Motivated by the boundary separation of SB_{NWFE} and considering the necessity to reduce the computational complexity, a new between-cluster scatter matrix is defined for SDM in this research study. If there are R clusters, two clusters out of R are evaluated at a time for the separation between clusters. Therefore, the number of possible permutations, $Perm$, from R clusters is :

$$Perm = P_{R,2} = \frac{R!}{(R-2)!} \tag{4.26}$$

By taking two clusters, j and l, at a time, let

$$Com_{j,l} = \min\left(\text{dist}\left(c_l, x_i^{(j)}\right)\right), \forall i \in \{1,2,\dots,N_j\} \tag{4.27}$$

and

$$Com_{l,j} = \min\left(\text{dist}\left(c_j, x_k^{(l)}\right)\right), \forall k \in \{1,2,\dots,N_l\} \tag{4.28}$$

where $Com_{j,l}, Com_{l,j} \in Perm, j \neq l$, then

$$SB_{SDM} = \sum_{z=1}^{Perm} \min\left(Com_{j,l}\,,\, Com_{l,j}\right)\,, \forall z \in \{1,2,\dots,Perm\} \tag{4.29}$$

Figure 4.5 depicts the relationship of data samples in cluster j to the centre of cluster l and vice versa. In this figure, the minimum distance between the data sample in cluster j and the centre of the other cluster (e.g. cluster l) is used to estimate the nearest point of the respective cluster to the centre of the other cluster. In this way, both pairs of minimum distances ($Com_{j,l}$ and $Com_{l,j}$) are compared to obtain the closest possible distance between two clusters. Additionally, this process is repeated for $Perm$ times depending on the number of cluster combinations. Although the boundary is not uniformly separated, the minimum distance obtained indicates that there are no other segments of the boundary that have a narrower separation based on the estimation towards the centre of the other cluster. The proposed SB_{SDM} measure avoids tedious point-to-point distance calculation between clusters in SB_{NWFE} that can exponentially increase the computational complexity during the segmentation. It provides a closer estimation pertaining to the cluster separation than the conventional between-cluster evaluation, which is purely based on the distance of the cluster centre towards the mutual centre of all clusters ($c_j - c$), as shown in FCS (Wu et al., 2005) and LDA.

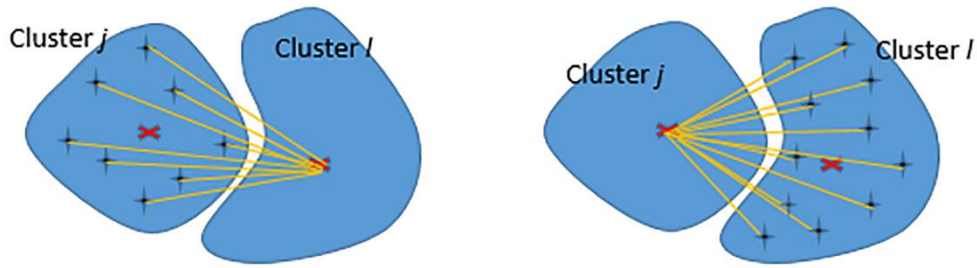


Figure 4.5 The distance of data samples (the yellow lines) to the centre of the other cluster, the red ✕ symbols are centre of each cluster, and the ✦ symbols represents data samples.

4.3.3. SDM-based Clustering for the Segmentation of Nucleus and Cytoplasm of Lymphocyte and Lymphoblast Cell Image

In this research, SDM is embedded into the GA to improve the FCM approach in separating the nucleus and cytoplasm from the identified lymphocyte/lymphoblast images obtained from ALL-IDB2. From the literature review in Chapter 2, Section 2.5, different researchers have applied different colour spaces for the segmentation of nucleus and cytoplasm of lymphocyte and lymphoblast cell images. For example, Putzu et al. (2014) used the combination of the green component of RGB colour space and a* component of CIELAB colour space for the selection of nucleus and cytoplasm via threshold operation. Mohapatra et al. (2014) proposed the use of a* and b* components of CIELAB colour space to segment nucleus and cytoplasm using shadow C-means (SCM), whereas Madhloom et al. (2012b) claimed that the S-component of the HSV made the nucleus of lymphoblasts become the brightest objects during segmentation.

In this study, the clustering algorithm is performed on the L^* component of CIELAB colour space, because the L^* component is able to show more differences between nucleus and cytoplasm whereby nucleus is normally darker owing to the existence of chromatin, whereas cytoplasm is relatively brighter. Although the luminance across images varies, the luminance in a particular image during clustering creates a difference between nucleus and cytoplasm.

The proposed SDM-based clustering embedded into the GA for segmentation of nucleus, cytoplasm and background of lymphocytic cell image is illustrated in Algorithm 4.1. It aims to improve the segmentation capability of conventional FCM. The algorithm starts with a random initialisation of population, P, which consists of chromosomes, S_i, where $i=1,2,\ldots k$, that represents the threshold value of three clusters: nucleus, cytoplasm and the background.

During the initialisation, one of the chromosomes, S_m, is obtained as a seed from the converged solution of FCM to accelerate the process of optimisation, where $S_m \in P$. By referring to the threshold value represented by each chromosome, all pixels in the original image are grouped into three clusters, i.e, A, B, and C, which represent clusters of cytoplasm, nucleus and the background, respectively. In this case, each pixel represents a data sample in a cluster, and a pixel can only belong to one cluster at a time. After separating the pixels, the next step of this algorithm is chromosome evaluation, whereby the chromosome fitness function, $F(S_i)$, is obtained based on SB_{SDM} and SW_{SDM}, defined as follows:

$$F(S_i) = \begin{cases} \frac{SW_{SDM}}{SB_{SDM}} + \alpha, \text{If constraints not satisfied} \\ \frac{SW_{SDM}}{SB_{SDM}}, \text{Otherwise} \end{cases} \tag{4.30}$$

Algorithm 4.1: The SDM-based clustering embedded into the GA for segmentation of nucleus, cytoplasm and background of lymphocytic cell image.

```
Input:
    I_in : Cropped identified lymphocyte/lymphoblast sub-image;
Output:
    The best proposed clustering threshold for separating nucleus, cytoplasm and
    background
Begin
  //initialisation
      I_L = a luminance component of CIELAB colour space of input image I_in;
    //initialisation seeds from the converged solution of conventional
    //FCM, where ~ are ignored return data from the FCM and P(S_1, S_2, ..., S_k).
      (~, P, ~) = FCM(I_L, 3);
    For each S_i ∈ P
      Divide pixels into three clusters according to S_i;
      Evaluate within-cluster and between-cluster variation using SDM;
      //Determine fitness, F(S_i).
    End for
    While maximum iteration not reached
      Rank P;
      SE = select (P); //Select chromosomes based on stochastic universal sampling
      OF = crossover (SE); // Crossover operation to generate offspring
      OF = mutation (OF); //Mutation operation to generate offspring
      Evaluate each S_i ∈ OF; //Determine fitness, F(S_i), for each new S_i in OF
      P = merge (P, OF);
    End while
    Return new updated P;
End.
```

We aim to obtain smaller SW_{SDM} and larger SB_{SDM}, which indicate a higher degree of similarity for within-cluster evaluation and larger separation between clusters, respectively.

Given the aim, the smaller the fitness of the chromosome, $F(S_i)$, the better the solution is. As mentioned previously and demonstrated in Figure 4.3, there are cases where the pixel intensities of nucleus and cytoplasm get very close to each other, therefore, implying a greater degree of difficulty to separate both clusters. In this situation, two constraints are used to assist the segmentation process, as follows: (i) the nucleus/cytoplasm area should not be less than 10% of the corresponding cytoplasm/nucleus area; (ii) the background area should not be larger than the area of the whole membrane (nucleus area + cytoplasm area). If the constraints are not satisfied, a penalty value, α, is applied to increase the fitness of the chromosome, $F(S_i)$.

After evaluating $F(S_i)$, the chromosomes are ranked according to their fitness. Then, a stochastic universal sampling technique is used to avoid bias during the selection of chromosomes for reproduction. The smaller the value of $F(S_i)$, the higher is the chance to be selected. In this study, single-point crossover and mutation are used as the genetic operators to produce new offspring with the probability of 0.7 and 0.3 respectively. The newly-generated offspring are used to divide the pixels into separated clusters (i.e. nucleus, cytoplasm and background). Based on the division of data samples, each offspring is further evaluated with the fitness function, $F(S_i)$, as shown in Eq (4.30). Then, with a generation gap of 0.9, offspring and parent solutions are merged into the new generation. Based on several trials, the GA is able to converge to a good separation between nucleus and cytoplasm when the maximum number of generation is set to 100. Therefore, the processes of evaluation, crossover and mutation are repeated until the maximum number of generations (i.e. 100) is achieved.

This research employs SDM as the objective function to guide the search towards a better segmentation performance. In order to evaluate the discriminant capability of SDM, the segmentation results are compared with those obtained using LDA, FCM and FCS. For a fair comparison, LDA-based clustering and SDM-based clustering use the same GA parameter settings. The only difference in LDA is that SW_{LDA} and SB_{LDA}, as shown in Eq (4.4) and Eq (4.5), are used instead of SW_{SDM} and SB_{SDM}. Two types of FCS evaluations are implemented in this research study: (i) FCS1 based on Li et al. (2011) according to Eq (4.7) and Eq (4.8); (ii) FCS2 based on Wu et al. (2005) according to Eq (4.15) and Eq (4.16). In order to avoid negative membership values for FCS, as mentioned in Section 4.3.1, Eq (4.14) is modified according to the traditional FCM membership calculation shown in Eq (4.3). This ensures the

range of membership values lies within [0, 1]. The revised membership function of FCS is defined as follows:

$$\mu_{ij_{FCS}} = \frac{\left(\|x_i - c_j\| - \eta_j \|c_j - c\|\right)^{-2/(m-1)}}{\sum_{k=1}^{C} \left(\|x_i - c_k\| - \eta_k \|c_k - c\|\right)^{-2/(m-1)}} \tag{4.31}$$

On the subject of η_j, due to the large variation of tuning, the setting of $\eta_j = \frac{1}{(C(C-1))}$ in Yin et al. (2006) is adopted, where C is the number of clusters (Li et al., 2011).

Overall, this study compares SDM-based clustering with LDA-based clustering, FCM, FCS1 and FCS2 qualitatively and quantitatively. Qualitative comparison is based on visual inspection of the segmented nucleus and cytoplasm, whereas quantitative evaluation is based on a 2-dimensional correlation coefficient between automatic segmentation and ideal segmentation from manual cropping in consultation with haematologists, as shown in Section 4.6.1. Eq (4.32) depicts the formula of the correlation coefficient, $Corr$.

$$Corr = \frac{\sum_r \sum_s (Y_{rs} - \bar{Y})(T_{rs} - \bar{T})}{\sqrt{(\sum_r \sum_s (Y_{rs} - \bar{Y})^2)(\sum_r \sum_s (T_{rs} - \bar{T})^2)}} \tag{4.32}$$

where *r* and *s* refer to the row and column pixels, while $\bar{Y}$ and $\bar{T}$ refer to the mean of matrix elements (pixels) in images *Y* and *T*, respectively.

The separation results of nucleus, cytoplasm and the background are discussed in Section 4.6. Empirical results indicate that SDM-based clustering outperforms other algorithms in terms of nucleus and cytoplasm selection. It is observed that there are high numbers of mis-clustered pixels in the segmented images when the existing clustering algorithms are applied. The proposed SDM-based method, however, only shows very small numbers of mis-clustered pixels (i.e. the so-called “salt and pepper” conditions) in the segmented regions; for example, the pixels belonging to nucleus have been clustered into cytoplasm or vice versa. Such “salt and pepper” conditions can easily be solved by further conducting simple morphological operations focusing on nucleus/cytoplasm to identify small hole areas in nucleus/cytoplasm, fill the holes, and remove the filled pixels from the corresponding cytoplasm/nucleus cluster. Matlab functions “imfill” and “bwareaopen” are used for these morphological operations. The results of SDM-based clustering, with and without morphological improvement, are compared and discussed in Section 4.6. After the identified lymphocytic images have separated nucleus and cytoplasm by the proposed SDM-based clustering technique, the features extraction of

both nucleus and cytoplasm images of each identified lymphocyte and lymphoblast image is processed. The detail of the extracted feature sets is explained in the next section.

4.4. Feature Extraction from the Separated Nucleus and Cytoplasm Images

According to Meer et al. (2007), cell size, amount and colour of cytoplasm, shape and chromatin structure are important to characterise lymphocytes. Also, the consultation with haematologists at the Royal Victoria Infirmary (RVI Hospital at Newcastle-Upon-Tyne, United Kingdom) about the criteria for diagnosis of ALL in terms of clinical diagnosis and haematologist experiences was probed. We found that most of the consultation information is similar to the clinical diagnosis of acute lymphoblastic leukaemia, as described in Chapter 2, Section 2.2. Then, we incorporate the consultation information with the descriptors of cell image analysis for the ALL detection from state-of-the-art researches (as mentioned in Chapter 2, Section 2.6) as crucial information to form a set of descriptors for this research study. In order to differentiate normal and abnormal lymphocyte cells, 80 features that comprise 16 shape, 54 texture, and 10 colour-based descriptors are extracted from the segmented nucleus and cytoplasm. The 16 shape-based descriptors are: cytoplasm area, nucleus area, nucleus to cytoplasm ratio, length to diameter ratio, major axis length, orientation, filled area, perimeter, solidity, eccentricity, minor axis length, convex area, form factor, compactness1 based on Mohapatra et al. (2014), compactness2 based on Mohapatra et al. (2010) and roundness of nucleus region. These features mainly aim to extract information on the cell size, nucleus size, nucleus shape and amount of cytoplasm. As for the 54 texture-based features, 13 descriptors from the GLCM matrix, including correlation, sum of variance, normalised inverse difference moment, sum of average, contrast, difference variance, entropy, cluster prominence, cluster shade, dissimilarity, energy, homogeneity, and normalised inverse difference are computed in four different angles (i.e. 0, 45, 90 and 135). In addition to the GLCM features, skewness and kurtosis are included in the texture-based descriptors. Chromatin pattern and the existence of nucleoli and vacuole will change the textural information in GLCM. Therefore, these texture descriptors are used to distinguish normal and unhealthy lymphocyte cells. Finally, 10 colour-based features that involve mean and standard deviations of the a* and b* components of the CIELAB colour space are evaluated for both nucleus and cytoplasm, with two descriptors referred to the ratio of the mean of a* and b* components between cytoplasm and nucleus. The summary of all features used in this research study is shown in Table 4.1.

Table 4.1 Summary of all 80 features in this research.

Groups of feature	Number of features	Feature details
Shape-based features	16	Cytoplasm: cytoplasm area Nucleus: nucleus area, ratio of nucleus to cytoplasm, major axis length, minor axis length, ratio of length (major axis length) to diameter (minor axis length), orientation, filled area, perimeter, solidity, eccentricity, convex area, form factor, compactness1, compactness2 and roundness of nucleus region
Texture-based features	54	Texture from GLCM matrix angle 0 degree: correlation, sum of variance, normalised inverse difference moment, sum of average, contrast, different variance, entropy, cluster prominence, cluster shade, dissimilarity, energy, homogeneity, normalised inverse difference, skewness and kurtosis Texture from GLCM matrix angle 45, 90, 135 degrees: correlation, sum of variance, normalised inverse difference moment, sum of average, contrast, different variance, entropy, cluster prominence, cluster shade, dissimilarity, energy, homogeneity and normalised inverse difference
Colour-based features	10	convert images to CIELAB or CIE L*a*b* colour space Cytoplasm: mean of a* and b*components, standard deviation of a* and b* components and ratio of mean a* to mean b* components Nucleus: mean of a* and b*components, standard deviation of a* and b* components and ratio of mean a* to mean b* components

4.5. ALL Detection and Classification

4.5.1. Feature Dataset for Training and Testing All Classifiers

The dataset for this research study includes 180 lymphocytic sub-images, which are 60 lymphocyte (normal or healthy) and 120 lymphoblast (abnormal or unhealthy) images. We categorise a dataset for training and testing samples randomly with a balanced number of

lymphocytes and lymphoblasts in both training and testing images. Therefore, this dataset, which is used for all experiments in this chapter, includes 90 training (i.e. 30 healthy and 60 unhealthy lymphocytes) and 90 testing (i.e. 30 healthy and 60 unhealthy lymphocytes) sample images.

4.5.2. Finding the Optimal Configuration Parameters for Classifiers

In this research study, we employ a number of classifiers, i.e. MLP, SVM, and ensembles with diverse weighting combination methods, for classifying normal and abnormal lymphocyte cells. Before classification, the 80 features, comprising texture, colour and shape-based information, as mentioned in Section 4.4, are scaled into the range of [-1, 1]. These scaled features are then used as the inputs of each classifier for recognising normal and abnormal lymphocyte cells. Moreover, all experiments are implemented based on MATLAB software versions 8.1 (R2013a) and using CPU Intel Core i7 3.6 GHz personal computer with memory 16 GB running on Microsoft Windows 7 Enterprise operating system.

For the MLP classifier, we first conduct a test to find the optimal network topology in order to achieve a good classification rate. Input data normalisation is also performed to avoid the dominance of large input values to the learning process. A logarithmic sigmoid transfer function is used as the activation function for the hidden layer, while a linear transfer function is used for the output layer. The Lavenberg-Marquardt algorithm is also used to train the MLP. We are setting the variation of hidden nodes for each hidden layer of the MLP models as follows: the model of one hidden layer, the number of hidden nodes, is ranged in 2 to 70 nodes; the model of two hidden layers, the number of hidden nodes of each hidden layer, is ranged in 2 to 50; the model of three hidden layers, the number of hidden nodes for each hidden layers, is ranged in 2 to 10. The model of MLP for this research study is shown in Figure 4.6.

For the SVM classifier, the RBF kernel is used, since it supports non-linear mapping of data samples and possesses fewer hyper-parameters (Chih-Wei Hsu, Chih-Chung Chang, 2008). In order to achieve a good setting of the RBF kernel, the scaling factor, γ, and the soft margin constant, *Co*, are determined using the grid search method (Chih-Wei Hsu, Chih-Chung Chang, 2008). By using exponentially growing sequences, the ranges from 2^{-5} to 2^{15} and 2^{-10} to 2^{5} are searched for Co and γ, respectively.

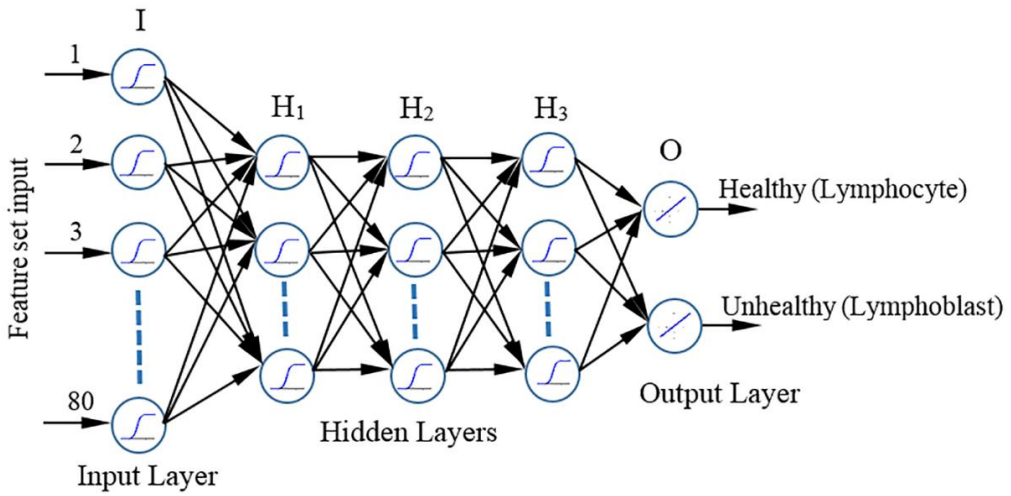


Figure 4.6 The MLP model to find the optimal network topology for one hidden layer, two hidden layers and three hidden layers.

In addition to single MLP and single SVM classifiers, the ensemble classifiers are implemented with the aim to improve classification accuracy. In this research study, a series of ensembles with nine weighting strategies are employed, i.e. majority voting, minimum and maximum probability, distribution summation, average of probabilities, product of probabilities, Bayesian combination, decision templates and Dempster-Shafer (Kuncheva, 2004, 2014; Polikar, 2006; Rokach, 2010). To make a feasible comparison study, all these weighting strategies are implemented using the same number of base classifiers with the same setting for each base model, as shown in Figure 4.7.

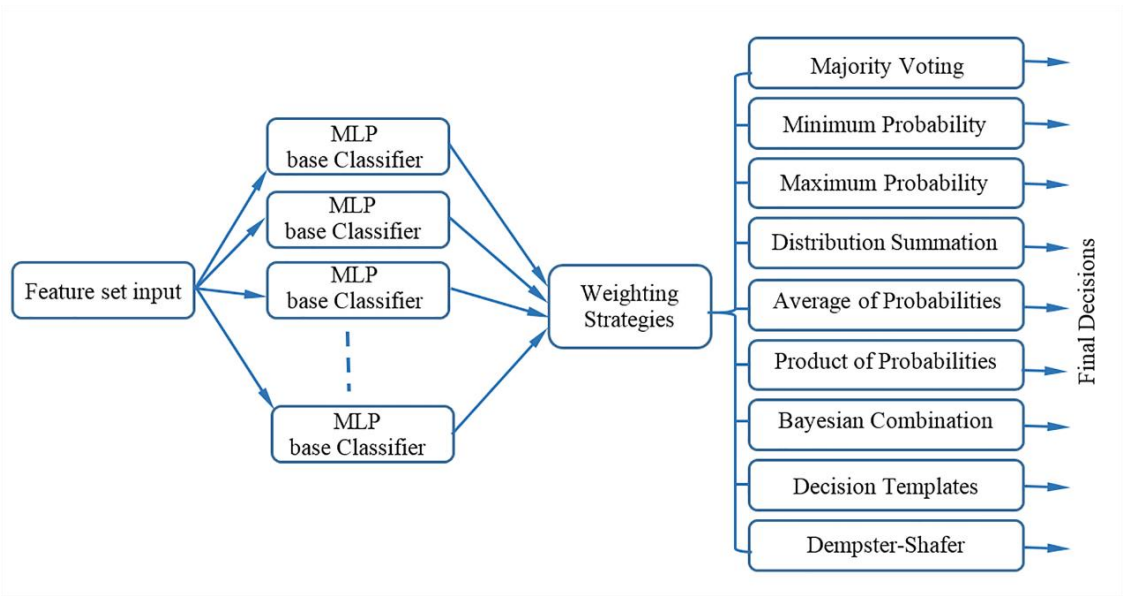


Figure 4.7 The ensemble model built with MLP base classifiers.

Empirical results indicate that the best accuracy is achieved by Dempster-Shafer, followed by majority voting. Therefore, the results from the Dempster-Shafer ensemble and two single classifiers (i.e. MLP and SVM) are presented and discussed in Section 4.6.

Two case studies are conducted in this chapter as follows:

(i) The 80 lymphocytic images for comparison with the work of Khashman and Abbas (2013).

(ii) The 180 lymphocytic images for the overall performance evaluation. The best setting of each classifier for different case studies is given below.

In the first case study, three evaluation schemes, comprising different training and testing data ratios, i.e. 75%:25%, 50%:50% and 25%:75%, are used for evaluating a total of 80 lymphocytic sub-images extracted from ALL-IDB2, respectively.

The MLP has the following settings, i.e. two hidden layers, each layer with 8 and 43 hidden nodes for the first and second evaluation schemes; and one hidden layer with 13 hidden nodes for the third evaluation strategy.

As for the SVM, the best parameter settings of tuple values (i.e. scaling factor (γ), soft margin constant (Co)) obtained from grid search are (8, 0.5), (8, 4) and (16, 32) and, for Dempster-Shafer ensemble, there are 10, 11 and 10 MLP base models employed for the first, second and third schemes, respectively. In particular, such ensembles are constructed based on the best trade-off between computational complexity and system performance.

In the second case study, two types of validation methods are used: (i) 10-fold cross validation and (ii) 500 bootstrap sampling validation. We employ 10-fold cross validation for evaluating 180 images segmented using the proposed SDM clustering and morphological operations with SVM for ALL classification. In the experiments of this research study, 90 images are used for training with the remaining independent 90 images for testing, as mentioned previously.

The settings of SVM are tuned by conducting grid search based on 10-fold cross validation purely on the training set of 90 images. The optimum values of the scaling factor, γ, and the soft margin constant, Co, are identified, respectively, as $\gamma = 8$, $Co = 8$. Subsequently, these settings are applied to 90 unseen test images for evaluation.

Although 10-fold cross validation is widely used, over-fitting can occur in some cases, since cross validation may over-estimate a classifier's performance. In order to provide more reliable performance using a more comprehensive evaluation strategy, bootstrap sampling

validation is further employed for performance comparison using the MLP, SVM and Dempster-Shafer ensemble. In this research study, we employ .632 bootstrap (Han, Kamber & Pei, 2012), with the dataset sampled 500 times with replacement. For each bootstrap sampling, we obtain a training set of 180 lymphocytic images, where some images in the original dataset can occur more than once (because of sampling with replacement). The remaining data samples that are not included in the training set form the test set (Han et al., 2012).

Finally, the overall accuracy of the bootstrap model, C, is calculated as shown in Eq (4.33).

$$Acc(C) = \frac{1}{n}\sum_{i=1}^{n}\left(0.632 \times Acc(C_i)_{test\_set} + 0.368 \times Acc(C_i)_{train\_set}\right) \quad (4.33)$$

where $Acc(C_i)_{test\_set}$ and $Acc(C_i)_{train\_set}$ represent the accuracy rates of the model obtained with bootstrap sample *i* when it is tested using test set *i* and the original dataset of 180 images (Han et al., 2012), respectively. In this research study, the n = 500 represents 500 times of sampling with replacement.

In order to ensure a similar parameter tuning procedure is used for all the classifiers in bootstrap validation, 10-fold validation tuning as used for the single SVM is employed to identify optimal settings of the single MLP and Dempster-Shafer ensemble. Based on the results, the MLP has two hidden layers with 16 and 30 hidden nodes, respectively, in the first and second layers. For the Dempster-Shafer ensemble, the five MLP base models are identified. Both single MLP and each base model of the Dempster-Shafer ensemble share the same topology setting and use a learning rate of 0.1, a momentum rate of 0.8 and a termination error of 0.01, to achieve a balance between accuracy and generalisation performance.

4.6. Evaluation and Discussion

We employ the microscopic sub-image database, ALL-IDB2, for the evaluation of this research study. In order to test the discriminant capability of SDM, experiments have been conducted in comparison to other classic and advanced clustering algorithms, including LDA, FCM and FCS. The 80 features have also been extracted for single and ensemble classifiers to test the robustness and efficiency of the proposed clustering algorithm. Experiments indicate that the proposed system in this chapter outperforms typical methods and related research reported in the literature.

4.6.1. Evaluation of the Proposed SDM-based Clustering

As mentioned previously in Section 4.5.1, the system evaluation experiments of this research study use 180 sub-images of 60 lymphocyte (healthy) and 120 lymphoblast (unhealthy) cells extracted from ALL-IDB2. The ground truth of these selected images has been established based on the database annotation in further consultation with haematologists from the Royal Victoria Infirmary (RVI) Hospital, Newcastle-Upon-Tyne, United Kingdom, as shown in some examples in Figure 4.8. The ground truths and annotations of the lymphocytic sub-images as depicted in Figure 4.8, particularly the cell nucleus and cell cytoplasm columns of each of the lymphocytic cell sub-images, are utilised in the evaluation of the separation nucleus and cytoplasm results of the novel SDM algorithm and other base line algorithms using the two-dimensional correlation coefficient. Figure 4.9 shows the example results of the segmented nucleus (N) and cytoplasm (C) of the lymphocytic membrane samples obtained from ALL-IDB2 using different clustering techniques.

The separation of nucleus and cytoplasm by the proposed SDM clustering and SDM with morphological operations has the best accuracy as compared with those obtained from other prevalent methods. The SDM-based clustering gives better results in terms of complete separation of nucleus and cytoplasm, as well as recognition of the chromatin texture in the segmented nucleus. In particular, the chromatin texture is one of the important features in the nucleus and possesses a similar colour to that of the cytoplasm (e.g. the first and third blast cells in Figure 4.9); therefore, the extraction of nucleus becomes very difficult because the chromatin texture tends to be mis-clustered as the cytoplasm by FCS1 (Li et al., 2011), FCS2 (Wu et al., 2005), LDA and FCM clustering algorithms. In comparison with these methods, the proposed SDM-based clustering is able to identify most of the chromatin texture in the nucleus with relatively less mis-clustered pixels (“salt and pepper” conditions). To further improve the segmentation from SDM-based clustering, simple morphological operations are conducted on the segmented nucleus and cytoplasm in a vice versa manner to identify small hole areas in nucleus/cytoplasm, fill the holes and remove the filled pixels from the corresponding cytoplasm/nucleus cluster. As can be observed in the last column of Figure 4.9, combining the SDM-based clustering with the morphological operations manages to produce clean and precise separation results.

In order to validate the separation results of nucleus and cytoplasm in a quantitative manner, a correlation coefficient, as shows in Eq (4.32), is used to measure the degree of similarity against manually segmented nucleus and cytoplasm obtained in consultation with haematologists, as shown in some examples in Figure 4.8.

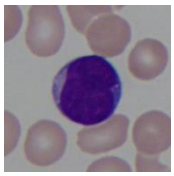
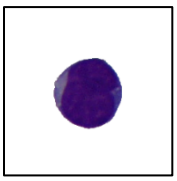
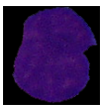
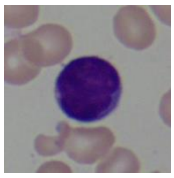
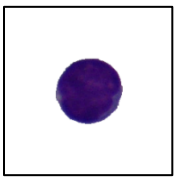
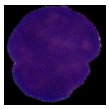
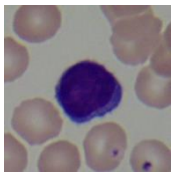
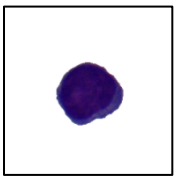
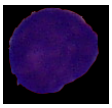
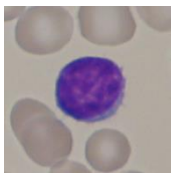
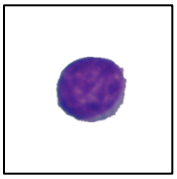
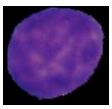
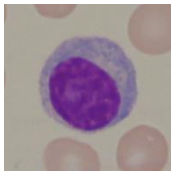
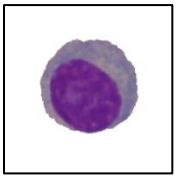
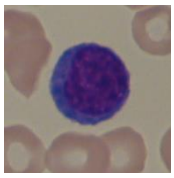
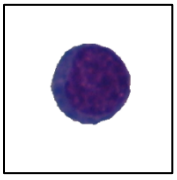
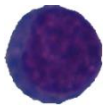
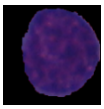
Lymphocytic cells file name and condition	Original WBC sub-images	Ground Truths			
		WBC membranes	Cropped membranes	Cell Nucleus	Cell Cytoplasm
Unhealthy (lymphoblast) Im004_1					
Unhealthy (lymphoblast) Im015_1					
Unhealthy (lymphoblast) Im024_1					
Healthy (lymphocyte) Im156_0					
Healthy (lymphocyte) Im196_0					
Healthy (lymphocyte) Im204_0					

Figure 4.8 The sub-image samples of the lymphocytic cells with ground truths and annotations from the haematologists.

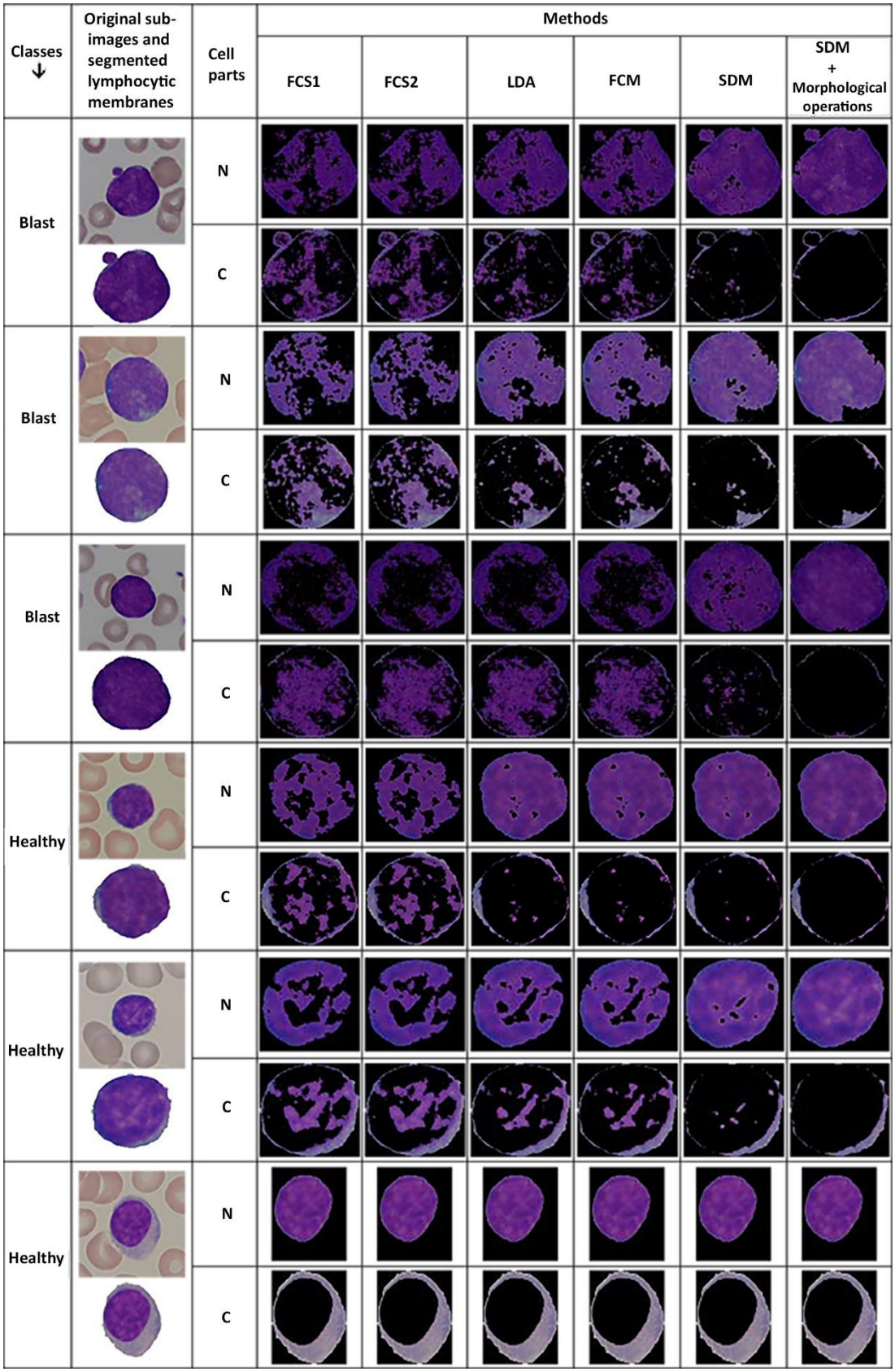


Figure 4.9 Comparison of the separation of nucleus and cytoplasm between the proposed SDM clustering and other clustering methods.

The average correlation coefficient of each compared method is shown in Table 4.1. From the results, the proposed SDM method with morphological operations performs the best with the highest correlation to human segmentation results relating to both nucleus and cytoplasm selection. Moreover, the SDM achieves better correlation results for both nucleus and cytoplasm and outperforms other compared segmentation methods. It is also interesting to note that, although FCM does not include between-cluster scatter evaluation, its robust membership function based on within-cluster scatter is able to produce comparable results to those of LDA, which employs both within- and between-cluster matrices, but without the implementation of any fuzzy membership. Even though efforts have been made to include between-cluster scatter, together with fuzzy membership, in the proposal of FCS, the developed fuzzy membership of FCS requires subjective tuning of the parameter η_j. As a result, FCS does not seem to perform well in the segmentation when η_j is fixed.

However, FCS2 (Wu et al., 2005) performs slightly better than FCS1 (Li et al., 2011) owing to the consideration of $(a_j - c)$ instead of $(x_i - c)$ in the between-cluster scatter evaluation, where a_j represents the fuzzy sample mean for j^{th} cluster, while x_i indicates the corresponding data sample and c represents the mutual centre of all clusters. The reason is mainly owing to the involvement of all data samples (i.e. all x_i) in FCS1 (Li et al., 2011), where very large and very small values can affect the evaluation of between-cluster evaluation. The proposed SDM-based clustering does not employ any fuzzy membership; therefore, it is not restricted to subjective tuning of parameter η_j. Overall, the proposed SDM is able to produce more promising segmentation results of nucleus and cytoplasm with a higher correlation coefficient as compared with those from other clustering algorithms.

Table 4.1. The correlation coefficient values of the proposed and several selected clustering methods in comparison to manual separation of nucleus (CorrN) and cytoplasm (CorrC) for 180 sub-images.

Methods	**CorrN**	**CorrC**
FCS1	0.627	0.624
FCS2	0.633	0.627
LDA	0.773	0.705
FCM	0.774	0.706
SDM	0.841	0.744
SDM + morphological operation	0.865	0.756

4.6.2. Evaluation of ALL Detection

In this study, we employ the MLP, SVM and Dempster-Shafer ensemble for ALL classification. Several evaluation strategies are applied to assess the system efficiency. We compare our research with other related work in the literature. To the best of our knowledge, Khashman and Abbas (2013), Putzu et al. (2014) and Madhukar et al. (2012) have achieved high recognition performances using the same ALL-IDB database. First of all, we analyse the results from our work and those from Khashman and Abbas (2013) because of their impressive system performance. Khashman and Abbas (2013) employed three different schemes of the training and testing data ratios for evaluating a total of 80 normal and abnormal lymphocyte images extracted from ALL-IDB2, i.e. 75%:25%, 50%:50% and 25%:75%. In each scheme, a balanced number of normal and abnormal samples in the training and testing sets were used. In order to have a fair comparison, we also employ the same three schemes of training and testing data ratios to evaluate our system performance using 80 randomly selected lymphocytic images from the ALL-IDB2 database. The details of comparison results are shown in Table 4.2. It can be clearly observed that SDM+SVM/MLP/Dempster-Shafer in this research significantly outperforms those of Khashman and Abbas (2013). The Dempster-Shafer results are better by 10%, 18.33% and 19.9% for the first, second and third schemes, respectively. Since the MLP is applied in both the work of this research study and that of Khashman and Abbas (2013), the MLP results achieved across the three schemes also clearly reveal the strength of the proposed SDM-based method, which provides more efficient nucleus-cytoplasm separation to achieve high ALL classification rates.

Table 4.2 Comparison of the recognition accuracy according to the three testing strategies used in Khashman and Abbas (2013) (N: Normal, A: Abnormal).

Training &Testing Split	**ALL Detection Accuracy**			
	Khashman and Abbas (2013) (%)	**SDM+SVM (%)**	**SDM+MLP (%)**	**SDM+ Dempster-Shafer (%)**
Training 75% (30(N):30(A)) Testing 25% (10(N):10(A))	90	90	95	100
Training 50% (20(N):20(A)) Testing 50% (20(N):20(A))	80	100	96.75	98.33
Training 25% (10(N):10(A)) Testing 75% (30(N):30(A))	75.1	86.67	91	95

Madhukar et al. (2012) and Putzu et al. (2014) are another two related studies in ALL diagnosis. Putzu et al. (2014) achieved 93.2% accuracy using SVM with RBF based on 10-fold cross validation, whereas Madhukar et al. (2012) achieved 93.5% accuracy with SVM using leave-one-out cross validation. Since SVM was used in both studies, and 10-fold cross validation is a better bias-variance trade-off method as compared with leave-one-out cross validation, we employ 10-fold cross validation for evaluating 180 lymphocytic images segmented using the proposed SDM clustering and morphological operations with SVM for ALL classification. Based on the experimental setting given to find the optimal configuration parameters for classifiers in Section 4.5.2, we achieve an accuracy rate of 96.67% for 10-fold cross validation using SVM classifier.

Even though 10-fold cross validation is widely implemented, it is undeniable that cross validation might over-estimate classifier performance owing to the issue of over-fitting. As a result, a more comprehensive evaluation, i.e. bootstrap sampling validation, is further conducted across MLP, SVM and Dempster-Shafer in this research study and the results of bootstrap validation are shown in Table 4.3.

Table 4.3 Comparison of ALL detection accuracy using the bootstrap validation method.

Validation Method	**Classifiers**		
	MLP (%)	**SVM (%)**	**Dempster-Shafer (%)**
Bootstrap Validation	95.96	95.61	96.72

Table 4.3 depicts the classifier performances for the original dataset of 180 images for bootstrapping. As can be observed, the Dempster-Shafer produces the highest accuracy of 96.72%, followed by the single MLP and SVM with 95.96% and 95.61% accuracies, respectively. Figure 4.10 shows the boxplot for 500 bootstrap sampling validation for each classifier. It can be seen that the Dempster-Shafer shows a better accuracy distribution with comparatively smaller variations between the 25% and 75% percentiles, as compared with those from the single SVM and MLP classifiers. Even though there are slight differences in terms of classification rate across different classifiers, significant ALL recognition is observed in both 10-fold cross validation and 500 bootstrap sampling validation. Overall, the proposed SDM clustering segmentation works well, and is able to produce high accuracy for normal and abnormal lymphocytes detection.

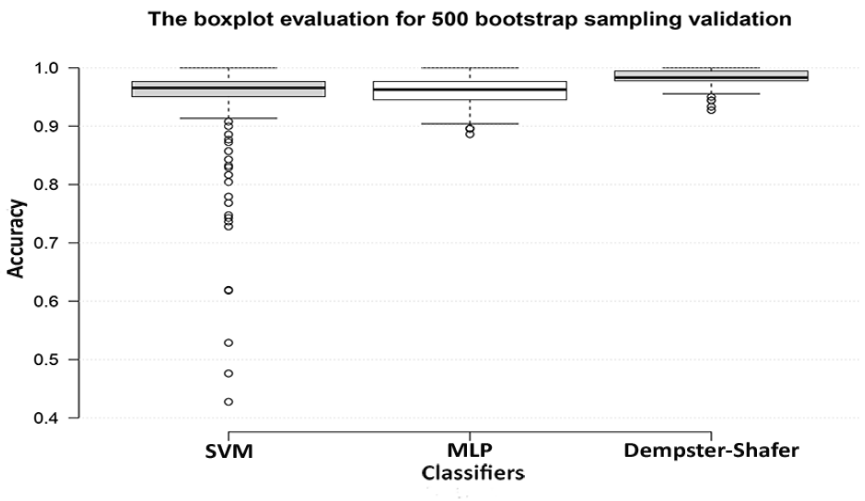


Figure 4.10 The boxplot evaluation for 500 bootstrap sampling validation.

4.7. Chapter Summary

This chapter has introduced the design and development of the novel clustering algorithm with stimulating discriminant measures of both within- and between-cluster scatter variances for the robust separation of nucleus and cytoplasm of the lymphocyte/lymphoblast cell membrane images. The SDM-based clustering overcomes the limitations of classical FCM, which only considers the within-cluster scatter variance. The between-cluster scatter criteria are designed based on the trade-off pertaining to several between-cluster measures (SB_{NWFE} and SB_{LDA}) through the application of the GA. The SDM-based clustering method achieves the highest correlation coefficient scores for the selection of nucleus and cytoplasm and outperforms LDA, FCM and FCS. A total of 80 feature descriptors are extracted from the segmented nucleus and cytoplasm. These features are used for the experiments in this chapter as the inputs to the MLP, SVM and Dempster-Shafer for lymphocyte and lymphoblast identification.

For comparison with the work of Khashman and Abbas (2013) using three evaluation schemes, the proposed SDM-based clustering integrated with Dempster-Shafer ensemble achieves the best accuracy rates of 100%, 98.33% and 95% and outperforms the results in Khashman and Abbas (2013) by 10%, 18.33% and 19.9%, corresponding to the three evaluation schemes, as has been shown in Section 4.6.2.

To provide a comprehensive evaluation study on the proposed SDM-based clustering method with feature extraction and recognition techniques, another case study is carried out using 180 lymphocytic images, as has also been revealed in Section 4.6.2. The results show that 10-fold cross validation, together with SVM, is able to produce an accuracy rate of 96.67%. In order to prevent over-estimation of the classifier performance, 500 bootstrap sampling validation is

further conducted using the SVM, MLP and Dempster-Shafer ensemble classifiers. The Dempster-Shafer ensemble achieves the highest accuracy rate of 96.72%.

Overall, the SDM-based clustering algorithm with a total of 80 features incorporated with single SVM, single MLP and Dempster-Shafer ensemble classifiers achieves better recognition accuracy in distinguishing normal and blast lymphocyte cells as compared with reported results in the literature.

Chapter 5: The Proposed BBPSO Variant for Feature Optimisation

5.1 Introduction

In this chapter, we propose a feature optimisation algorithm, namely a variant of Bare-Bones Particle Swarm Optimisation (BBPSO), to identify the most significant discriminative characteristic of the cell nucleus and cell cytoplasm segmented by the SDM-based clustering algorithm. The goal of this chapter is that we try to achieve the improvement of classification accuracy in terms of the performance of classifier, e.g. SVM, which employed the best-selected feature subsets from the extracted raw 80 features in Chapter 4 and then compare the obtained classification accuracy to the baseline algorithms and the existing researches. In addition, we aim at the development of a novel algorithm to find the most relevant features for the highly accurate and robust acute lymphoblastic leukaemia detection. The proposed BBPSO variant algorithm incorporates cuckoo search, dragonfly algorithm, BBPSO, and local and global random walk operations of uniform combination and Lévy flights, to diversify the search and mitigate the premature convergence problem of conventional BBPSO. It also employs subswarm concepts, self-adaptive parameters, and convergence degree monitoring mechanisms to enable a fast convergence rate. The several proposed strategies above work in a co-operative manner to guide the search to the global optima. For the evaluation of the proposed method, we conduct experiments for the ALL detection and classification using the SVM classifier with the identified optimal feature subsets, as shown in the green rectangle dashed-line in Figure 5.1. The reason of using SVM classifier for the evaluation of the proposed method owing to we aim to compare the outcome of the proposed BBPSO variant algorithm to the state-of-the-art researches.

This chapter is organised in the following way. Section 5.2 introduces the background and knowledge of the evolutionary algorithms. Section 5.3 presents the proposed BBPSO variant algorithm. The ALL detection and classification using SVM are discussed in Section 5.4. Evaluation and discussion of the proposed algorithm in comparison with other baseline optimisation methods using the ALL-IDB2 database (Labati et al., 2011a) are presented in Section 5.5. Finally, the summary of this chapter is described in Section 5.6.

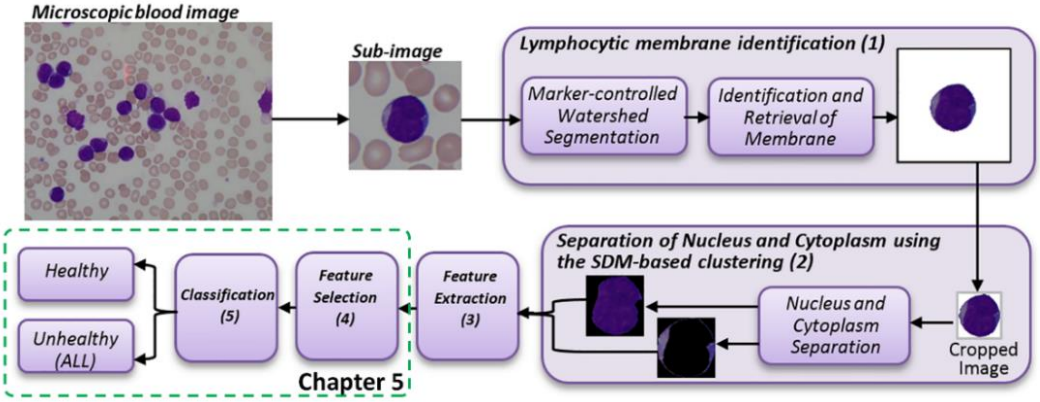


Figure 5.1 The proposed BBPSO variant algorithm for robust ALL detection.

5.2 Evolutionary Algorithms

This section introduces the background and knowledge of the evolutionary algorithms, which are employed in this research study. According to Eiben and Rudolph (1999) and Iglesia (2013), evolutionary algorithms have been developed to solve optimisation problems based on the iterative evolution of a population of solutions that mimics principles of biological evolution. Moreover, the function optimisation capability of the evolutionary algorithms is emphasised owing to its high adaptability to different problems and to which researchers cannot apply traditional optimisation techniques (Wong, 2016).

In the general process of an evolutionary algorithm, it starts with a randomly initialised population. In addition, the population, which has a variety of solutions, then evolves across several generations. In each generation, the fittest individuals or solutions are selected to become parents of other individuals. Next, they crossover with each other to generate new individuals called offspring. The new offspring individuals are randomly selected and then mutated at particular point. Afterwards, the algorithm selects the strong individuals or optimal solutions for survival to the next generation according to the method of survival selection, which is designed in advance, i.e. the overlapping condition of parents. The selected surviving individuals, parents, are employed to reproduce offspring for the next generation. Such a process is repeated until a satisfaction condition or a termination condition is met. Finally, the strongest individual or optimal solution is achieved (Wong, 2016), as illustrated in Algorithm 5.1.

Algorithm 5.1: A Typical Evolutionary Algorithm

```
Choose suitable representation methods;
 P(t): Parent population at time t
 O(t): Offspring population at time t

t ← 0;
Initialise P(t);
While not termination condition do
{   temp = Parent Selection from P(t);
      O(t+1) = Crossover in temp;
      O(t+1) = Mutate O(t+1);
          If overlapping then
              P(t+1) = Survival Selection from O(t+1) ∪ P(t);
          Else
              P(t+1) = Survival Selection from O(t+1);
          End if
       t ← t+1;
    }End while
Good individuals can then be found in P(t);
```

The design of an evolutionary algorithm combines with several components, including representation, parent selection, crossover operator, mutation operator, survival selection and termination condition. Examples of evolutionary algorithms are Swarm Intelligence (SI) algorithms such as GA, and their details are as follows.

SI is a special class of evolutionary algorithm, which was first introduced by Beni and Wang (1993). It is the artificial simulation or implementation of the collective behaviour and social behaviour intelligence of a group of animals in nature (Bonabeau, Dorigo, & Theraulaz, 1999). Moreover, the local rules for interaction between the individuals in social intelligence are decentralised controllers, such that the simulation of the social behaviour of the population can help to find the simple rules between them (Bonabeau et al., 1999; Mirjalili, 2015). In the process of SI, it maintains fixed-size population of individuals for search across generations. After each generation, the individuals have to evaluate their fitness, which is recorded and used to adjust the search strategy in the next generation. The search process stops when it finds the best individual or the maximum generation is reached (Wong, 2016).

As aforementioned in Section 2.3.3 in Chapter 2, GA is the most classic evolutionary algorithm. It draws inspiration from the Darwinian evolution theory, i.e. the concept of survival of the fittest in human or animal societies. In the GA, each individual has an information chain, which is a fixed-length binary array or bit string or binary string, as its genotype. Then the fitness of each individual is calculated for the selection process. After that, the algorithm then processes to select parents for one-point crossover to produce offspring, which subsequently undergo mutation operations. The offspring individuals become the population in the next generation. The process stops when the fittest reaches satisfaction or

the maximum number of generations is achieved (Bäck & Schwefel, 1993; Wong, 2016; Zhang et al., 2015d).

5.3 The Proposed BBPSO Variant Algorithm

In this research, we propose a BBPSO variant algorithm for feature optimisation, which mitigates the premature convergence problem of conventional BBPSO. The proposed algorithm incorporates multiple search strategies such as BBPSO, CS and DA to diversify the search in the primary and subswarms, respectively. An adaptive mechanism is also applied to identify stagnant situations and convergence degrees of each of the search algorithms employed. The proposed BBPSO variant employs Lévy flights and the uniform combination to increase particle swarm diversity if the primary or subswarm based search stagnates. In particular, the search strategies of the CS and DA algorithms possess two search capabilities, i.e. local and global search. CS employs Lévy flights and a discovery probability to control global (3/4 of the lifetime of CS) and local search (1/4 of the lifetime of CS) to satisfy global convergence requirements (Yang & Deb, 2009) whereas the DA employs static and dynamic swarming behaviours of dragonflies and models their social interaction behaviours (in search for food and evading enemies etc.) to balance between exploitation and exploration. Therefore, CS and DA are selected to extend the global and local search capabilities of conventional BBPSO and applied to the subswarms, respectively, to guide the search of global optimum. In particular, the discovery probability of CS is also dynamically adjusted in our algorithm as the number of iterations increases to adjust its effect and make initial attempts to overcome drawbacks of constant parameter setting in traditional CS to improve performance. Overall, BBPSO, CS, DA, Lévy flights and the uniform combination work collaboratively to drive the search out of local optimum trap and find the global optimal solutions. For example, if any of the above search strategies fail to increase global best solutions for a number of iterations and stagnate, other algorithms are used to escalate search and population diversity to enable the proposed approach to escape from local optimum trap. The proposed BBPSO variant algorithm is illustrated in Algorithm 5.2. Additionally, the flowchart of the algorithm is shown in Figure 5.2. We introduce the proposed algorithm in detail as follows.

As illustrated in Algorithm 5.2, the algorithm first of all performs a conventional BBPSO for N number of iterations to identify a global best solution, *gbest_bbpso*. A combination probability, p_{BBPSO}, is also embedded in BBPSO to observe its convergence rate and stagnant situations. If p_{BBPSO} > a random value to indicate stagnation occurred in BBPSO in the primary swarm, then Lévy flights are applied to the current primary swarm to increase particle swarm diversity. Moreover, this newly generated diversified swarm using Lévy flights is then

Algorithm 5.2: Pseudo-code of the Proposed BBPSO Variant Algorithm

```
Start
Initialise the position of each particle in the swarm;

Repeat {
   Repeat{//perform original BBPSO operation
        For each particle in the overall population do
         {
           Evaluate each particle by the defined fitness function;
           Update the position of each particle;
          Update the individual best particle pbest and the best particle
          gbest_bbpso in the overall population;
         }End for
       Update the combination probability p_BBPSO; //to observe stagnation
       in BBPSO
   }Until (stopping condition);
   If BBPSO stagnates (i.e p_BBPSO> rand)
     Apply Lévy flights to the overall swarm;
   End if
   Divide the population into two subswarms s1 and s2;

   //perform CS in subswarm s1
   Repeat{
     Perform CS as illustrated in Algorithm 5.3 in subswarm s1;
     Update the global best solution, gbest_cs;
     Update the combination probability, p_CS; //to observe stagnation in CS
   }Until (stopping condition);

   //perform DA in subswarm s2
   Repeat
   {
     Perform DA as indicated in Algorithm 5.4 in subswarm s2;
     Update the global best solution, gbest_da
     Update the combination probability, p_DA; //to observe stagnation in
     DA
   }Until (stopping condition);

     Compare the fitness of gbest_bbpso, gbest_cs, and gbest_da and assign
   the best leader to gbest.
     If CS stagnates (i.e. p_CS> rand)
         Conduct uniform combination as shown in Algorithm 5.5 to s1;
     End if
     If DA stagnates (i.e. p_DA> rand)
         Conduct uniform combination as shown in Algorithm 5.5 to s2;
     End if

     Discard the worst leader among gbest_bbpso, gbest_da, and gbest_da
   and combine s1 and s2 into one whole population again;
     Reinsert gbest into the updated population as the swarm leader;
}Until (stopping condition);
Return the most optimal solution(s);
End.
```

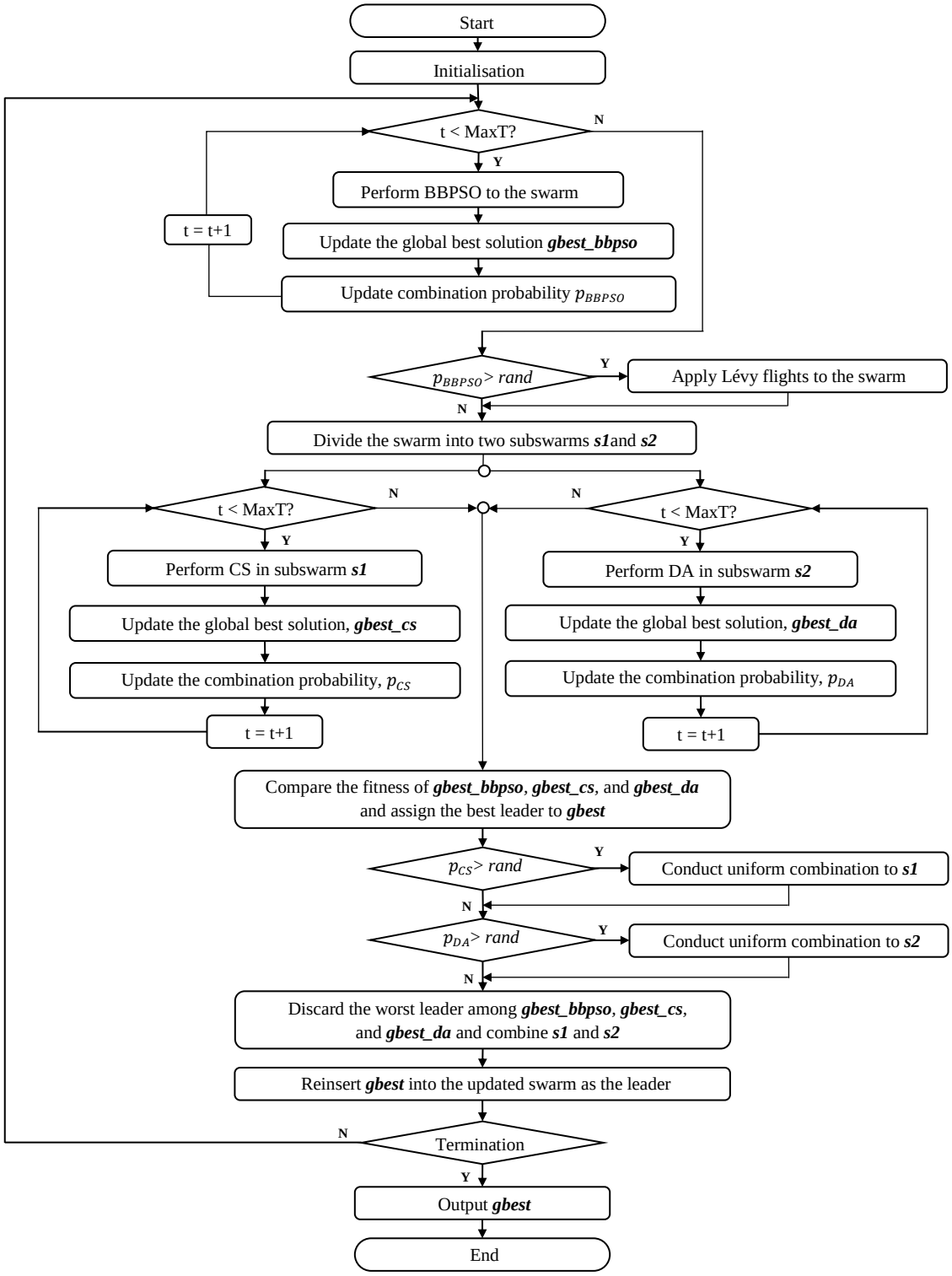


Figure 5.2 The flowchart of the proposed BBPSO variant algorithm.

divided into two subswarms, *s1* and *s2*. We employ CS and DA in each of the subswarm based search, respectively. After *N* number of iterations, the global best solutions, *gbest_cs* and *gbest_da*, are identified by CS and DA, respectively, in each subswarm. Then the three optimal solutions, i.e. *gbest_bbpso*, *gbest_cs* and *gbest_da*, obtained from BBPSO, CS and DA, respectively, are compared with each other and the one with the highest fitness value is assigned as the global best solution, i.e. *gbest*, whereas in the meantime, the worst leader among the three is discarded.

Furthermore, as illustrated in Algorithm 5.2, we also observe stagnant iterations of CS and DA by including combination probabilities, p_{CS} and p_{DA}, in the subswarm based search, respectively. If CS or DA stagnates (i.e. p_{CS} or p_{DA} is more than a random value), then uniform combination integrated with opposition-based mutation is used to diversify the population of the subswarms. Subsequently, the two updated subwarms by the uniform combination are merged to form an overall swarm. This newly formed primary swarm and the most recently identified *gbest* are subsequently passed on to the next generation for the search of global optimal solution(s). The search process iterates until the termination criteria are met, i.e. (1) the number of generations reaches 200, or (2) the fitness value of the identified global best solution equals to or is more than 0.98. The best solution, i.e. the selected feature subset, is obtained when either termination criteria is satisfied.

Most importantly, the three search strategies of BBPSO, CS, and DA, and the long and short jump mutation mechanisms of Lévy flights and uniform combination work in a collaborative manner to drive the search out of local optimum trap. For example, if BBPSO fails to generate a fitter leader and stagnates, Lévy flights are used to diversify the population of the primary swarm, which may further enhance subsequent CS and DA based search in each subswarm to avoid local optimum trap. Moreover, the BBPSO algorithm is also able to contribute to the retrieval of fitter global optimal solutions in a scale of the overall swarm to reduce the probability of premature convergence when the subswarm based search using CS or DA or both stagnate.

Furthermore, if the subswarm based search using CS stagnates, empirical results indicate that the DA-based search in the other subswarm employs static and dynamic swarming behaviours of dragonflies and is capable of achieving dramatic fitness improvement in comparatively later stage of iterations to identify a fitter leader and drive the search out of local optimum, whereas if the DA stagnates, results indicate that the CS algorithm with either a fixed or a dynamically changing discovery probability shows great capabilities to reach global optimality and guide the search to escape from local optimum. The diversified updated subswarms using the

uniform combination, incurred by the stagnation of CS or DA, may also enable the BBPSO-based search in the primary swarm in the next generation to achieve global optimality. Overall, the above co-operative strategy of the proposed algorithm enables primary and subswarm based search mechanisms, and the local (uniform combination) and global (Lévy flights) random walk operations to work in a collaborative manner to retain the population diversity, increase local exploitation and global exploration, overcome premature convergence of conventional BBPSO and guide the search to the ultimate global optima. Subsequently, each strategy of the proposed algorithm is introduced in detail as follows.

5.3.1 Bare-Bones Particle Swarm Optimisation (BBPSO)

PSO is proposed by Kennedy and Eberhart (Kennedy & Eberhart, 1995) and has been widely used as an efficient technique for feature selection. In PSO, each particle has a position in the search space represented by a vector $x_i = (x_{i1}, x_{i2}, ..., x_{iD})$ and a velocity denoted as $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$, where D denotes the dimensionality of the search space. Particles move in the search space in order to search for the optimal solution(s). Additionally, in PSO, the best position ever achieved by a particle, i.e. the personal best, *pbest*, and the best position of the overall swarm, i.e. the global best, *gbest*, are used to update the velocity and position of each particle. Eq (5.1) and Eq (5.2) define the velocity and position updating in PSO.

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_1 * \left(p_{id} - x_{id}^{t}\right) + c_2 * r_2 * \left(p_{gd} - x_{id}^{t}\right) \tag{5.1}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{5.2}$$

where w indicates the inertia weight and c_1 and c_2 denote acceleration constants with r_1 and r_2 as random values uniformly distributed within [0, 1], $d \in D$ denotes the dth dimension of the particle while t represents the iterations. p_{id} and p_{gd} refer to the elements of *pbest* and *gbest* in the dth dimension, respectively.

Furthermore, BBPSO is a PSO variant (Kennedy, 2003). Compared with conventional PSO, it does not consider the velocity, but only updates the positions of particles. Gaussian distribution is employed for position updating in BBPSO, as illustrated in Eq (5.3).

$$x_{id}^{t+1} = \phi\left(\frac{pbest_{id}^{t} + gbest_{d}^{t}}{2}, \left|pbest_{id}^{t} - gbest_{d}^{t}\right|\right) \tag{5.3}$$

where ϕ denotes Gaussian distribution and $\frac{pbest_{id}^{t} + gbest_{d}^{t}}{2}$ represents the mean or expectation of the distribution with $\left|pbest_{id}^{t} - gbest_{d}^{t}\right|$ as the standard deviation. Using Eq (5.3), the

new position of a particle is distributed according to Gaussian distribution, although other distribution techniques other than Gaussian could also be applied. Compared with conventional PSO, BBPSO does not require any operating parameters, is more efficient and has also been extensively applied to real-world single and multi-objective optimisation problems (Zhang, Gong, & Ding, 2012a).

In this research, the following settings are employed for the proposed BBPSO algorithm. We set the initial swarm with a population size of 50 and the maximum number of overall (external) generations is 5 with another 10 iterations employed for each internal search algorithm (i.e. CS, BBPSO and DA). As illustrated in Algorithm 5.2, at the initial stage of the algorithm, we employ 10 iterations for BBPSO to obtain the initial swarm leader. In order to observe the convergence degree of BBPSO, the algorithm includes a combination probability, p_{BBPSO}. Motivated by Zhang et al. (2015d), this combination probability is defined in Eq (5.4) and can be adjusted dynamically based on the number of stagnant iterations.

$$p_{BBPSO} = \frac{0.2}{1+e^{(5-num\_bbpso)}} \tag{5.4}$$

where num_bbpso represents the number of stagnant iterations in BBPSO. If the fitness of the *gbest* identified by BBPSO does not show obvious improvement between two successive iterations, num_bbpso is incremented by 1. When this combination probability, p_{BBPSO}, is more than a random value, Lévy flights defined in Eq (5.5) are applied to diversify the overall population.

$$x_{id}^{t+1} = x_{id}^{t} + (x_{max}^{d} - x_{min}^{d}) \times \psi(\lambda) \tag{5.5}$$

where ψ represents Lévy flights with λ as the random step length. x_{min}^{d} and x_{max}^{d} represent the minimum and maximum values in the dth dimension, respectively. According to Eq (5.4), p_{BBPSO} is increased from 0 to 0.2 along with the increase of the number of stagnant iterations, num_bbpso (p_{BBPSO} approaches 0.2 when $num\_bbpso \geq 10$). Therefore, the global random walk of Lévy flights is more likely to be activated when stagnation iterations increase in BBPSO. Then we divide the primary swarm into two subswarms, as illustrated in Algorithm 5.2. CS and DA are used to guide the search in each subswarm, respectively. A combination probability is embedded in CS and DA as well to observe stagnation in each algorithm. A local random walk operation (i.e. uniform combination) will be activated to increase particle swarm diversity to overcome local optimum when CS or DA stagnates.

In this research, the following fitness function, commonly applied to many other related applications (Zhang et al., 2015d), defined in Eq (5.6) is used to evaluate the proposed algorithm and other comparable optimisation methods.

$$\text{fitness}(C) = \mu * \text{accuracy}_C + (1 - \mu) * (\text{number\_features}_C)^{-1} \tag{5.6}$$

In Eq (5.6), μ and $1 - \mu$ denote the weights for classification accuracy, $accuracy_c$, and the number of selected features, $number\_features_c$, respectively. Since the classification accuracy is regarded as more important than the number of selected features, μ is assigned a higher value than that of $1 - \mu$.

5.3.2 Cuckoo Search Algorithm (CS)

The CS algorithm is initially proposed by Yang and Deb (Yang & Deb, 2009). Theoretical studies indicated that CS possesses local and global search mechanisms to fulfil global convergence. Research also indicated that CS is far more efficient and outperforms other meta-heuristic algorithms (such as GA, PSO, etc.) (Ljouad et al., 2014; Yang & Deb, 2010). Therefore, it is selected to guide the subswarm search in this research study.

The CS algorithm employs the following three main principle rules for the search of the global optimal solutions. Firstly, each cuckoo lays one egg (solution) at a time which is discarded in a randomly chosen nest. Secondly, the best nests with high-quality eggs are selected for the next generations. Thirdly, the host bird discovers the egg laid by a cuckoo with a probability, p_a. For example, a fraction p_a of worse nests will be abandoned and replaced by new nests. In CS, the number of available host nests is also usually set to a constant number during the search. The algorithm aims to replace not-so-good solutions in the nests with new and potentially better solutions. The pseudo-code of CS is provided as Algorithm 5.3 (Yang & Deb, 2009).

In this research, the initial population of CS has 25 randomly selected particles (i.e. half of the original swarm). In each generation, a global random walk strategy using Lévy flights defined in Eq (5.7) is applied in order to generate a new solution, x_i^{t+1}.

$$x_i^{t+1} = x_i^t + \alpha \times \psi(\lambda) \tag{5.7}$$

where x_i^{t+1} and x_i^t denote the ith solution in t+1 and t generations, respectively. ψ represents Lévy flights operation with λ as the random step length ($1 < \lambda \leq 3$) and α is the step-size scaling factor. Lévy flights perform a random walk where their random step-lengths are distributed based on a Lévy probability distribution. Lévy flights thus enable an offspring solution to

jump further away from its parent solution to increase global exploration and avoid local optimum trap (Hakli & Uğuz, 2013; Levy, 1954).

Algorithm 5.3: Pseudo-code of Cuckoo Search

```
Start
Initialise a population of n host nests x_i (i = 1, 2, …, n);

While (termination criteria are not met)
{
    Get a cuckoo i randomly by Lévy flights (using Eq (5.7));
    Evaluate the fitness f_i of the solution;
    Choose a nest j among n randomly ( j ∈ {1,2, ... , n} );
        If (f_i > f_j)
            Replace j by the new solution i;
        End if
    Abandon a fraction (p_a) of worse nests and build new nests (using Eq (5.8));
    Keep best solutions (i.e. nests);
    Rank the solutions and find the current best;
}End while
End
```

Moreover, a fraction of worse nests in CS is discovered with a probability p_a, which will be replaced by new nests (solutions). The following local random walk operator, defined in Eq (5.8), is applied to generate new solutions.

$$x_i^{t+1} = x_i^t + \alpha s \times H(p_a - \varepsilon) \times (x_k^t - x_l^t) \tag{5.8}$$

where x_k^t and x_l^t denote solutions selected randomly by random permutation, s denotes the step size and $H(v)$ represents a Heaviside function[3] while ε is a random number drawn from a uniform distribution. This new solution, x_i^{t+1}, is accepted as a new solution if it has a better fitness value than that of x_i^t. This random walk strategy defined in Eq (5.8) increases the local exploitation capability of the CS algorithm. Also, the population diversity in CS is determined by the discovery probability, p_a. For example, a higher value of the parameter, p_a, leads to the increase of population diversity and thus fast convergence speed whereas a lower value of the parameter may lead to premature convergence and a slow convergence rate. As illustrated in Eq (5.7) and Eq (5.8), the CS algorithm employs both local and global search mechanisms, controlled by a switching/discovery probability, to achieve global convergence.

[3] Heaviside function is a discontinuous function whose value is zero, '0', for negative argument and one, '1', for positive argument, i.e. $H(-1) = 0$; $H(3) = 1$.

Also existing research indicates that the setting of p_a and α plays very important roles in fine-tuning the solution vectors and adjusting the convergence rate of CS (Valian, Tavakoli, Mohanna, & Haghi, 2013). In particular, the search process will require large values for both of these parameters at the beginning stage of iterations to increase global search capabilities and comparatively smaller values for these parameters to fine-tune potential solution vectors in final iterations. Therefore, constant parameter setting in CS may have limitations. For example, existing research indicates that the fixed settings for both p_a and α may lead to either increased iterations (e.g. when both values are small) or incapability in exploiting in local search space efficiently to find the best solution(s) (e.g. when both values are large) (Valian et al., 2013). Therefore, in this research, we make initial attempts and propose a self-adaptive parameter tuning strategy defined in Eq (5.9) to dynamically adjust the discovery probability, p_a, which changes as the number of generation increases during the execution of CS in the proposed algorithm.

$$p_a^{t+1} = p_a^t - \frac{1}{MaxT} \tag{5.9}$$

where $MaxT$ represents the maximum number of iterations for CS and p_a^{t+1} and p_a^t represent the discovery probability in t+1 and t iterations, respectively. In this way, p_a is decreased as the number of iterations increases which enables CS to start the search with a comparatively larger value of p_a to increase search and population diversity and apply a much smaller p_a to fine-tune the identified solution vectors in the final stage to identify the most optimal solutions. In this research, experiments have been conducted using both a fixed and self-adaptive p_a within CS in the proposed algorithm in order to identify the effect of a dynamic changing p_a under different experimental settings. The initial parameter setting of CS in the proposed algorithm is illustrated as below, which is recommended by our experimental trials and other research (Yang & Deb, 2009).

n (i.e. the number of host nests) = 25; p_a (i.e. initial discovery probability) = 0.3; λ (i.e. the random step length for Lévy flights) = 1.5; α (i.e. the step-size scaling factor) = 0.08;

Furthermore, as indicated in Algorithm 5.2, similar to BBPSO, we also include a combination probability, p_{CS}, defined in Eq (5.10) in CS to observe its convergence degree.

$$p_{CS} = \frac{0.2}{1+e^{(5-num\_cs)}} \tag{5.10}$$

where num_cs represents the number of stagnant iterations in CS. When the combination probability, p_{CS}, is more than a random value, instead of using Lévy flights as in BBPSO,

uniform combination integrated with opposition-based mutation is applied to increase the population diversity of CS. Subsequently, we introduce the DA based search in the other subswarm as follows.

5.3.3 Dragonfly Algorithm (DA)

The DA is proposed by Mirjalili (2015). It simulates and implements static and dynamic swarming behaviours of dragonflies to balance between global exploration and local exploitation. The search process of the DA employs five social interaction behaviours of dragonflies, including separation, alignment, cohesion, attraction (towards food) and distraction (outwards enemies), which distinguishes DA from PSO and other SI algorithms. The swarming factors (i.e. weights) associated with each of these five social behaviours play important roles in affecting exploration and exploitation capabilities of the algorithm in the search space. Experimental results of DA in this research study also indicate its tendency of a performance surge in final generations when more neighbouring dragonflies are gathered. Because of its impressive performance and satisfying global convergence requirement, it is selected in this research to guide the subswarm search.

We introduce the modelling of the five social interaction behaviours of dragonflies as follows. Firstly, the separation behaviour, denoted as S_i, indicates the static collision avoidance of the individuals from other neighbourhood individuals, which is defined in Eq (5.11) (Mirjalili, 2015).

$$S_i = -\sum_{k=1}^{n} x - x_k \tag{5.11}$$

where x and x_k denote the positions of the current individual and the kth neighbourhood artificial dragonfly, respectively, while n represents the number of individuals in the neighbourhood.

The alignment behaviour, represented as A_i, is calculated using Eq (5.12), which refers to the velocity matching among neighbourhood individuals (Mirjalili, 2015).

$$A_i = \frac{\sum_{k=1}^{n} V_k}{n} \tag{5.12}$$

where V_k denotes the velocity of the kth neighbourhood individual.

Cohesion, C_i, is defined in Eq (5.13), which models the inclination of individuals to travel towards the centre of the mass of the neighbourhood (Mirjalili, 2015).

$$C_i = \frac{\sum_{k=1}^{n} x_k}{n} - x \tag{5.13}$$

Motivated by survival tactics, dragonflies are attracted towards a food source, F_i, and flee away from an enemy, E_i, which are simulated using Eq (5.14) and Eq (5.15), respectively (Mirjalili, 2015).

$$F_i = x^* - x \tag{5.14}$$

$$E_i = x^\epsilon + x \tag{5.15}$$

where x^* in Eq (5.14) and x^ϵ in Eq (5.15) represent the positions of a food source and an enemy, respectively.

Similar to the PSO algorithm, the movement of artificial dragonflies in the search space is carried out by updating the step/velocity and position vectors, which are defined in Eq (5.16) and Eq (5.17), respectively (Mirjalili, 2015).

$$\Delta x_{t+1} = (sS_i + aA_i + cC_i + fF_i + eE_i) + w\Delta x_t \tag{5.16}$$

In Eq (5.16), *s*, *a*, *c*, *f*, and *e* represent the swarming weights/factors for separation, alignment, cohesion, attraction (towards food) and distraction (outwards enemies), respectively, with *w* as the inertia weight whereas Δx_{t+1} and Δx_t represent the step/velocity vector in *t*+1 and *t* iterations, respectively. Eq (5.17) shows the position updating based on the step vector calculated by Eq (5.16).

$$x_{t+1} = x_t + \Delta x_{t+1} \tag{5.17}$$

where x_{t+1} and x_t indicate the positions of an individual in *t*+1 and *t* iterations, respectively.

Eq (5.16) and (5.17) simulate social behaviours of an artificial dragonfly when it has at least one neighbouring individual (Mirjalili, 2015). However, in order to increase global exploration of the DA, a random walk such as Lévy flights is applied to model its flying around behaviour in the search space when there is no neighbouring solution (dragonfly) available. Eq (5.18) shows the random walk behaviour using Lévy flights (Mirjalili, 2015).

$$x_{t+1} = x_t + x_t \times \psi(d) \tag{5.18}$$

where ψ represents Lévy flights with *d* as the dimension of the position vectors.

The pseudo-code of the DA algorithm is provided in Algorithm 5.4 (Mirjalili, 2015). In this study, the algorithm starts with the initialisation of a set of 25 random solutions (i.e. half of the overall swarm). The position and velocity vectors are also assigned with random values

within the lower and upper bounds of the variables initially. Eq (5.16) - Eq (5.17) or Eq (5.18) is used to update the velocity/step and position vectors, respectively, in each iteration. Since the DA simulates not only a dynamic swarm where alignment for flying is high while maintaining proper separation and cohesion, but also a static swarm where alignment is low and cohesion is high while attacking prey, the swarming factors *a* and *c* are adjusted accordingly to enable the effective exploring and exploiting of the search space. Furthermore, the five swarming factors, *s*, *a*, *c*, *f*, and *e*, also enable different global and local search behaviours to be achieved during optimisation in DA. The above position updating process continues until the termination criteria are fulfilled.

Algorithm 5.4: Pseudo-code of Dragonfly Algorithm

```
Start
Initialise a population of n dragonflies x_i (i = 1, 2, …, n);
Initialise step vectors Δx_i (i = 1, 2, …, n);
While (termination criteria are not met)
{
    Evaluate the fitness of all dragonflies;
    Update the food source and enemy;
    Update the swarming factors, i.e. w, s, a, c, f, and e
    Generate S, A, C, F and E (using Eq (5.11)-(5.15));
    Update neighbouring radius;
      If (there is at least one neighbouring dragonfly to the current individual)
            Update velocity and position vectors (using Eq (5.16)-(5.17));
      Else
            Update position vector (using Eq (5.18));
      End if
    Correct the new positions based on the boundaries of variables if required.
}End while
End
```

Furthermore, as discussed earlier and indicated in Algorithm 5.2, we have also embedded a combination probability, p_{DA}, in DA to observe its convergence, which is defined in Eq (5.19).

$$p_{DA} = \frac{0.2}{1+e^{(5-num\_da)}} \tag{5.19}$$

where num_da represents the number of stagnant iterations in DA. Similar to the case for CS and BBPSO, when the combination probability, p_{DA}, is more than a random value, uniform combination integrated with opposition-based mutation is applied to diversify the subswarm in order to enable the search to escape from local optimum. As the cases for BBPSO and CS,

the influence of uniform combination is strengthened when more stagnation iterations occur in DA. We introduce the random walk mechanism for population diversity preservation using uniform combination in the next section.

5.3.4 Uniform Combination

When CS and DA in subswarm based search stagnate, uniform combination is activated to increase population diversity to overcome premature convergence. In this research, this local random walk operator combines opposition-based mutation to diversify the particles.

In uniform combination, firstly, we select a range of elements from a particle, UC_n, based on the combination probability of CS and DA. Then we randomly identify a starting point, l, where the combination process starts. For each element of a particle from the starting point, l, to the dynamically adjusted range, UC_n, the opposition-based mutation illustrated in Eq (5.20) is applied. For example, this opposition-based mutation is applied separately to the specific dimensions of each particle to increase its diversity.

$$x_{ij}^{t+1} = x_{min}^{j} + x_{max}^{j} - x_{ij}^{t} \tag{5.20}$$

where $j \in [l, l+UC_n]$ denotes the jth dimension of the ith particle while x_{min}^{j} and x_{max}^{j} represent the minimum and maximum values in the jth dimension, respectively. The pseudo-code for uniform combination is illustrated in Algorithm 5.5.

Algorithm 5.5: Pseudo-code of the Proposed Uniform Combination

```
Start
//N_sub: the size of the subswarm. D_s: the size of dimensions for each
//particle. pc: the combination probability which is replaced by pCS or pDA
//accordingly for each subswarm.
For (i=1 to N_sub)
{
   If (pc> rand)) //rand generates a random number within [0, 1]
   {
      UCn = [pc * D_s]; //calculate the range
      l = rand(D_s - UCn , 1); //select a starting point randomly
         For (j = l to (l+UCn))
            x_ij = x_min^j + x_max^j - x_ij; //apply opposition-based mutation
         End for
   }End if
}End for
Output the new subswarm;
End
```

Empirical results indicate that the uniform combination with opposition-based mutation increases the subswarm population diversity and enables the proposed algorithm to overcome local optimum. Subsequently, the identified discriminative feature subsets for healthy and blast cells are used for the detection of lymphocytes and lymphoblasts. We have also compared the proposed algorithm with other state-of-the-art PSO variants and advanced and conventional search methods to identify its efficiency. Empirical results indicate the superiority of the proposed BBPSO variant algorithm in comparison to other methods. Detailed evaluation results are discussed in Section 5.5.

5.4 The ALL Detection and Classification

5.4.1 Evaluation Datasets

The dataset for this study includes 180 lymphocytic sub-images, which contain 60 lymphocyte (normal or healthy) and 120 lymphoblast (abnormal or unhealthy) images extracted from the ALL-IDB2 (Labati et al., 2011a) in consultation with the haematologists, as aforementioned in Chapter 3, Section 3.3, and also as utilised in the experiments in Chapter 4. Furthermore, we categorise two datasets for all experiments in this chapter. The details of each dataset are as follows. The first dataset, which is as well as used in Chapter 4, contains a balanced number of lymphocytes and lymphoblasts for both training and testing. Therefore, the first dataset, known as first experiment setting, includes 90 training (30 healthy and 60 unhealthy lymphocytes) and 90 (30 healthy and 60 unhealthy lymphocytes) unseen testing sample images. Moreover, for the second dataset, the training samples include a balanced number of lymphocyte and lymphoblast cell images, while the test samples are the remaining unseen images. Therefore, the second dataset, known as the second experiment setting, includes 100 training (50 healthy and 50 unhealthy) images, and 80 (10 healthy and 70 unhealthy) unseen test sample images.

5.4.2 Finding the Optimal Configuration Parameters for a Classifier

In this research study, we employ an SVM for classifying normal and abnormal lymphocyte cell images. Before classification, we conduct normalisation to the selected feature subsets, which are scaled into the range of [-1, 1]. These scaled features are then used as the inputs of a classifier for recognising normal and abnormal lymphocyte cell images. Moreover, all experiments are implemented based on MATLAB software versions 8.5 (R2015a) and using CPU Intel Core i7 3.6 GHz personal computer with memory 16 GB running on Microsoft Windows 7 Enterprise operating system.

In order to recognise normal and abnormal lymphocyte cell images, the SVM with radial basis function kernel (RBF) is employed in this research due to the fact that it supports nonlinear mapping of samples and has fewer hyper-parameters (Hsu & Chang, 2008). Moreover, the kernel parameter setting plays a very important role in achieving optimal classification performance (Ding & Chen, 2010). In this research, we employ grid search method (Hsu & Chang, 2008) to determine the scaling factor, γ, and the soft margin constant, *Co*, to make SVM achieve optimal performance based on RBF kernel. By using exponentially growing sequences, the ranges from 2^{-5} to 2^{15} and 2^{-10} to 2^{5} are searched for Co and γ, respectively. Furthermore, a 10-fold cross validation has been conducted in order to find the best combinations of parameters, known as tuple value (i.e. scaling factor (γ), soft margin constant (*Co*)), and also to avoid overfitting for the SVM. Additionally, the parameter settings, which achieve the best performance from the training dataset, are employed for the evaluation of the unseen testing images at the test stage for the proposed algorithm and other comparable baseline optimisation algorithms.

In this study, we conduct two experiments according to the two fitness functions, i.e. the fitness function 1 defined in Eq (5.6) and the fitness function 2 written in Eq (5.21), as below. The fitness function 1 illustrated in Eq (5.6) indicates stronger focus on classification accuracy in comparison to the emphasis given to the number of selected features. We have also provided the fitness function 2, as defined in Eq (5.21), which indicates a comparatively more balanced trade-off between classification accuracy and the number of selected features.

$$\text{fitness}_2(C) = \mu * \text{accuracy}_C + (1-\mu) * \left(1 - \frac{\text{number}_{\text{features}_C}}{\text{number}_{\text{all}}}\right) \qquad (5.21)$$

where $number_{all}$ and $number_{features_C}$ indicates the overall number of raw features (i.e. 80) and the number of selected features, respectively. The second part of the above equation focusing on the number of selected features has more influence on the overall fitness calculation than the corresponding part of the original fitness function defined in Eq (5.6). In addition, the same weight settings of μ and $1-\mu$ as those used in Eq (5.6) are also applied to this newly defined fitness function.

Furthermore, as mentioned earlier, in each experiment, we also conduct two experiment settings to evaluate the performance of the proposed method and the comparable baseline optimisation methods as following: (i) the first experiment setting, which uses the 90 unbalanced training images, i.e. 60 unhealthy and 30 healthy lymphocyte cell images, and 90 unseen testing images and (ii) the second experiment setting, which employs the 100 balanced

training images, i.e. 50 unhealthy and 50 healthy lymphocyte cell images and 80 unseen testing images, i.e. 70 unhealthy and 10 healthy lymphocyte cell images.

5.5 Evaluation and Discussion

In order to compare the proposed method with other comparable baseline algorithms, we have implemented the following optimisation methods, including binary BBPSO (Zhang et al., 2015d), ELPSO (Jordehi, 2015), conventional BBPSO, CS, DA, PSO and GA for the comparison. The 180 lymphocytic images under the abovementioned two experiment settings are employed to evaluate the efficiency of the proposed BBPSO algorithm and compared to all other methods. Since the proposed method and the other baseline optimisation methods are all stochastic algorithms, we also employ 30 independent experiment runs for each algorithm to identify discriminative feature subsets and use the average of 30 trials for comparison. Moreover, the proposed method with both a fixed and a self-adaptive *pa* within the CS is tested under the two experiments. The evaluation results of all experiments are as follows.

5.5.1 Experiment 1 Using Fitness Function 1

In experiment 1, first, we compare convergence rates of all algorithms at the training stage for both experiment settings, i.e. 90 unbalanced and 100 balanced training images, owing to a convergence rate being of practical importance for an iterative method, which is the speed at which an optimisation sequence approaches its limit. Thus, this experiment illustrates the convergence rates, known as convergence curves, of all algorithms, which employed the 90 unbalanced and 100 balanced training datasets, as depicted in Figure 5.3 and Figure 5.4, respectively. In addition, a higher convergence rate means that an optimisation algorithm requires fewer iterations to produce a useful approximation of the problem domain, whereas a lower convergence rate means that an optimisation algorithm needs more iterations to produce a useful approximation solution. Figures 5.3 and 5.4 show the averaged convergence curves of all algorithms over 30 experiment runs employing the 90 unbalanced and 100 balanced training lymphocytic cell images, respectively. In particular, when trained with the 100 balanced images, the proposed method embedded with a self-adaptive (or changing) *pa* balances well between global exploration and exploitation, as indicated in Figure 5.4, and has the fastest convergence rate in comparison with the one that embedded a fixed *pa* in the CS, and vice versa when trained with the 90 unbalanced images, as shown in Figure 5.3. On the other hand, in comparison to the other baseline algorithms, the proposed method embedded with both a fixed and a self-adaptive *pa* shows the efficient exploration and exploitation capabilities and achieves the superior and fastest convergence rate. Overall, the proposed method has a comparatively higher convergence rate with a fewer number of iterations needed

to achieve a useful approximation of healthy and unhealthy lymphocyte cell images, whereas the other comparable baseline methods illustrate comparatively lower convergence rates with more iterations required to reach a useful approximation of the lymphocytic cell images.

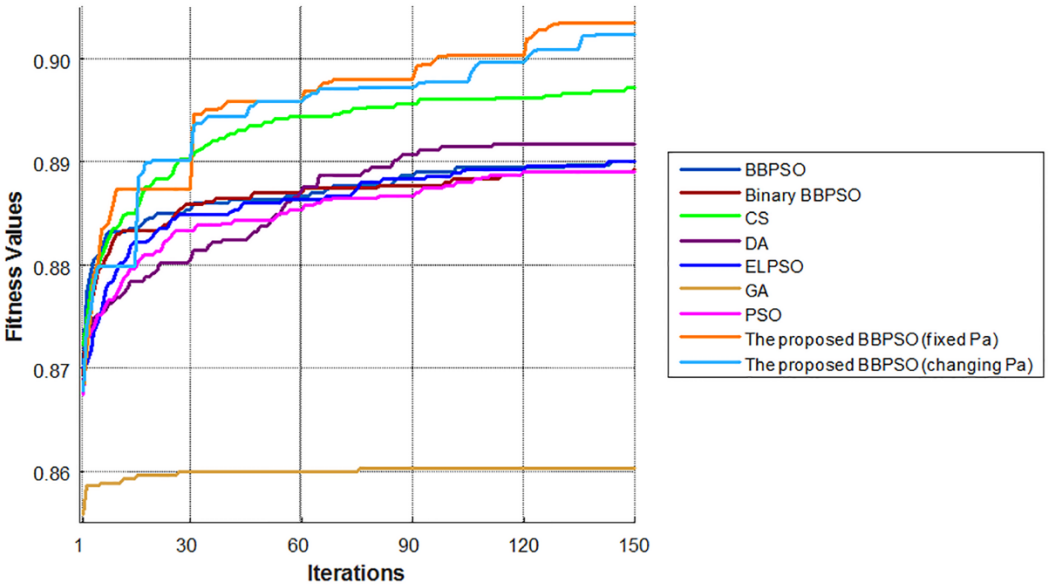


Figure 5.3 The convergence curve of the proposed algorithm over 30 experiment runs using 90 unbalanced training lymphocytic cell images.

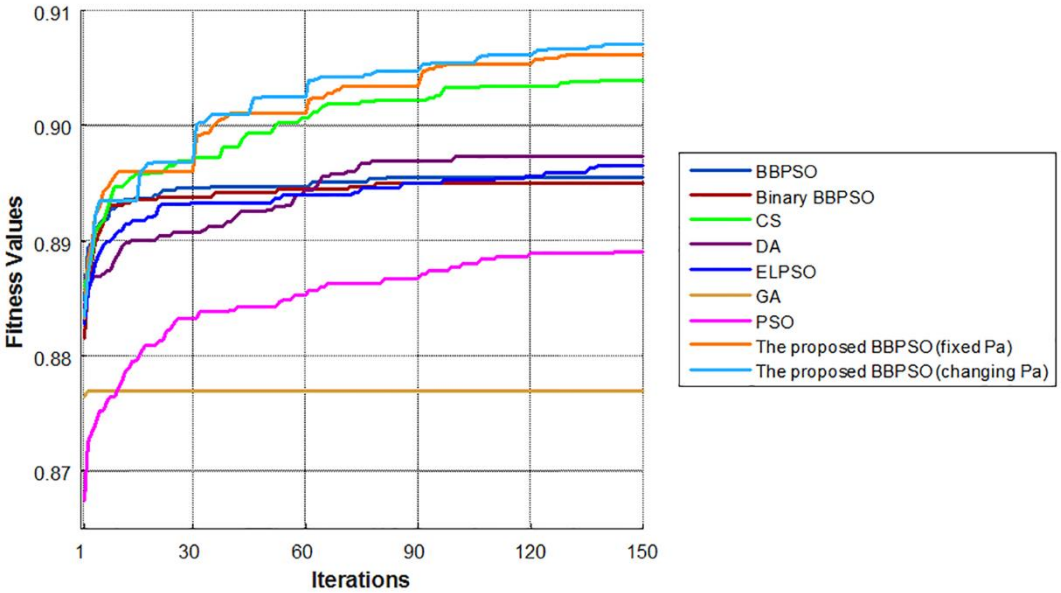


Figure 5.4 The convergence curve of the proposed algorithm over 30 experiment runs using 100 balanced training lymphocytic cell images.

By observation of both experimental settings, the convergence rates of the proposed methods are the best, followed by those of comparable methods, advanced and conventional search algorithms, such as the CS, DA, ELPSO, BBPSO, binary BBPSO, PSO and GA, respectively. The performance differences between balanced and unbalanced training sets as shown in Figure 5.3 and Figure 5.4, respectively, are summarised in the following two points. First, the fitness value of a balanced training set tends to be higher than the one obtained using an unbalanced training set, which means that training with a balanced training set can support the optimisation algorithm to find the best solution (a set of best-selected features), which satisfies the objective function indicating the high fitness values close to 1. Secondly, the trend of convergence rates of all algorithms in Figure 5.4 tend to be faster in fewer iterations when training with a balanced training set compared with an unbalanced training set. In order to fairly compare the proposed method to those baseline search strategies, the iterations in both Figure 5.3 and Figure 5.4 are the real iteration numbers for comparison of all search strategies. In particular, the real iteration numbers of the proposed method, which includes both the external loop, i.e. the repeat-until loop controlling the *iteration* variable, and the internal loop, i.e. the for-loop controlling the *t* variable of each search strategies, as depicted in Algorithm 5.2, are indicated in both Figure 5.3 and Figure 5.4. E.g. the iteration number 30 is calculated from the multiplication of the number of external loops, i.e. 1, and the summation of the three internal loops, i.e. 30 is the summation of the three search strategies that each search has defined as 10 maximum iterations of its for-loop. The convergence curves of the proposed method have already achieved the highest average fitness value over 30 experiment runs at iteration 30 in both Figure 5.3 and Figure 5.4 for both experimental settings compared to the advanced baseline comparable methods, which take at least approximately a further 30, 60 and 100 iterations for the CS, DA and ELPSO, respectively, to have performance improvements or to reach the similar fitness values of the proposed method. Moreover, when observing the convergence rate of the CS, BBPSO and DA algorithms, we found that the CS and BBPSO have fast convergence at the initial iterations owing to their local and global searching strategies, whereas the DA improves its performance quickly at the final iterations because the searching mechanism has benefits during the swarming/moving of each dragonfly in their population towards the best food sources (solutions), when taking more neighbourhoods of dragonflies into account. Therefore, the compatible combination of the BBPSO and CS, which have the fastest convergence rate, and DA, which has the static and dynamic social behaviours, with surge in performance to achieve more mature iterations, benefits the efficient convergence speed and strong exploration and exploitation capabilities of the proposed algorithm.

In addition to the above convergence experiment, we then employ the first experimental setting, i.e. 90 unbalanced training and 90 unseen testing images, for testing. Moreover, we have compared the proposed methods with both embedded fixed and self-adaptive (or changing) *pa* in the CS to the baseline state-of-the-art PSO variants and meta-heuristic optimisation methods over 30 experiment runs. The SVM-based RBF kernel is employed for training the 90 unbalanced dataset and evaluating with the unseen 90 lymphocytic cell images. Table 5.1 depicts the average empirical results of each algorithm over 30 trials. In order to identify the efficiency of the feature optimisation processes, the classification result of employing the 80 raw features without using feature optimisation process is taken into account for comparison with other feature optimisation methods and illustrated in the last row of Table 5.1.

Table 5.1 The average classification performance of each optimisation algorithm utilising the fitness function 1, as defined in Eq (5.6), over 30 experiment runs and using 90 unseen testing images as well as the classification result employing the entire set of 80 raw features.

Methods	**Number of selected features**	**SVM (10-fold)**
GA	26-46	0.8115
PSO	20-43	0.8267
DA	22-41	0.8722
CS	19-41	0.8889
BBPSO	25-44	0.8981
ELPSO	27-46	0.8922
Binary BBPSO	35-58	0.9007
The proposed BBPSO (fixed *pa*)	**6-27**	**0.9393**
The proposed BBPSO (changing *pa*)	**9-33**	**0.9356**
80 raw features (entire set)	**-**	0.9089

As shown in Table 5.1, the proposed method embedded with either a fixed or a self-adaptive (or changing) *pa* achieves the highest average classification accuracy of 93.93% and 93.56%, respectively, and outperforms all comparable baseline methods. Moreover, the proposed BBPSO variant algorithm is able to converge within 100 to 120 iterations, on average, over

30 experiment runs with the number of selected feature subsets of 6-27, when employing a fixed *pa*, or 9-33, when using a self-adaptive *pa*, respectively; whereas the other comparable baseline methods reach convergence within 130 to 200 iterations with comparatively larger identified feature subsets, for example, 22-41 for the DA, 19-41 for the CS, 25-44 for the BBPSO, 27-46 for the ELPSO, 35-58 for the binary BBPSO, 20-43 for the PSO, and 26-46 for the GA. Overall, the proposed method compares favourably with the comparable baseline methods and outperforms the DA, CS, BBPSO, ELPSO, binary BBPSO, PSO and GA by 6.71%, 5.04%, 4.12%, 3.86%, 11.26% and 12.78%, respectively, on average of SVM classification accuracy over 30 experiment runs, when a fixed *pa* is employed, and by 6.34%, 4.67%, 3.75%, 4.34%, 3.49%, 10.89% and 12.41%, respectively, when embedded with a self-adaptive *pa*. In addition to comparison with the SVM classification of using original 80 raw features, the proposed method embedded with either a fixed or a self-adaptive *pa* outperforms that employing the 80 raw features without any feature selection process, whereas, under the same experimental setting, i.e. using 90 unbalanced training and 90 unseen testing datasets, the classification accuracies of all other baseline algorithms are comparable to, or sometimes lower than, the classification performance obtained utilising the original 80 raw features.

In terms of the clinical perspective, the important characteristics of lymphocytic cell images for ALL diagnosis include nucleus area, cytoplasm area, ratio of nucleus area to cytoplasm area, form factor and compactness (supporting the diagnosis in terms of an irregularity of cell shape in the nucleus region), perimeter, texture changes related to open or close of the chromatin pattern in the nucleus, eccentricity, etc. By inspection of the experimental results of the proposed method embedded with either a fixed or a self-adaptive *pa*, it indicates that the important characteristics of lymphocytic cell image, as mentioned previously, for ALL diagnosis are commonly included in the selected feature subsets. However, a few of the clinical important features of the lymphocytic cell image, as aforementioned, such as nucleus area and ratio of nucleus area to cytoplasm area, are often missed out or not co-existing in the selected feature subsets by the other comparable baseline optimisation methods; in fact, they sometimes select comparatively more features, which may be redundant for the training process of the classifier and cause the classification performance to decline.

Furthermore, we have conducted a boxplot diagram for comparison between the proposed method, embedded with either a fixed or a self-adaptive (or changing) *pa* and the comparable based-line optimisation methods, which depicts the details of SVM classification accuracy variations for all algorithms over 30 experiment runs for 90 unseen testing images, as illustrated in Figure 5.5.

In Figure 5.5, the first two boxplots from left-hand side represent the classification accuracy variations of the proposed algorithms embedded with a self-adaptive (or changing) and a fixed *pa* in the CS, respectively. Firstly, comparing the proposed method embedded with a fixed *pa* to other baseline methods, the proposed method achieves the highest average classification accuracy of 93.93% over 30 experiment runs. Moreover, 25% of the results of the proposed method, with the third quartile of 97%, are higher than the maximum accuracy results of the CS (with 96%), DA (with 94%), BBPSO (with 97%), binary BBPSO (with 94%), ELPSO (with 97%), PSO (with 96%) and GA (with 96%), respectively. In terms of the median values, the proposed method (with 95%) also outperforms those of the CS (with 91%), DA (with 90%), BBPSO (with 91%), binary BBPSO (with 91%), ELPSO (with 91%), PSO (with 83%) and GA (with 81%) and is different to the proposed method by 4%, 5%, 4%, 4%, 4%, 12% and 14%, respectively, over 30 experiment runs. Except for the outliners, the minimum accuracy of the proposed method, with a lower whisker of 89%, is higher than 25% of the results of the CS, DA, BBPSO, binary BBPSO and ELPSO and 75% of the results of the PSO and GA.

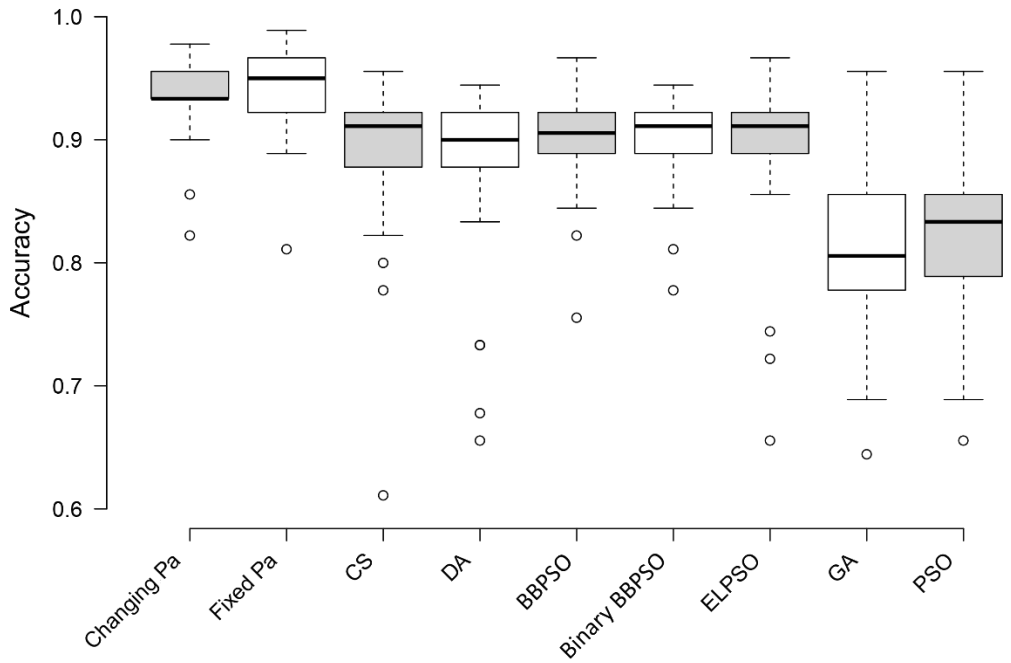


Figure 5.5 A boxplot diagram for each optimisation method integrated with SVM over 30 experiment runs for 90 unseen testing images employing the fitness function 1, as defined in Eq (5.6).

Secondly, when a self-adaptive (or changing) *pa* is embedded in the proposed algorithm, the boxplot also shows that our proposed algorithm achieves the highest average classification accuracy of 93.56% over 30 experiment runs; however, it has an average accuracy result slightly lower than the one embedded with a fixed *pa* due to the limited trials of initial setting of the *pa* value. In contrast, when we compare the accuracy result of the proposed algorithm embedded with a changing *pa* to all the comparable baseline methods, it has a better classification accuracy distribution with comparatively smaller variations, between 25% and 75% percentiles. Moreover, a group of 25% of the proposed method, with the third quartile of 96%, is also higher than the maximum classification accuracy results of the CS (with 96%), DA (with 94%), binary BBPSO (with 94%), PSO (with 96%) and GA (with 96%), respectively. In terms of the median values, the proposed algorithm (with 93%) also outperforms those of the PSO (with 83%) and GA (with 81%) by 10% and 12%, respectively, and is different to the proposed method by 2%-4% for all the other comparable baseline methods. Furthermore, the minimum classification accuracy result of the proposed algorithm, with a lower whisker of 90%, is higher than the 50% classification accuracy results of the DA, by at least 25% of the results of the CS, BBPSO, binary BBPSO and ELPSO and 75% of the results of PSO and GA. Overall, the proposed algorithms embedded with both a fixed and a self-adaptive (or changing) *pa* outperform all the comparable baseline optimisation algorithms greatly in the first experimental setting, i.e. 90 unbalanced training and 90 unseen testing images.

Next, we take the second experimental setting, i.e. 100 balanced training and 80 unseen testing images, into account for evaluation over 30 experiment runs of each of the algorithms. Table 5.2 illustrates the average SVM classification accuracy results of each method for evaluation of the 80 unseen testing images over 30 experiment runs, as well as the classification accuracy results obtained employing the original 80 raw features. As shown in Table 5.2, the proposed method, equipped with a fixed *pa*, achieves an average classification accuracy of 95.54%, whereas, when a self-adaptive (or changing) *pa* is employed in the proposed algorithm, it achieves the highest average classification accuracy of 95.88% over 30 experiment runs.

Moreover, the results in Table 5.2 indicate that the proposed method, embedded with either a fixed or a self-adaptive *pa*, outperforms all the comparable baseline optimisation algorithms and has the least number of selected feature subsets. Overall, in terms of the average classification accuracy results over 30 experiment runs, the proposed algorithm equipped with a fixed *pa* outperforms the PSO and GA by 8.35% and 11.5%, and all the other comparable methods by 1.04-2.41%, respectively, and the proposed algorithm embedded with a self-

adaptive (or changing) *pa* also outperforms the PSO and GA by 8.69% and 11.84%, and all other baseline optimisation methods by 1.38-2.75%, respectively.

Table 5.2 The average classification performance of each optimisation algorithm utilising the fitness function 1, as defined in Eq (5.6), over 30 experiment runs and using 80 unseen testing images as well as the classification result employing the entire set of 80 raw features.

Methods	**Number of selected features**	**SVM (10-fold)**
GA	27-45	0.8404
PSO	29-42	0.8719
DA	23-40	0.9329
CS	22-40	0.9450
BBPSO	26-38	0.9392
ELPSO	26-40	0.9313
Binary BBPSO	31-49	0.9358
The proposed BBPSO (fixed *pa*)	**10-26**	**0.9554**
The proposed BBPSO (changing *pa*)	**9-28**	**0.9588**
80 raw features (entire set)	**-**	0.9375

By inspecting the boxplot diagram in Figure 5.6, it indicates that 25% of the classification accuracy results of the proposed algorithm embedded with a fixed *pa*, with third quartile of 96%, are higher than the maximum classification accuracy results of the CS (with 96%), binary BBPSO (with 96%), PSO (with 94%), and GA (with 95%), respectively, whereas 25% of the accuracy results of the proposed method embedded with a self-adaptive (or changing) *pa*, with the third quartile of 97%, are also higher than the maximum classification accuracy results of all other comparable baseline methods.

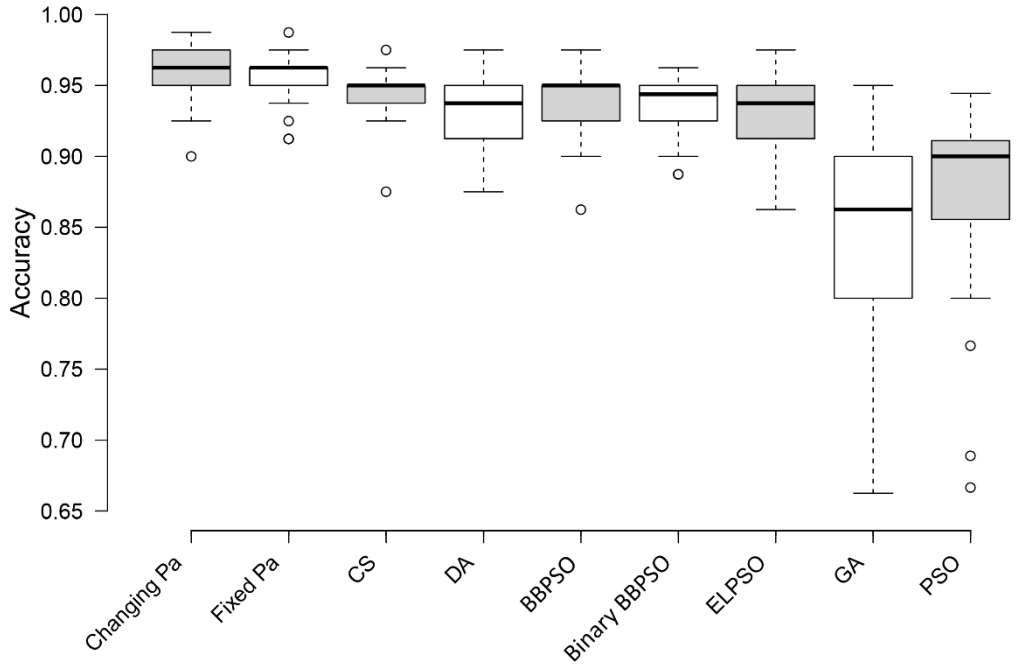


Figure 5.6 A boxplot diagram for each optimisation method integrated with SVM over 30 experiment runs for 80 unseen testing images employing the fitness function 1, as defined in Eq (5.6).

In addition, the proposed algorithm, embedded with either a fixed or a self-adaptive *pa*, has the same median values of 96%, which are 1%-2% higher than those of the CS, DA, BBPSO, binary BBPSO and ELPSO, and also higher than those of the PSO and GA by 6% and 10%, respectively, over 30 experiment runs. In terms of the minimum classification accuracy results, the proposed algorithm, equipped with either a fixed or a self-adaptive *pa*, is higher than at least 25% of the classification accuracy results of the DA, BBPSO, binary BBPSO, ELPSO, PSO and GA, whereas the proposed algorithm embedded with a fixed *pa* is also higher than 25% of the classification accuracy results of the CS. Overall, the proposed method, equipped with either a fixed or a self-adaptive (or changing) *pa*, also outperforms the other comparable baseline optimisation algorithms in the second experiment setting, i.e. 100 balanced training and 80 unseen testing lymphocytic cell images.

5.5.2 Experiment 2 employing Fitness Function 2

We also test the proposed algorithms using fitness function 2. First, we take the first experimental setting, i.e. 90 unbalanced training and 90 unseen testing images, into consideration for evaluating the proposed method embedded with either a fixed or a self-adaptive (or changing) *pa* and all comparable baseline optimisation methods over 30

experiment runs. The SVM-based RBF kernel is also employed for training the 90 unbalanced lymphocytic images and evaluating with the unseen 90 sample images. In addition, the classification accuracy result of employing the original 80 raw features without using feature selection process accounts for comparison with other feature optimisation methods and is depicted in the last row of Table 5.3.

As illustrated in Table 5.3, the proposed algorithm equipped with a self-adaptive (or changing) *pa* achieves a higher average classification accuracy of 93.78%, whereas the proposed method embedded with a fixed *pa* achieves an average classification accuracy of 93.26%. They also outperform all other comparable baseline feature optimisation methods. Moreover, the experimental results indicate that the proposed algorithm embedded with a changing *pa* outperforms the CS, DA, BBPSO, binary BBPSO, ELPSO, PSO and GA by 2.08%, 5.3 %, 3.82%, 4.97%, 5.71%, 8.37% and 11.41%, respectively, over 30 experiment runs, whereas the proposed algorithm equipped with a fixed *pa* outperforms the CS, DA, BBPSO, binary BBPSO, ELPSO, PSO and GA by 1.56%, 4.78%, 3.3%, 4.45%, 5.19%, 7.85% and 10.89%, respectively, over 30 trials.

Table 5.3 The average classification performance of each optimisation algorithm utilising the fitness function 2, as defined in Eq (5.21), over 30 experiment runs and using 90 unseen testing images as well as the classification result employing the entire set of 80 raw features.

Methods	**Number of selected features**	**SVM (10-fold)**
GA	28-46	0.8237
PSO	27-46	0.8541
DA	21-39	0.8848
CS	23-36	0.9170
BBPSO	25-40	0.8996
ELPSO	25-42	0.8807
Binary BBPSO	30-52	0.8881
The proposed BBPSO (fixed *pa*)	**10-28**	**0.9326**
The proposed BBPSO (changing *pa*)	**5-15**	**0.9378**
80 raw features (entire set)	-	0.9089

Furthermore, a boxplot diagram, as depicted in Figure 5.7, is obtained to investigate the classification accuracy result variations of all algorithms, when the fitness function 2 is employed. Figure 5.7 illustrates that 25% of the classification accuracy results of the proposed algorithm embedded with either a fixed or a self-adaptive (or changing) *pa*, with the same third quartile of 96%, are higher than the maximum classification performances of the DA (with 94%), BBPSO (with 96%), binary BBPSO (with 96%), ELPSO (with 96%), PSO (with 93%) and GA (with 93%). In terms of the median values, the proposed method, equipped with either a fixed or a self-adaptive *pa*, achieves with 94% and is higher than those median values of the CS (with 91%), DA (with 90%), BBPSO (with 92%), binary BBPSO (with 92%), ELPSO (wit 89%), PSO (with 88%) and GA (with 83%) by 3%, 4%, 2%, 4%, 5%, 6% and 11%, respectively, over 30 experiment runs. Moreover, a minimum classification accuracy of the proposed algorithm embedded with a fixed *pa*, with a lower whisker of 86%, is higher than 75% of the classification accuracy results of the GA and at least 25% accuracy results of the DA, binary BBPSO, ELPSO and PSO, whereas the minimum classification accuracy result of the proposed method equipped with a self-adaptive (or changing) *pa*, with a lower whisker of 88%, is higher than at least 75% of the accuracy results of the GA, 50% of accuracy results of the PSO and at least 25% of the accuracy results of the BBPSO, DA, binary BBPSO and ELPSO.

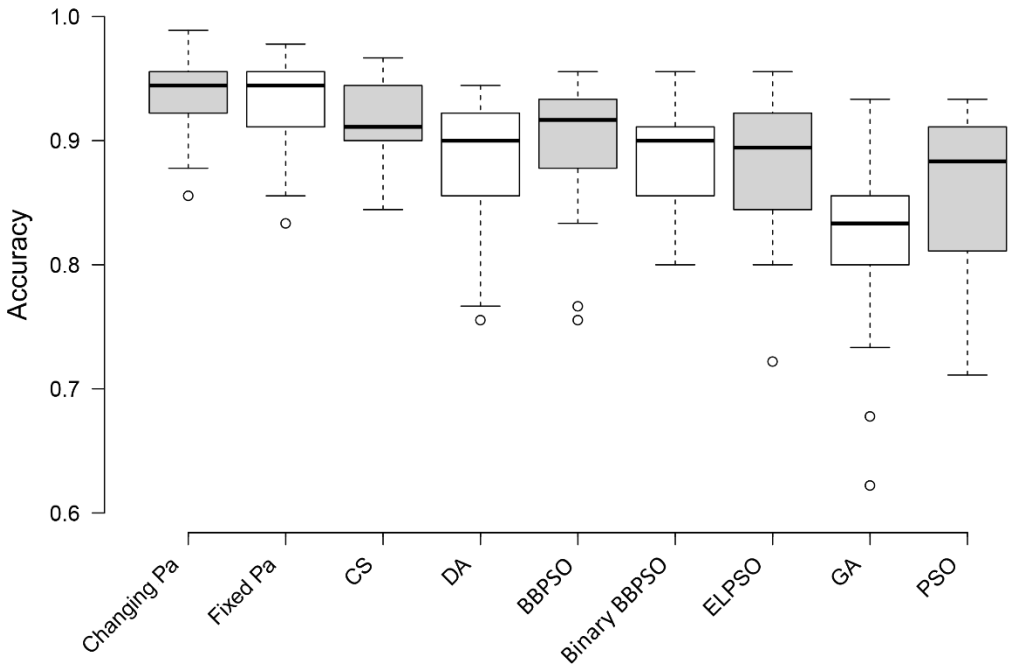


Figure 5.7 A boxplot diagram for each optimisation method integrated with SVM over 30 experiment runs for 90 unseen testing images employing the fitness function 2, as defined in Eq (5.21).

We have also taken the fitness function 2, as defined in Eq (5.21), and the second experimental setting, i.e. 100 balanced training and 80 unseen testing images, into account to test the efficiency of all feature optimisation algorithms. In addition, we also conduct the experiments of each algorithm for a fair comparison over 30 experiment runs as well as the previous experiments. The SVM with RBF kernel is also employed for training the 100 balanced lymphocytic images and evaluating with the unseen 80 sample images. Table 5.4 and Figure 5.8 illustrate the average classification accuracy of all feature optimisation methods and the classification accuracy result variations of all algorithms over 30 trials, respectively.

Table 5.4 The average classification performance of each optimisation algorithm utilising the fitness function 2, as defined in Eq (5.21), over 30 experiment runs and using 80 unseen testing images as well as the classification result employing the entire set of 80 raw features.

Methods	**Number of selected features**	**SVM (10-fold)**
GA	22-45	0.8654
PSO	23-38	0.8922
DA	20-37	0.9258
CS	20-38	0.9396
BBPSO	26-38	0.9354
ELPSO	24-37	0.9421
Binary BBPSO	27-44	0.9338
The proposed BBPSO (fixed *pa*)	**14-19**	**0.9604**
The proposed BBPSO (changing *pa*)	**12-17**	**0.9583**
80 raw features (entire set)	-	0.9375

Inspection of Table 5.4 indicates that the proposed algorithm embedded with a fixed *pa* achieves the highest average classification accuracy results of 96.04%, whereas the one embedded with a self-adaptive (or changing) *pa* achieves an average classification accuracy of 95.83% over 30 experiment runs. Moreover, the proposed algorithm, embedded with either a fixed or a self-adaptive *pa*, outperforms all the other comparable baseline optimisation methods greatly. The proposed method equipped with a fixed *pa* outperforms the PSO and

GA by 6.82% and 9.5%, respectively, and all other baseline methods by 1.83%-3.46%, whereas the proposed algorithm embedded with a self-adaptive *pa* outperforms the PSO and GA by 6.61% and 9.29%, respectively, and the other comparable feature optimisation algorithms by 1.62%-3.25%, respectively, on average classification accuracy results over 30 trials.

As depicted in Figure 5.8, a boxplot diagram for all algorithms employing the fitness function 2 reveals that the classification accuracy results of the proposed method, embedded with either a fixed or a self-adaptive (or changing) *pa*, with a third quartile of 97% and 96%, respectively, are higher than the maximum classification accuracy results of the binary BBPSO (with 96%), PSO (with 96%) and GA (with 96%). In terms of the median values, the proposed method (with the median values of 96%), embedded with either a fixed or a self-adaptive *pa*, is also higher than those of all other comparable based-line methods. In particular, the proposed method equipped with a self-adaptive (or changing) *pa* has the best classification accuracy distribution with comparatively smaller variations between the 25% and 75% percentiles, as compared to those from all other baseline optimisation methods. Furthermore, the minimum

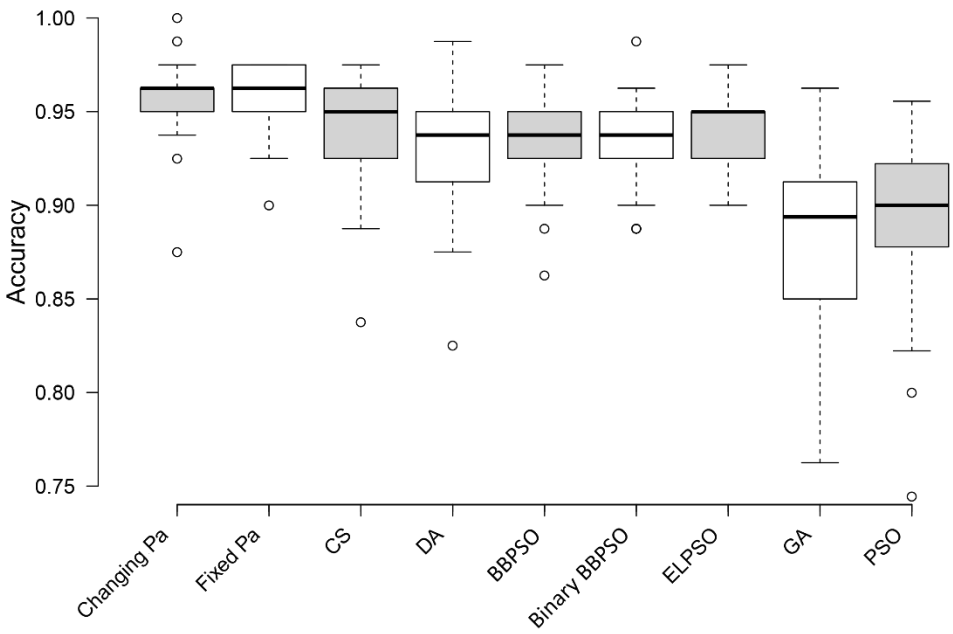


Figure 5.8 A boxplot diagram for each optimisation method integrated with SVM over 30 experiment runs for 80 unseen testing images employing the fitness function 2, as defined in Eq. (5.21).

classification accuracies of the proposed algorithm embedded with a self-adaptive *pa*, with a lower whisker of 94%, are higher than 50% of the classification accuracy results of the DA, BBPSO and binary BBPSO, with the same median values of 94%, and at least 25% of the accuracy results of the CS and ELPSO, with a first quartile of 93%, whereas the minimum classification accuracies of the proposed method equipped with a fixed *pa*, with a lower whisker of 93%, are higher than at least 25% of the classification accuracy results of the CS, DA, BBPSO, binary BBPSO and ELPSO. Moreover, the minimum classification accuracy results of the proposed method, embedded with either a fixed or a self-adaptive *pa*, are higher than 75% of the classification accuracy results of the PSO, with a third quartile of 92%, and the GA, with a third quartile of 91%.

By observation, the empirical results of the proposed method evaluated with the fitness function 2 utilising both experimental settings, i.e. 90 unbalanced and 100 balanced training datasets, further strengthen the superiority of the proposed algorithm. Overall, the proposed BBPSO variant algorithm, embedded with either a fixed or a self-adaptive *pa*, reveals great efficiency and outperforms all the other comparable baseline feature optimisation algorithms, i.e. state-of-the-art PSO variants, meta-heuristic and conventional optimisation algorithms, across the different experimental settings and under the different fitness function evaluations.

In comparing the proposed BBPSO variant algorithm to the other related studies in ALL diagnosis reported in the literature in Chapter 2, Section 2.8, to the best of our knowledge, Madhukar et al. (2012) and Putzu et al. (2014) have achieved high recognition performances employing the same ALL-IDB database. Putzu et al. (2014) achieved 93.2% classification accuracy using SVM with RBF kernel based on 10-fold cross validation and evaluated with 131 extracted features of the lymphocytic cell images, whereas Madhukar et al. (2012) obtained 93.5% classification accuracy employing SVM with leave-one-out cross validation and evaluated with a high dimension vector of shape, texture, and HD features of the nuclei extracted to distinguish normal and blast cells. The experimental results of the proposed BBPSO variant algorithm embedded with a fixed *pa* using SVM-based RBF kernel with 10-fold cross validation and evaluated with the 100 balanced training samples and testing with 80 unseen images under the fitness function 2, as defined in Eq (5.21), achieves an average classification accuracy of 96.04% over 30 experiment runs and identifies comparatively far fewer discriminative feature subsets for recognition of healthy and unhealthy lymphocytic cell images, whereas the proposed algorithm embedded with a self-adaptive *pa* employing the fitness function 1, as defined in Eq (5.6), achieves an average classification accuracy of 95.88% over 30 trials. In addition, the comparison results of the proposed algorithm are obtained by the average classification accuracy results over 30 experiment runs in each

experimental setting, i.e. 90 unbalanced training and 90 unseen testing images or 100 balanced training and 80 unseen testing samples. Therefore, the proposed BBPSO variant algorithm compares favourably with other related studies for ALL diagnosis reported in the literature and indicates the efficiency of the feature optimisation algorithm for acute lymphoblastic leukaemia detection.

Overall, in comparison to all other comparable baseline feature optimisation algorithms, the proposed method has the benefit of combining the three different search strategies to work in a collaborative manner. In particular, it embeds two operations, including the local random walk, which is a uniform combination, and the global random walk, which is Lévy flights, to diversify both primary and subswarm populations and to jump-out from local traps and to increase local exploitation and global exploration capabilities. Moreover, it integrates single swarm based BBPSO algorithm, and multi-subswarm based CS and DA algorithms, which work together under the compatible mechanisms to guide the search towards the optimal solutions.

5.6 Chapter Summary

This chapter has introduced the proposed evolutionary BBPSO variant algorithm for feature optimisation. The unique contribution in this chapter is a novel combination of the two complementary search algorithms, i.e. CS and DA algorithms, which are used to enhance and diversify the search behaviour of the original BBPSO algorithm, in an attempt to overcome the local optimum trap and guide the search toward the global optimal solution(s). The proposed algorithm enables both primary and subswarm-based searches employing the BBPSO, CS and DA searching algorithms and local and global random walk operations of a uniform combination and Lévy flights to work co-operatively to increase local exploitation and global exploration and mitigate (avoid) premature convergence problems of the conventional BBPSO. It was evaluated using the two experiment settings of a dataset of 180 lymphocytic images obtained in consultation with the haematologists, i.e. (i) 90 unbalanced training and 90 unseen testing images, and (ii) 100 balanced training and 80 testing samples. The proposed method embedded with either a fixed or a self-adaptive *pa* indicates great efficiency and greatly outperforms the comparable baseline optimisation algorithms, including state-of-the-art PSO variants, meta-heuristic and conventional optimisation algorithms, across different experimental settings under the two different fitness function evaluations.

For comparison with the other baseline optimisation algorithms, the proposed BBPSO variant algorithm, with both a fixed and a self-adaptive *pa*, identify comparatively fewer feature subsets with the fastest convergence rates and outperforms these comparable algorithms.

Furthermore, in comparison with the other related studies, the proposed method outperforms these comparable related research studies for ALL diagnosis reported in the literature by a significant margin, as illustrated in Section 5.5.

Overall, the best experimental results have been achieved, when the proposed BBPSO variant algorithm has been trained upon balanced 100 images and evaluated with 80 unseen samples employing SVM-based RBF kernel with 10-fold cross validation. In particular, if it is embedded with a fixed *pa*, it achieves a superior average classification accuracy of 96.04% under the fitness function defined in Eq (5.21), whereas embedded with a self-adaptive *pa* achieves an average classification accuracy of 95.88% under the fitness function defined in Eq (5.6) over 30 experiment runs, respectively. The empirical results indicate the efficiency of the proposed feature optimisation algorithm for ALL detection.

Chapter 6: Conclusion and Future Work

6.1. Introduction

In this chapter, brief summaries of each chapter of this research study are presented. The contributions of this research are also discussed, followed by identified limitations of the thesis. Moreover, recommendations for future investigations to overcome the deficiencies of this research study are also outlined.

6.2. Summary of This PhD Research

First, Chapter 1 described the relevant background of the multidisciplinary areas of this research study, including biomedical engineering, haematology and computer science. Information about cancer and recent incidence of cancer, in particular, blood cancer, or leukaemia, was also presented. Moreover, the model of screening or early state diagnosis of acute leukaemia of individuals from remote areas in order to receive full diagnosis at an advanced clinical laboratory for accurate diagnosis and appropriate treatments and therapies (Figure 1.1 of Chapter 1) was introduced. This is a crucial process of screening or early diagnosis which may lead to an increasing rate of survival among those cured of a severe illness such as acute leukaemia. Although modern hospitals and clinics have advanced laboratories with powerful medical equipment used in diagnosis of blood cancer/leukaemia, for the resource-limited regions there remain major barriers to such facilities. Therefore, microscopic examination of peripheral blood smear samples remains a necessary screening or early process for blood cancer, especially acute leukaemia. In this chapter, we also introduced the research problems and the motivation behind the decision to investigate and develop an intelligent decision support system for ALL detection using microscopic blood smear images. Then, the aims and objectives were explained, and brief details of the research contributions were described. Finally, the details and the structure of each chapter were also presented.

Chapter 2 reviewed the biological background of leukaemia, ALL, laboratory diagnosis of ALL and the limitations of traditional methods. Moreover, an image analysis on blood smear samples using computer technology and image processing was provided to indicate the benefit of using a quantitative microscopic for image analysis to reduce human operation error and assist the experts in diagnosis of ALL. This chapter also described the state-of-the-art of development for ALL detection, by organising the related literature review under five sequential processes: image segmentation; image separation of nucleus and cytoplasm of the

cell membrane images; feature extraction of cell nucleus and cell cytoplasm of the segmented cells; feature selection of the extracted descriptors to reduce the redundancy of the non-significant features; and ALL detection/classification. As observed in the related works, we found challenging tasks for the improvement of the above diagnosis stages for the quantitative analysis of ALL detection/classification. Finally, we provided the scope of this research study.

Chapter 3 introduced the first key stage of this research study, i.e. the segmentation of WBCs membranes, particularly lymphocyte and lymphoblast cells, using the proposed modified marker-controlled watershed algorithm integrated with the morphological operations using the microscopic sub-images of ALL-IDB2 database. The unique contribution of this chapter is a novel combination of existing techniques, i.e. watershed transform and morphological operations, and the proposed algorithm of generating the good markers for watershed transform to isolate the lymphocytic cell membranes with promising results. This stage focused on the isolation of lymphocyte/lymphoblast cell membrane from touching and overlapping of the RBCs, platelets and artefacts of the microscopic peripheral blood smear sub-images. Moreover, the overall system architecture of this PhD research was also introduced. In addition, the details of the microscopic peripheral blood smear image database, the ground truths and annotations of the lymphocytic images, and clinical diagnosis criteria of the ALL according to the consultations with the haematologists were described. In evaluation using the 180 lymphocytic sub-images from the ALL-IDB2 database and comparison between the proposed method and the traditional marker-controlled watershed transformation using the correlation coefficient, the proposed method using Gaussian low-pass filter achieves segmentation results with the highest correlation coefficient scores to the ground truth images of 0.9374. Furthermore, it is able to produce promising segmentation results of the whole lymphocytic cell membrane, including nucleus and cytoplasm. Therefore, the segmentation results of the proposed method are of benefit for the next step of blood cell image analysis.

Chapter 4 presented the second and third key stages of this research study. The second key stage is a novel SDM-based clustering algorithm with both within- and between-cluster scatter variances. It used to produce robust separation of the nucleus and cytoplasm of lymphocyte/lymphoblast cell images. Additionally, to overcome the limitation of the conventional FCM algorithm, the motivation and development of the proposed SDM algorithm were explained. The third key stage is concerned with the extraction of the eighty features consisting of shape, texture and colour information of the nucleus and cytoplasm sub-images. This chapter revealed the simulation and evaluation results of the SDM clustering compared with state-of-the-art clustering techniques reported in the literature. A number of classifiers (MLP, SVM and Dempster-Shafer ensemble) were employed for

lymphocyte/lymphoblast classification. Evaluated using the ALL-IDB2 database, the proposed SDM-based clustering overcomes the shortcomings of FCM, which focuses purely on within-cluster scatter variance. Additionally, it achieves the highest correlation coefficient scores for the separation of nucleus and cytoplasm and outperforms FCM, FCS and LDA. Finally, the overall system, as shown in Figure 4.1, achieves superior recognition rates of 96.72% and 96.67% accuracies using bootstrapping and 10-fold cross validation with Dempster-Shafer ensemble and SVM, respectively. This indicates the usefulness of the proposed SDM-based clustering method.

Chapter 5 introduced the fourth key stage of this thesis, the proposed BBPSO variant algorithm, to identify the most significant discriminative characteristics of healthy/lymphocyte and unhealthy/lymphoblast cell images to enable efficient ALL recognition. The unique contribution of this chapter is a novel hybridisation of the two complementary search algorithms, i.e. CS and DA algorithms, which are employed to enhance and diversify the search behaviour of the traditional BBPSO algorithm, in an attempt to overcome the local optimum trap and lead the search toward the global optimal solution(s). The proposed BBPSO-based feature optimisation with the two objective functions for fitness evaluation, and ALL identification using SVM classifier, were described. This chapter also revealed the simulation and evaluation results of the proposed BBPSO variant algorithm compared with state-of-the-art nature-inspired meta-heuristic algorithms reported in the literature. Evaluated using the ALL-IDB2 database, it achieves superior recognition accuracy of 95.88% and 96.04% using two different fitness evaluation strategies, respectively. Moreover, the proposed BBPSO variant algorithm outperforms the baseline state-of-the-art optimization algorithms and related research for ALL detection.

6.3. Summary Contribution to Knowledge of this Research

The achievements of this research study described in the above section enable us to make three contributions to the field of quantitative image analysis of ALL.

Contribution 1:

White blood cell membranes segmentation using a modified marker-controlled watershed method and morphological operations:

a. White blood cell membranes segmentation for microscopic blood smear sub-images, particularly lymphocyte (healthy lymphocyte cell) and lymphoblast (unhealthy lymphocyte cell) sub-images, using integration of the modified marker-controlled watershed method and morphological operations is

presented. This method can segment and identify WBC membrane from a noisy background sub-image, which is touching and overlapping with RBCs, to retrieve the original RGB pixels' colour of the identified cell membrane in the white background sub-image.

Contribution 2:

The separation of nucleus and cytoplasm of the identified lymphocyte and lymphoblast cell membrane using a novel SDM-based clustering technique and the feature extraction from the separated nucleus and cytoplasm cell images:

a. The novel clustering technique to separate nucleus and cytoplasm of lymphocytic (lymphocyte and lymphoblast) cell membrane images, namely SDM-based clustering, which takes both within- and between-cluster scatter variants into consideration, overcomes the limitation of the objective function of conventional Fuzzy C-mean (FCM) clustering, which focuses on only within-cluster scatter variance. It also outperforms other clustering methods, including Linear Discriminant Analysis (LDA) and Fuzzy Compactness and Separation (FCS) (Wu et al., 2005) for robust identification of cell nucleus and cell cytoplasm. This clustering technique can also produce robust results of the separation of nucleus and cytoplasm of the lymphocytic cell membrane images.

b. A total of 80 features, which include shape-based features, texture-based Gray Level Co-occurrence Matrix (GLCM) features, colour-based CIELAB colour space features, and the statistical measurement of these feature sets, is identified and used to discriminate healthy and unhealthy lymphocyte cells, as well as being used for ALL screening or an early detection system with image processing and artificial intelligent machine learning techniques.

c. Diverse single and ensemble classifiers are used in the experimental study for lymphocyte and lymphoblast detection. In this research study, Dempster-Shafer ensemble achieves the highest accuracy of 96.72% for bootstrap validation, whereas SVM with Gaussian Radial Basis Function kernel (RBF) achieves an accuracy of 96.67% for 10-fold cross validation.

Contribution 3:

The identification of the most significant discriminative characteristics of lymphocyte and lymphoblast cells to enable efficient ALL recognition using a proposed evolutionary Bare-Bones Particle Swarm Optimisation (BBPSO) variant algorithm:

a. The proposed BBPSO variant algorithm for feature selection incorporates the following search mechanisms, i.e. cuckoo search (CS), dragonfly algorithm (DA), convergence speed monitoring mechanisms, self-adaptive parameter settings and subswarm concepts to reduce the premature convergence problem of the conventional BBPSO.

b. The proposed algorithm incorporates BBPSO, CS and DA to diversify the primary and subswarm based search, respectively. An adaptive mechanism is also used to observe stagnant iterations and convergence degrees of each of the aforementioned search algorithms. The proposed algorithm employs Lévy flights and uniform combination to increase particle swarm diversity if the primary or subswarm based search stagnates. A self-adaptive discovery probability of the CS is also employed in the proposed method to further fine-tune solution vectors to overcome drawbacks of constant parameter setting in a traditional CS in order to further improve performance. Most importantly, the previous mentioned diverse search strategies, i.e. BBPSO, CS and DA, and local and global random walk operations, i.e. uniform combination and Lévy flights, work in a cooperative manner to increase local exploitation and global exploration and overcome the local optimum.

c. In comparison with advanced and classic nature-inspired and meta-heuristic algorithms, e.g. ELPSO, PSO, BPSO, Genetic Algorithm, CS, DA, etc., the proposed BBPSO-based feature optimisation algorithm has efficient discriminative capabilities in which the significant discriminating feature subsets for lymphocytes and lymphoblasts are revealed. Evaluated using the ALL-IDB2 dataset, the proposed algorithm, with either a fixed or a dynamically changing parameter setting, shows great efficiency and significantly outperforms all other baseline search algorithms across different experimental settings with two different fitness evaluations. It also compares favourably with related researches for ALL detection as reported in the literature.

6.4 Limitations and Future Work

Our research study has taken us on a long journey through the multidisciplinary areas involving biomedical engineering, haematology and computer science. In particular, we are specialising in the computer science discipline. Therefore, this project outcome is more in the computer science area. This research study is limited to ALL blood cancer, focusing on the lymphocyte and lymphoblast white blood cells. Additionally, the experimental results reported in this research study are evaluated using the ALL-IDB2 microscopic blood smear sub-images database (Labati et al., 2011b). Moreover, the ground truths and annotations about each of the microscopic sub-images are derived from both the publication of the ALL-IDB database and consultation with the haematologists in the Royal Victoria Infirmary (RVI Hospital at Newcastle-Upon-Tyne, United Kingdom). There are several possible directions for future investigation appearing from the implementation of this thesis, as follows:

First, regarding the segmentation of the WBCs cell membrane images (Chapter 3), the results of segmentation of the lymphocyte and lymphoblast cell membrane sub-images using the proposed method are promising, in which the segmented cell membrane includes nucleus and cytoplasm. However, an alternative technique that is interesting for future investigation to improve WBCs membrane segmentation is to use adaptive location and iteration (Liu, Cao, Zhao, & Chu, 2016) to identify the location of the WBC with adaptive adjustment and, then, an iterative GrabCut based on the dilation method could be employed for segmentation of the blood cell membrane.

Secondly, for the separation of cell nucleus and cell cytoplasm sub-images using SDM-based clustering algorithm (Chapter 4), since the SDM-based discriminant measure can be used as a fitness/cost function for different optimisation algorithms, the SDM-based clustering method with different optimisation algorithms, such as BBPSO, CS, Firefly Algorithm, DA, etc., are interesting to explore. Moreover, ensemble classifiers integrated with clustering techniques are also interesting for further investigation to detect the arrival of novel unseen classes, e.g. AML cell images, without prior training required (Farid et al., 2013; Neoh, Zhang, et al., 2015d) and it may help to reduce the difficulty in collecting huge samples of microscopic blood smear images to cover all types of leukaemia cells.

Thirdly, for the feature subsets optimisation using the BBPSO variant algorithm (Chapter 5), the proposed BBPSO variant algorithm contains diverse search strategies to improve its performance. In order to further release the burden of identifying the upfront optimal setting for the CS search strategy in the proposed method, a self-adaptive step-size parameter, α, together with a self-adaptive *pa*, is of interest to employ to further fine-tune solution vectors

in order to improve performance of the proposed algorithm. Moreover, to date, we have employed a single objective fitness function for the two different fitness evaluation strategies to search for the global best solution(s). However, challenging real-world optimisation problems tend to have multiple constraints and competing objectives, such as computational cost, the number of selected feature subset, convergence speed, swarm diversity, etc. Consequently, multiple criteria decision-making is required. The multi-objective evolutionary algorithms, such as multi-objective PSO (Coello Coello & Lechuga, 2002) and CS (Yang & Deb, 2013), nondominated sorting PSO (NSPSO) (Yang Liu, 2008), Strength Pareto Evolutionary Algorithm2 (SPEA2) (Zitzler, Laumanns, & Thiele, 2001), and Pareto Achieved Evolutionary Algorithms (PAES) (Knowles & Corne, 1999), etc., are interesting to explore to further improve performance of the proposed algorithm.

Finally, the other possible application of this research study could be applied in the field of computerised-aid technology for health care disease/cancer screening from microscopic images. For example, the early screening of cervical cancer using Pap smear images. Cervical cancer is the second leading cause of cancer death in females across the globe (Torre et al., 2015). Most affected group are younger women in many countries, including Europe, Central Asia, Japan, and China (Bray et al., 2013; Vaccarella et al., 2013). The screening program can efficiently reduce the mortality rates of cervical cancer. Torre et al. (2015) reported that in many Western countries, where the long-time existed screening programs have been used, the rates of cervical cancer have decreased by almost 65% over the past 40 years. The individuals can be cured by early detection or diagnosed in the pre-cancerous lesion stage. Papanicolaou test or Pap test is a widely used physical examination technique to prevent cervical cancer by finding cells that either reveal the significant characteristics to indicate the cancer or have the high possibility to turn cancerous. Therefore, the highly accurate automated intelligent screening systems for cervical cancer can be used as an aid-tool for the experts' decision of helping suspected individuals to have more chance to be cured and live longer.

Bibliography

Abbas, N., & Mohamad, D. (2014). Automatic color nuclei segmentation of leukocytes for acute Leukemia. *Research Journal of Applied Sciences, Engineering and Technology*, *7*(14), 2987–2993.

Abbott. (2008). *Atlas of Pediatric Peripheral Blood Smears*.

Abdul Nasir, A. S., Mashor, M. Y., & Rosline, H. (2011). Unsupervised colour segmentation of white blood cell for acute leukaemia images. In *The IEEE International Conference on Imaging Systems and Techniques*. Penang, Malaysia: IEEE. http://doi.org/10.1109/IST.2011.5962188

Abdul-Hamid, G. (2011). Classification of Acute Leukemia. In M. Antica (Ed.), *Acute Leukemia - The Scientist's Perspective and Challenge* (p. 428). InTech. http://doi.org/10.5772/19848

Agaian, S., Madhukar, M., & Chronopoulos, A. T. (2014). Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Images. *Ieee Systems Journal*, *8*(3), 995–1004. http://doi.org/10.1109/JSYST.2014.2308452

Agrawal, S., & Deardean, C. (2014). Chronic lymphocytic leukaemia (CLL). Retrieved May 9, 2016, from https://bloodwise.org.uk/sites/default/files/documents/CLL_patient_info_booklet.pdf

Albitar, M., Giles, F. J., & Kantarjian, H. (2008). Acute Leukemias (pp. 119–130). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-72304-2_8

American Cancer Society. (2015). *Global Cancer Facts & Figures 3rd Edition*. *American Cancer Society*. Atlanta. Retrieved from http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-044738.pdf

American Cancer Society. (2016a). *Childhood Leukemia*. Retrieved from http://www.cancer.org/acs/groups/cid/documents/webcontent/003095-pdf.pdf

American Cancer Society. (2016b). *Leukemia-Acute Lymphocytic (Adults)*. Retrieved from http://www.cancer.org/acs/groups/cid/documents/webcontent/003109-pdf.pdf

Amnis corporation. (2010). *IDEAS® image data exploration and analysis software user's manual*. Seattle, WA.

Amoda, N., & Kulkarni, R. K. (2013). Image Segmentation and Detection using Watershed Transform and Region Based Image Retrieval. *International Journal of Electronics Communication and Computer Engineering*, *2*, 89–94. http://doi.org/10.1.1.24.5229

Andrews, J. M., Holm, M. T., & Myers, J. B. (2005). Clinical Predictors of Abnormal Peripheral Blood Lymphocytoses Diagnosed by Flow Cytometry: An Algorithmic Approach That Can Be Applied in Routine Clinical Practice. *Blood*, *106*(11), 3932–3932.

Apperley, J. (2015). Chronic myeloid leukaemia (CML). Retrieved May 9, 2016, from https://bloodwise.org.uk/sites/default/files/documents/CML_patient_info_booklet.pdf

Argyle, J. C., Benjamin, D. R., Lampkin, B., & Hammond, D. (1989). Acute Nonlymphocytic Leukemias of childhood. inter-observer variability and problems in the use of the fab classification. *Cancer*, *63*(2), 295–301.

Bäck, T., & Schwefel, H.-P. (1993). An Overview of Evolutionary Algorithms for Parameter Optimization. *Evolutionary Computation*, *1*(1), 1–23. http://doi.org/10.1162/evco.1993.1.1.1

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support Vector Regression. *Neuronal Information Processing - Letters and Reviews*, *11*(10), 203–224. http://doi.org/10.4258/hir.2010.16.4.224

Beni, G., & Wang, J. (1993). Swarm Intelligence in Cellular Robotic Systems. In P. Dario, G. Sandini, & P. Aebischer (Eds.), *Robots and Biological Systems: Towards a New Bionics?* (NATO ASI S, p. pp 703–712). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-58069-7_38

Bennett, J. M., Catovsky, D., Daniel, M. T., Flandrin, G., Galton, D. A., Gralnick, H. R., & 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. http://doi.org/10.1111/j.1365-2141.1976.tb03563.x

Beucher, S., & Meyer, F. (1992). The morphological approach to segmentation: the watershed transformation. In E. R. Dougherty (Ed.), *Mathematical Morphology in image processing, Optical Science and Engineering* (pp. 433–481). New York: Marcel Dekker Inc. http://doi.org/Export Date 6 May 2013

Bezdex, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, *34*(3), 483–519. http://doi.org/10.1007/s10115-012-0487-8

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. New York: Oxford University Press.

Bray, F., Lortet-Tieulent, J., Znaor, A., Brotons, M., Poljak, M., & Arbyn, M. (2013). Patterns and trends in human papillomavirus-related diseases in Central and Eastern Europe and Central Asia. *Vaccine*, *31*, H32-H45. https://doi.org/10.1016/j.vaccine.2013.02.071

Breslauer, D. N., Maamari, R. N., Switz, N. A., Lam, W. A., & Fletcher, D. A. (2009). Mobile Phone Based Clinical Microscopy for Global Health Applications. *PLoS ONE*, *4*(7), 1–7. http://doi.org/10.1371/journal.pone.0006320

Buavirat, S., & Srisa-an, C. (2008). Classification for Acute Lymphocytic Leukemia using Feature Extraction and Neural Networks in White Blood Cell stained images. In *The 3rd International Symposium on Biomedical Engineering (ISBME 2008)* (pp. 1–4). Bangkok, Thailand.

Campbell, K. (2011). Understanding blood cancers: What is blood cancer? Retrieved December 27, 2015, from https://bloodwise.org.uk/sites/default/files/documents/Young_adults_with_a_blood_cancer_what_do_I_need_to_know.pdf

CancerResearchUK. (2008). Press release: Scientists predict three quarters of children with leukaemia will be cured. Retrieved December 27, 2015, from http://www.cancerresearchuk.org/about-us/cancer-news/press-release/2008-07-02-scientists-predict-three-quarters-of-children-with-leukaemia-will-be-cured

CancerResearchUK. (2015a). Leukaemia (all subtypes combined) statistics. Retrieved December 27, 2015, from http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia#heading-One

CancerResearchUK. (2015b). Screening for acute lymphoblastic leukaemia. Retrieved December 28, 2015, from http://www.cancerresearchuk.org/about-cancer/type/all/about/screening-for-acute-lymphoblastic-leukaemia

CancerResearchUK. (2015c). Screening for acute myeloid leukaemia. Retrieved December 28, 2015, from http://www.cancerresearchuk.org/about-cancer/type/aml/about/screening-for-acute-myeloid-leukaemia

CancerResearchUK. (2016a). Cancer Statistics for the UK. Retrieved June 13, 2016, from http://www.cancerresearchuk.org/health-professional/cancer-statistics

CancerResearchUK. (2016b). Causes of cancer and reducing your risk. Retrieved June 13, 2016, from http://www.cancerresearchuk.org/about-cancer/causes-of-cancer

CancerResearchUK. (2017). Acute Lymphoblastic Leukaemia (C91.0):2012-2014 Average Number of New Cases per Year and Age-Specific Incidence Rates per 100,000 Population, UK. Retrieved May 26, 2017, from http://www.cancerresearchuk.org/sites/default/files/cstream-node/cases_crude_leuka_all_I14.pdf

Cason, J. D., Trujillo, J. M., Estey, E. H., Huh, Y. O., Freireich, E. J., & Stass, S. A. (1989). Peripheral acute leukemia: high peripheral but low-marrow blast count. *Blood*, *74*(5), 1758-1761.

Chamberlain, J., & Moss, S. (1996). *Evaluation of Cancer Screening -(Focus on Cancer Series)*. (C. Jocelyn & M. Susan, Eds.). Springer-Verlag London.

Chih-Wei Hsu, Chih-Chung Chang, and C.-J. L. (2008). A Practical Guide to Support Vector Classification. *BJU International*, *101*(1), 1396–400. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Cleveland Clinic. (2016). Benign Hematology. Retrieved May 8, 2016, from https://my.clevelandclinic.org/health/diseases_conditions/benign-hematology-overview

Coello Coello, C. A., & Lechuga, M. S. (2002). MOPSO: A proposal for multiple objective particle swarm optimization. In *Proceedings of the 2002 Congress on Evolutionary Computation, CEC 2002* (Vol. 2, pp. 1051–1056). http://doi.org/10.1109/CEC.2002.1004388

Cooper, L. A. D., Carter, A. B., Farris, A. B., Wang, F., Kong, J., Gutman, D. A., … Saltz, J. H. (2012). Digital pathology: Data-intensive frontier in medical imaging. *Proceedings of the IEEE*, *100*(4), 991–1003. http://doi.org/10.1109/JPROC.2011.2182074

Craig, F. E., & Foon, K. A. (2008). Flow cytometric immunophenotyping for hematologic neoplasms. *Blood*, *111*(8), 3941–3967. http://doi.org/https://doi.org/10.1182/blood-2007-11-120535

Das, D. K., Chakraborty, C., Mitra, B., Maiti, A. K., & Ray, A. K. (2013). Quantitative microscopy approach for shape-based erythrocytes characterization in anaemia. *Journal of Microscopy*, *249*(2), 136–149. http://doi.org/10.1111/jmi.12002

de la Iglesia, B. (2013). Evolutionary computation for feature selection in classification problems. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, *3*(6), 381–407. http://doi.org/Doi 10.1002/Widm.1106

Ding, S., & Chen, L. (2010). Intelligent Optimization Methods for High-Dimensional Data Classification for Support Vector Machines. *Intelligent Information Management*, *02*, 159–169. http://doi.org/10.4236/iim.2010.26043

Dorini, L. B., Minetto, R., & Leite, N. J. (2007). White blood cell segmentation using morphological operators and scale-space analysis. In *XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)* (pp. 294–304). IEEE. http://doi.org/10.1109/SIBGRAPI.2007.33

Dougherty, G. (2009). *Digital Image Processing for Medical Applications* (1st ed.). New York: Cambridge University Press.

Eiben, A. E., & Rudolph, G. (1999). Theory of Evolutionary Algorithms: A Bird's Eye View. *Theoretical Computer Science*, *229*(1), 3–9.

El Rassi, F., Little, B. P., Holloway, S., Roberts, D., & Khoury, H. J. (2012). Early diagnosis of acute myeloid leukemia by computed tomography scan. *J Clin Oncol*, *30*(23), e207–8. http://doi.org/10.1200/JCO.2011.41.0506

Elsheikh, T. M., Asa, S. L., Chan, J. K. C., DeLellis, R. A., Heffess, C. S., LiVolsi, V. A., & Wenig, B. M. (2008). Interobserver and intraobserver variation among experts in the diagnosis of thyroid follicular lesions with borderline nuclear features of papillary carcinoma. *American Journal of Clinical Pathology*, *130*(5), 736–744. http://doi.org/10.1309/AJCPKP2QUVN4RCCP

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. *Artificial Intelligence in Medicine*, *55*(3), 163–175. http://doi.org/10.1016/j.artmed.2012.03.005

F. Rodak, B., & H. Carr, J. (2012). *Clinical Hematology Atlas* (4th ed.). Elsevier Health Sciences.

Farid, D. M., Zhang, L., Hossain, A., Rahman, C. M., Strachan, R., Sexton, G., & Dahal, K. (2013). An adaptive ensemble classifier for mining concept drifting data streams. *Expert Systems with Applications*, *40*(15), 5895–5906. http://doi.org/10.1016/j.eswa.2013.05.001

Fatma, M. (2014). Leukemia Image Segmentation using K-Means Clustering and HSI Color Image Segmentation. *Ijca*, *94*(12), 6–9.

Fukuyama, Y., & Sugeno, M. (1989). A new method of choosing the number of clusters for the fuzzy c-means method. In *The 5th Fuzzy System Symposium* (pp. 247–250).

Gautam, A., Bhadauria, H. S., & Singh, A. (2014). White Blood Nucleus Segmentation Using an Automated Thresholding and Mathematical Morphing. In *The international conference on Advances in Engineering and Technology* (pp. 31–35). Roorkee, India.

Gonzalez, M. A., & Ballarin, V. L. (2009). Automatic marker determination algorithm for watershed segmentation using clustering. *Latin American Applied Research*, *39*, 225–229.

Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2004). *Digital Image Processing Using MATLAB*. New York: Prentice Hall.

Hakli, H., & Uğuz, H. (2013). Levy flight distribution for scout bee in artificial bee colony algorithm. In *Lecture Notes on Software Engineering* (Vol. 1, pp. 254–258). http://doi.org/10.7763/LNSE.2013.V1.55

Halim, N. H. A., Mashor, M. Y., Abdul Nasir, A. S., Mokhtar, N. R., & Rosline, H. (2011). Nucleus segmentation technique for acute leukemia. *Proceedings - 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications, CSPA 2011*, 192–197. http://doi.org/10.1109/CSPA.2011.5759871

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques* (3rd Ed.). Morgan Kaufmann.

Harris, N. L., Jaffe, E. S., Diebold, J., Flandrin, G., Muller-Hermelink, H. K., Vardiman, J., … Bloomfield, C. D. (2000). 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. *Histopathology*, *36*, 69–89.

Held, P., & Banks, P. (2013). *Analysis of Nuclear Stained Cells- Using the Cytation3 Cell Imaging Multi-Mode Microplate Reader with DAPI-Stained Cells*. Vermont, USA. Retrieved from http://e3f99a4a3891332df177-b511c07cae915b9d1958179eaa4432ea.r82.cf1.rackcdn.com/assets/tech_resources/Cytation3_Nuclear_Staining_App_Note.pdf

Hough, R. (2015a). Acute Lymphoblastic Leukaemia (ALL) in Children and Young Adults up to 16 Years. Retrieved September 13, 2016, from https://bloodwise.org.uk/sites/default/files/documents/chALL_patient_info_booklet.pdf

Hough, R. (2015b). Acute myeloid leukaemia (AML) in Children and Young Adults up to 16 years. Retrieved May 9, 2016, from https://bloodwise.org.uk/sites/default/files/documents/ChAML_patient_info_booklet.pdf

Howard, D., & Mark, B. (1998). Neural Network Toolbox: For use with the MATLAB. Natick, Massachusetts, United States.: The MathWorks, Inc.

Huang, D., & Hung, K.-D. (2012). Leukocyte Nucleus Segmentation and Recognition in Color Blood-smear Images. In *IEEE International Conference of Instrumentation and Measurement Technology* (pp. 171–176). http://doi.org/10.1109/I2MTC.2012.6229443

J. Bain, B. (2004). *A Beginner's Guide to Blood Cells* (2nd ed.). Blackwell Publishing.

J. Bain, B. (2010). *Leukaemia Diagnosis* (4th ed.). Blackwell Publishing.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37. http://doi.org/10.1109/34.824819

Jain, A. K., Murty, M. N., & Flynn, P. J. (2000). Data Clustering : A Review. *ACM Computing Surveys*, *31*(3), 60.

Jemal, A., Bray, F., & Ferlay, J. (2011). Global Cancer Statistics: 2011. *CA Cancer J Clin*, *61*(2), 69–90. http://doi.org/10.3322/caac.20107.

Jiang, K., Liao, Q.-M., & Dai, S.-Y. (2003). A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering. In *The IEEE 2003 International Conference on Machine Learning and Cybernetics* (Vol. 5, pp. 2–7). http://doi.org/10.1109/ICMLC.2003.1260033

Jordehi, A. R. (2015). Enhanced leader PSO (ELPSO): A new PSO variant for solving global optimisation problems. *Applied Soft Computing*, *26*, 401–417. http://doi.org/10.1016/j.asoc.2014.10.026

Joshi, M. M. D., Karode, P. A. H., & Suralkar, P. S. R. (2013). White Blood Cells Segmentation and Classification to Detect Acute Leukemia. *International Journal of Emerging Trends & Technology in Computer Science*, *2*, 147–151.

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, *1*(4), 321–331. http://doi.org/10.1007/BF00133570

Kebriaei, P., Anastasi, J., & Larson, R. A. (2002). Acute lymphoblastic leukaemia: diagnosis and classification. *Best Pract Res Clin Haematol*, *15*(4), 597–621. https://doi.org/10.1053/beha.2002.0224

Kennedy, J. (2003). Bare bones particle swarms. In *Proceedings of the 2003 IEEE Swarm Intelligence Symposium* (pp. 80–87). http://doi.org/10.1109/SIS.2003.1202251

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *IEEE International Conference on Neural Networks* (Vol. 4, pp. 1942–1948). http://doi.org/10.1109/ICNN.1995.488968

Khashman, A., & Abbas, H. H. (2013). Acute Lymphoblastic Leukemia Identification Using Blood SmearImages and A Neural Classifier. In *Advances in computational intelligence: 12th international conference on Artificial Neural Networks (IWANN 2013), Part II, LNCS 7903* (pp. 80–87). Springer Berlin Heidelberg.

Knowles, J., & Corne, D. (1999). The Pareto archived evolution strategy: A new baseline algorithm for Pareto multiobjective optimisation. In *Proceedings of the 1999 Congress on Evolutionary Computation, CEC 1999* (Vol. 1, pp. 98–105). http://doi.org/10.1109/CEC.1999.781913

Kulkarni, T. A., Bhosale, D. S., & Yadav, D. M. (2014). A Fast Segmentation Method for the Recognition of Acute Lymphoblastic Leukemia using Thresholding Algorithm. *International Journal of Electronics Communication and Computer Engineering*, *5*(4), 364–368.

Kumar, B. R., Joseph, D. K., & Sreenivas, T. V. (2002). Teager energy based blood cell segmentation. In *The 14th International Conference on Digital Signal Processing, 2002.* (pp. 619–622). Hellas, Greece: IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1028167

Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. *Combining Pattern Classifiers: Methods and Algorithms*. New Jersey: John Wiley & Sons. http://doi.org/10.1002/0471660264

Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms*. *Technometrics* (2nd Ed., Vol. 47). New Jersey: John Wiley & Sons. http://doi.org/10.1002/9781118914564

Kuo, B.-C., & Landgrebe, D. A. (2004). Nonparametric weighted feature extraction for classification. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(5), 1096–1105. http://doi.org/10.1109/TGRS.2004.825578

Labati, R. D., Piuri, V., & Scotti, F. (2011a). ALL-IDB : The acute lymphoblastic leukemia image database for image processing. In *18th IEEE International Conference on Image Processing* (pp. 2045–2048). Brussels, Belgium. http://doi.org/10.1109/ICIP.2011.6115881

Labati, R. D., Piuri, V., & Scotti, F. (2011b). ALL-IDB web site. Retrieved February 10, 2013, from http://homes.di.unimi.it/scotti/all/

Lakshmi, S., & Sankaranarayanan, D. V. (2010). A study of Edge Detection Techniques for Segmentation Computing Approaches. *International Journal of Computer Applications Special Issue on Computer Aided Soft Computing Techniques for Imaging and Biomedical Applications*, (1), 35–41. http://doi.org/10.5120/993-25

Leonard, B. (1993). *Leukemia: A Research Report*. (B. Leonard, Ed.). DIANE. Retrieved from https://books.google.co.uk/books/about/Leukemia.html?id=VfFCVvX9btYC&redir_esc=y

Lévy, P. (1954). *Théorie de l'Addition des Variables Aléatoires* (second). Paris: Gauthier-Villars.

Lewis, G., Sheringham, J., Lopez Bernal, J., & Crayford, T. (2014). *Mastering Public Health: A Postgraduate Guide to Examinations and Revalidation* (2nd ed.). CRC Press.

Li, C. H., Kuo, B. C., & Lin, C. T. (2011). LDA-Based Clustering Algorithm and Its Application to an Unsupervised Feature Extraction. *IEEE Transactions on Fuzzy Systems*, *19*(1), 152–163. http://doi.org/10.1109/TFUZZ.2010.2089631

Liao, Q., & Deng, Y. (2002). an Accurate Segmentation Method. In *IEEE International Symposium on Biomedical Imaging Proceedings* (pp. 245–248). IEEE.

Liu, Y. (2008). A fast and elitist multi-objective particle swarm algorithm: NSPSO. In *2008 IEEE International Conference on Granular Computing, GRC 2008* (pp. 470–475). http://doi.org/10.1109/GRC.2008.4664711

Liu, Y., Cao, F., Zhao, J., & Chu, J. (2016). Segmentation of White Blood Cells Image Using Adaptive Location and Iteration. *IEEE Journal of Biomedical and Health Informatics*, *X*(X), 1–12. http://doi.org/10.1109/JBHI.2016.2623421

Ljouad, T., Amine, A., & Rziza, M. (2014). A hybrid mobile object tracker based on the modified Cuckoo Search algorithm and the Kalman Filter. *Pattern Recognition*, *47*, 3597–3613. http://doi.org/10.1016/j.patcog.2014.04.003

Logan, A. C., Wang, C., Sahaf, B., Jones, C. D., Marshall, E. L., Buno, I., … Miklos, D. B. (2010). High-Throughput VDJ Sequencing Is Superior to Quantitative PCR and Flow Cytometry for the Quantification of Minimal Residual Disease In Chronic Lymphocytic Leukemia After Hematopoietic Cell Transplantation. *Blood*, *116*(21), 1290–1290.

Madhloom, H. T., Kareem, S. A., & Ariffin, H. (2012a). A robust feature extraction and selection method for the recognition of lymphocytes versus acute lymphoblastic leukemia. In *International Conference on Advanced Computer Science Applications and Technologies (ACSAT)* (pp. 330–335). http://doi.org/10.1109/ACSAT.2012.62

Madhloom, H. T., Kareem, S. A., & Ariffin, H. (2012b). An image processing application for the localization and segmentation of lymphoblast cell using peripheral blood images. *Journal of Medical Systems*, *36*, 2149–2158. http://doi.org/10.1007/s10916-011-9679-0

Madhukar, M., Agaian, S., & Chronopoulos, A. T. (2012). New decision support tool for acute lymphoblastic leukemia classification. In *The international conference on Image Processing: Algorithms and System X; and Parallel Processing for Imaging Application II* (pp. 829518–1 – 829518–12). http://doi.org/10.1117/12.905969

Mallath, M. K., Taylor, D. G., Badwe, R. A., Rath, G. K., Shanta, V., Pramesh, C. S., … Sullivan, R. (2014). The growing burden of cancer in India: Epidemiology and social context. *The Lancet Oncology*, *15*(6), e205–e212. http://doi.org/10.1016/S1470-2045(14)70115-9

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *An Introduction to Information Retrieval* (1st Ed.). Cambridge University Press.

Marques, O. (2011). *Practical Image and Video Processing Using MATLAB*. Hoboken, New Jersey: John Wiley & Sons.

MathWorks. (2016). RGB2GRAY. Retrieved August 15, 2016, from http://uk.mathworks.com/help/matlab/ref/rgb2gray.html?searchHighlight=rgb2gray

Meer, W. van der, Gelder, W. van, Keijzer, R. de, & Willems, H. (2007). The divergent morphological classification of variant lymphocytes in blood smears. *Journal of Clinical Pathology*, *60*(7), 837–838. http://doi.org/10.1136/jcp.2005.033787

Meera, V., & Mathew, S. A. (2014). Fuzzy Local Information C Means Clustering For Acute Myelogenous Leukemia Image Segmentation. In *The International Conference on Innovation and Advances in Science, Engineering and Technology* (Vol. 3, pp. 61–68). Arakunnam, Karela, India.

Mirjalili, S. (2015). Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. *Neural Computing and Applications*, 1–21. http://doi.org/10.1007/s00521-015-1920-1

Mohapatra, S., & Patra, D. (2010). Automated cell nucleus segmentation and acute leukemia detection in blood microscopic images. In *IEEE International Conference on Systems in Medicine and Biology* (pp. 49–54). Kharagpur, India. http://doi.org/10.1109/ICSMB.2010.5735344

Mohapatra, S., Patra, D., & Kumar, K. (2012). Unsupervised Leukocyte Image Segmentation Using Rough Fuzzy Clustering. *International Scholary Reserch ISRN Artificial Intelligence*, *2012*, 1–12. http://doi.org/10.5402/2012/923946

Mohapatra, S., Patra, D., Kumar, S., & Satpathi, S. (2012). Kernel Induced Rough c-means clustering for lymphocyte image segmentation. In *4th International Conference on Intelligent Human Computer Interaction: Advancing Technology for Humanity* (pp. 1–6). Kharagpur, India. http://doi.org/10.1109/IHCI.2012.6481865

Mohapatra, S., Patra, D., & Satpathi, S. (2010). Image analysis of blood microscopic images for acute leukemia detection. In *IEEE International Conference on Industrial Electronics, Control and Robotics* (pp. 215–219). http://doi.org/10.1109/IECR.2010.5720171

Mohapatra, S., Patra, D., & Satpathy, S. (2012). Unsupervised Blood Microscopic Image Segmentation and Leukemia Detection using Color based Clustering. *International Journal of Computer Information Systems and Industrial Management Applications*, *4*, 477–485.

Mohapatra, S., Patra, D., & Satpathy, S. (2014). An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Computing and Applications*, *24*(7-8), 1887–1904. http://doi.org/10.1007/s00521-013-1438-3

Mondal, P. K., Prodhan, U. K., Al Mamun, M. S., Rahim, M. A., & Hossain, K. K. (2014). Segmentation of White Blood Cells Using Fuzzy C Means Segmentation Algorithm. *IOSR Journal of Computer Engineering*, *16*(3), 01–05. Retrieved from http://www.iosrjournals.org/iosr-jce/papers/Vol16-issue3/Version-9/A016390105.pdf

Moor, Gary, D., Blann, A. D., & Knight, G. (2010). *Haematology* (1st ed.). Oxford University Press.

Muslimani, A., Kizilbash, S., Jaiyesimi, I., Nadeau, L., Zakalik, D., Margolis, J., & Huang, J. (2010). Evaluation of the Diagnostic Utility of Cerebral Spinal Fluid (CSF) Flow Cytometry (FC) In Detection of Central Nervous System (CNS) Involvement by Hematological Malignancy. *Blood*, *116*(21), 3837.

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 Journal of Information Technology*, *10*(4), 356–364. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84865030323&partnerID=40&md5=1625181c449e93b48c404f3efefddb24

National Cancer Institute Thailand. (2015). *Cancer in Thailand 2010-2012*. (W. Imsamran, A. Chaiwerawattana, S. Wiangnon, D. Pongnikorn, K. Suwanrungruang, S. Sangrajrang, & R. Buasom, Eds.) (Vol. 8). Bangkok, Thailand: New Thammada Press (Thailand) Co., Ltd. Retrieved from http://www.nci.go.th/th/File_download/Nci Cancer Registry/Cancer in Thailand8.pdf

Naz, S., Majeed, H., & Irshad, H. (2010). Image segmentation using fuzzy clustering: A survey. In *The 6th International Conference on Emerging Technologies (ICET)* (pp. 181–186). IEEE. http://doi.org/10.1109/ICET.2010.5638492

Neoh, S. C., Srisukkham, W., Zhang, L., Todryk, S., Greystoke, B., Peng Lim, C., … Aslam, N. (2015). An Intelligent Decision Support System for Leukaemia Diagnosis using Microscopic Blood Images. *Scientific Reports*, *5*, 14938. http://doi.org/10.1038/srep14938

Neoh, S. C., Zhang, L., Mistry, K., Hossain, M. A., Lim, C. P., Aslam, N., & Kinghorn, P. (2015). Intelligent facial emotion recognition using a layered encoding cascade optimization model. *Applied Soft Computing*, *34*, 72–93. http://doi.org/10.1016/j.asoc.2015.05.006

Nilima, S., Dhanesh, P., & Anjali, J. (2013). Review on Image Segmentation , Clustering and Boundary Encoding. *Ijirset*, *2*(11), 6309–6314.

Nixon, M. S., & Aguado, A. S. (2008). *Feature Extraction and Image Processing* (2nd Ed.). Oxford,: Elsevier Ltd.

Okada, D. R., & Blankstein, R. (2009). Digital Image Processing for Medical Applications. *Perspectives in Biology and Medicine*, *52*(4), 617–623. http://doi.org/10.1353/pbm.0.0123

Ongun, G., & Halici, U. (2001). An automated differential blood count system. *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, (2), 2583–2586. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1017309

Ongun, G., Halici, U., Leblebicioglu, K., Atalay, V., Beksac, M., & Beksac, S. (2001). Feature extraction and Classification of Blood Cells for an Automated Differential Blood Count System. In *IEEE International Joint Conference on Neural Networks* (Vol. 4, pp. 2461–2466). http://doi.org/10.1109/IJCNN.2001.938753

Osowski, S., Siroic, R., Markiewicz, T., & Siwek, K. (2009). Application of support vector machine and genetic algorithm for improved blood cell recognition. *Instrumentation and Measurement, IEEE Transactions on*, *58*(7), 2159–2168. http://doi.org/10.1109/TIM.2008.2006726

Pan, C., Zheng, C., & Wang, H. (2003). Robust color image segmentation based on mean shift and marker-controlled watershed algorithm. In *the 2nd International Conference on Machine Learning and Cybernetics* (pp. 2752–2756). Xian, China: IEEE. http://doi.org/10.1109/ICMLC.2003.1260013

Park, S., Parwani, A. V., Aller, R. D., Banach, L., Becich, M. J., Borkenfeld, S., … Pantanowitz, L. (2013). The history of pathology informatics: A global perspective. *Journal of Pathology Informatics*, *4*(1), 1–36. http://doi.org/10.4103/2153-3539.112689

Patil, D. D., & Deore, S. G. (2013). Medical Image Segmentation : A Review. *International Journal of Computer Science and Mobile Computing*, *2*(January), 22–27.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. http://doi.org/10.1109/TPAMI.2005.159

Piuri, V., & Scotti, F. (2004). Morphological classification of blood leucocytes by microscope images. In *IEEE International Conference onComputational Intelligence for Measurement Systems and Applications* (pp. 103–108). http://doi.org/10.1109/CIMSA.2004.1397242

Polikar, R. (2006). Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, *6*(3), 21–45. http://doi.org/10.1109/MCAS.2006.1688199

Poynton, C. A. (1996). *A Technical Introduction to Digital Video*. New York: John Wiley & Sons.

Priyadarsini, R. P., Valarmanthi, M. L., & Sivakumari, S. (2011). Gain Ratio Based Feature Selection Method for Privacy Preservation. *ICTACT Journal on Soft Computing*, *1*(4), 201–205.

Putzu, L., Caocci, G., & Di Ruberto, C. (2014). Leucocyte classification for leukaemia detection using image processing techniques. *Artificial Intelligence in Medicine*, *62*(3), 179–91. http://doi.org/10.1016/j.artmed.2014.09.002

Rawat, J., Singh, A., Bhadauria, H. S., & Virmani, J. (2015). Computer Aided Diagnostic System for Detection of Leukemia Using Microscopic Images. In *Procedia Computer Science* (Vol. 70, pp. 748–756). Elsevier Masson SAS. http://doi.org/10.1016/j.procs.2015.10.113

Rezatofighi, 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. http://doi.org/10.1016/j.compmedimag.2011.01.003

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*, 1–39. http://doi.org/10.1007/s10462-009-9124-7

Sadeghian, F., Seman, Z., Ramli, A. R., Abdul Kahar, B. H., & 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. http://doi.org/10.1007/s12575-009-9011-2

Scott, A. S., & Fong, E. (2014). *Body Structures and Functions* (12th Ed.). CENGAGE Learning, Stanford, USA.

Scotti, F. (2005). Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications* (pp. 96–101). Giardini Naxos, Italy. http://doi.org/10.1109/CIMSA.2005.1522835

Scotti, F. (2006). Robust segmentation and measurements techniques of white cells in blood microscope images. In *IEEE Instrumentation and Measurement Technology Conference* (pp. 43–48). Sorrento, Italy. http://doi.org/10.1109/IMTC.2006.235499

Shah, A., Stiller, C. a, Kenward, M. G., Vincent, T., Eden, T. O. B., & Coleman, M. P. (2008). Childhood leukaemia: long-term excess mortality and the proportion "cured". *British Journal of Cancer*, *99*(1), 219–23. http://doi.org/10.1038/sj.bjc.6604466

Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, *66*(1), 7–30. http://doi.org/10.3322/caac.21332.

Singhal, V., & Singh, P. (2014). Local Binary Pattern for automatic detection of Acute Lymphoblastic Leukemia. In *IEEE The Twentieth National Conference on Communications* (pp. 1–5). http://doi.org/10.1109/NCC.2014.6811261

Sinha, N., & Ramakrishnan, A. G. (2003). Automation of differential blood count. In *The IEEE Region 10 Technical Conference on Convergent Technologies for the Asia-Pacific Region (TENCON 2003)* (Vol. 2, pp. 547–551). http://doi.org/10.1109/TENCON.2003.1273221

Smith, Z. J., Chu, K., Espenson, A. R., Rahimzadeh, M., Gryshuk, A., Molinaro, M., … Wachsmann-Hogiu, S. (2011). Cell-phone-based platform for biomedical device development and education applications. *PloS One*, *6*(3), 1–11. http://doi.org/10.1371/journal.pone.0017150

Soille, P. (2004). *Morphological Image Analysis: Principle and Applications* (2 Ed.). Springer-Verlag Berlin Heidelberg. http://doi.org/10.1007/978-3-662-05088-0

Srisukkham, W., Lepcha, P., Hossain, A., Zhang, L., Jiang, R., & Lim, H. N. (2013). A mobile enable intelligent scheme to identify blood cancer for remote areas cell membrane segmentation using marker controlled watershed segmentation phase. In *the 7th International Conference on Software, Knowledge, Intelligent Management and Applications (SKIMA 2013)* (pp. 104–114). Chiang Mai, Thailand.

Stewart, B. W., & Wild, C. P. (2014). *World Cancer Report 2014*. (B. W. Stewart & C. P. Wild, Eds.). World Health Organization: The International Agency for Research on Cancer (IARC).

StLukes Cancer Center. (2016). Benign Blood Disorders. Retrieved May 8, 2016, from https://cancer.slhn.org/Cancers-We-Treat/Blood-Cancers/Benign-Blood-Disorders

Sun, H. Q., & Luo, Y. J. (2009). Adaptive watershed segmentation of binary particle image. *Journal of Microscopy*, *233*(2), 326–330. http://doi.org/10.1111/j.1365-2818.2009.03125.x

Tang, J., Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review. In Charu C . Aggarwal (Ed.), *Data Classification: Algorithms and Applications* (pp. 37–64). Chapman and Hall/CRC 2014. http://doi.org/10.1.1.409.5195

Thairath. (2014, October 12). Statistic Estimated of Cancers killed 1 Thai person in every 8 minutes. Bangkok, Thailand. Retrieved from http://www.thairath.co.th/content/456247

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. http://doi.org/10.1109/TITB.2007.892694

Theodoridis, S., & Koutroumbas, K. (2006). *Pattern Recognition* (3rd Ed.). San Diego, CA.: Academic Press.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, *65*(2), 87-108. http://doi.org/10.3322/caac.21262

Turgeon, M. L. (2012). *Clinical hematology: theory and procedures* (5th ed.). Lippincott Williams and Wikins.

Uthman, E. O. (2016). Blood Cells and the CBC. Retrieved May 2, 2016, from http://web2.airmail.net/uthman/blood_cells.html

Vaccarella, S., Lortet-Tieulent, J., Plummer, M., Franceschi, S., & Bray, F. (2013). Worldwide trends in cervical cancer incidence: impact of screening against changes in disease risk factors. *European journal of cancer*, *49*(15), 3262-3273. https://doi.org/10.1016/j.ejca.2013.04.024

Valian, E., Tavakoli, S., Mohanna, S., & Haghi, A. (2013). Improved cuckoo search for reliability optimization problems. *Computers and Industrial Engineering*, *64*(1), 459–468. http://doi.org/10.1016/j.cie.2012.07.011

Vincent, L. (1993). Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. *IEEE Transactions on Image Processing*, *2*(2), 176–201. http://doi.org/10.1109/83.217222

Vincent, L., & Soille, P. (1991). Watershed in digital spaces: an efficient algorithm based on immersion simulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, *13*(6), 583–598.

Wang, Z. (2010). Comparison of Four Kinds of Fuzzy C-Means Clustering Methods. In *Third International Symposium on Information Processing* (pp. 563–566). IEEE. http://doi.org/10.1109/ISIP.2010.133

Wong, K. (2016). Evolutionary Algorithms: Concepts, Designs, and Applications in Bioinformatics: Evolutionary Algorithms for Bioinformatics. In *Handbook of Research on Advanced Hybrid Intelligent Techniques and Applications* (p. 190). IGI Global. http://doi.org/10.4018/978-1-4666-9474-3.ch007

Wu, K.-L., Yu, J., & Yang, M.-S. (2005). A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests. *Pattern Recognition Letters*, *26*, 639–652. http://doi.org/10.1016/j.patrec.2004.09.016

Yang, X.-S., & Deb, S. (2009). Cuckoo Search via Levy Flights. In *World Congress on Nature & Biologically Inspired Computing* (pp. 210–214). India. http://doi.org/10.1109/NABIC.2009.5393690

Yang, X.-S., & Deb, S. (2010). Engineering Optimisation by Cuckoo Search. *International Journal of Mathematical Modelling and Numerical Optimization*, *1*(4), 330–343. Retrieved from http://arxiv.org/abs/1005.2908

Yang, X.-S., & Deb, S. (2013). Multiobjective cuckoo search for design optimization. *Computers and Operations Research*, *40*(6), 1616–1624. http://doi.org/10.1016/j.cor.2011.09.026

Yeoh, A. E. J., Tan, D., Li, C. K., Hori, H., Tse, E., & Pui, C. H. (2013). Management of adult and paediatric acute lymphoblastic leukaemia in Asia: Resource-stratified guidelines from the Asian Oncology Summit 2013. *The Lancet Oncology*, *14*(12), e508–e523. http://doi.org/10.1016/S1470-2045(13)70452-2

Yin, Z., Tang, Y., Sun, F., & Sun, Z. (2006). Fuzzy clustering with novel separable criterion. *Tsinghua Science and Technology*, *11*(1), 50–53. http://doi.org/10.1016/S1007-0214(06)70154-7

Zhang, Y., Gong, D.-W., & Ding, Z. (2012a). A bare-bones multi-objective particle swarm optimization algorithm for environmental/economic dispatch. *Information Sciences*, *192*, 213–227. http://doi.org/10.1016/j.ins.2011.06.004

Zhang, Y., Gong, D., Hu, Y., & Zhang, W. (2015b). Feature selection algorithm based on bare bones particle swarm optimization. *Neurocomputing*, *148*, 150–157. http://doi.org/10.1016/j.neucom.2012.09.049

Zhang, Y., Zhang, L., & Hossain, M. A. (2015c). Adaptive 3D facial action intensity estimation and emotion recognition. *Expert Systems with Applications*, *42*(3), 1446–1464. http://doi.org/10.1016/j.eswa.2014.08.042

Zhang, Y., Zhang, L., Neoh, S. C., Mistry, K., & Hossain, M. A. (2015d). Intelligent affect regression for bodily expressions using hybrid particle swarm optimization and adaptive ensembles. *Expert Systems with Applications*, *42*(22), 8678–8697. http://doi.org/10.1016/j.eswa.2015.07.022

Zini, G., Bain, B., Bettelheim, P., Cortez, J., D'Onofrio, G., Faber, E., … Bene, M. C. (2010). A European consensus report on blood cell identification: Terminology utilized and morphological diagnosis concordance among 28 experts from 17 countries within the European LeukemiaNet network WP10, on behalf of the ELN Morphology Faculty. *British Journal of Haematology*, *151*(4), 359–364. http://doi.org/10.1111/j.1365-2141.2010.08366.x

Zitzler, E., Laumanns, M., & Thiele, L. (2001). SPEA2: Improving the Strength Pareto Evolutionary Algorithm. In *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems* (pp. 95–100). http://doi.org/10.1.1.28.7571

Zuiderveld, K. (1994). *Contrast limited adaptive histogram equalization*. (P. S. Heckbert, Ed.). Pittsburgh: Academic Press Professional, Inc.