APPLICATION OF DOUBLE-STRATEGY RANDOM FOREST FOR DIAGNOSIS AND SUBTYPING OF CANCER CELLS IN DIGITAL CYTOLOGY IMAGES OF PLEURAL EFFUSION

KHIN YADANAR WIN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF ENGINEERING IN ELECTRICAL ENGINEERING
FACULTY OF ENGINEERING
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2019

KMITL-2019-EN-D-018-012

APPLICATION OF DOUBLE-STRATEGY RANDOM FOREST FOR DIAGNOSIS AND SUBTYPING OF CANCER CELLS IN DIGITAL CYTOLOGY IMAGES OF PLEURAL EFFUSION

KHIN YADANAR WIN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF ENGINEERING IN ELECTRICAL ENGINEERING
FACULTY OF ENGINEERING
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2019

KMITL-2019-EN-D-018-012

THESIS CERTIFICATION
FACULTY OF ENGINEERING
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

Thesis Title	Application of Double-Strategy Random Forest for Diagnosis and Subtyping of Cancer Cells in Digital Cytology Images of Pleural Effusion
Student	Miss Khin Yadanar Win
Student Id.	57601414
Degree	Doctor of Engineering
Program	Electrical Engineering
Thesis Advisor	Asst. Prof. Dr. Noppadol Maneerat
Thesis Co-Advisor	Prof. Dr. Kazuhiko Hamamoto
Thesis Reference Number	KMITL-2019-EN-D-018-012

EXAMINERS	SIGNATURES
Assoc. Prof. Dr. Manas Sangworasil	
Assoc. Prof. Dr. Surapan Airphaiboon	
Assoc. Prof. Dr. Chuchart Pintavirooj	
Asst. Prof. Dr. Yuttana Kitjaidure	
Asst. Prof. Dr. Noppadol Maneerat	

Date 18th February 2019 **Time** 12:00-02:00 pm.
Place Building A , 5th Floor Conference room no.3

(Assoc. Prof. Dr. Komsan Maleesee)
Dean, Faculty of Engineering
18th February 2019

THIS THESIS IS DEDICATED TO MY MOTHER AND THE MEMORY OF MY FAHTER. THE POEM ADDED BELOW ("WHEN GREAT TREES FALL" BY MAYA ANGELOU) IS IMPLICITLY DEDICATED TO MY FATHER.

From "When Great Trees Fall"

By Maya Angelou

"And when great souls die,
after a period peace blooms,
slowly and always
irregularly. Space fill
with a kind of
soothing electric vibration.
Our senses, restored never
to be the same, whisper to us.
They existed. They existed.
We can be. Be and be
Better. For they existed"

Thesis Title	Application of Double-Strategy Random Forest for Diagnosis and Subtyping of Cancer Cells in Digital Cytology Images of Pleural Effusion
Student Name	Ms. Khin Yadanar Win
Student ID.	57601414
Degree	Doctor of Engineering
Program	Electrical Engineering
Thesis Advisor	Asst. Prof. Dr. Noppadol Maneerat
Thesis Co-Advisor	Prof. Dr. Kazuhiko HAMAMOTO

ABSTRACT

Cytology examination of fluid specimens by a cytopathologist is deemed as a gold standard for diagnosing many cancers, including cancer cells in malignant pleural effusion. In routine cytology examination, cytopathologists visually examine every single cell in the fluid samples under the light microscope to determine the prevalence of the malignancy. Nevertheless, it is the manual procedure which is subjective, laborious and time-intensive in addition to inter-and intra-observers bias. With the advent of digital cytology and advanced computer vision systems, computer-assisted algorithms can alleviate the aforementioned issues. The main objective of this thesis is to develop the algorithms that can aid in building the fully computer-aided diagnosis (CAD) systems that detect cancer cells and classify subtypes of cancer from pleural effusion samples. Three novel frameworks for (i) segmentation of cell nuclei, (ii) detection of overlapping nuclei and (iii) observing the most significant features are proposed for the automated analysis of cytology pleural effusion images.

In order to obtain the precise delineation of nuclei which is one of the most challenging problems in cellular image analysis applications, a novel nuclei segmentation algorithm is proposed. The algorithm hybridizes SLIC superpixels

algorithm and K Means clustering. It minimizes the computational time while preserving the natural shapes of the cell nuclei. Since there is a significant presence of overlapping nuclei, it is crucial to detect them and isolate into individual ones. A novel framework for identification of overlapping nuclei is proposed, whereby double-strategy random-forest is employed to learn the visual appearance of overlapping nuclei using a new feature set that contains geometric and textural features in a composite manner. Once the overlapped nuclei are detected, they are decomposed into its constituent components using contour concavity analysis. Then, the classification framework of cancer cells is presented. From each delineated nucleus, we extracted a total of 201 features and select the most discriminant features using a novel hybrid simulated annealing based feature selection method. Using the selected features, ensemble bagging classifier is employed to classify between normal and cancer cells. The synergy between hybrid simulated annealing and ensemble bagging classifier has achieved 98.70% accuracy. At last, a multi-categorical classification of lung cancer subtypes in pleural effusion is further developed using deep convolutional neural networks. It yielded a very convincing result by given 97.93% accuracy.

All these frameworks can play a fundamental role in developing a CAD system of cancer cells in pleural effusion. CAD systems can provide useful information for the detection and diagnosis of cancer cells in pleural effusion images and have the potential to complement the interpretation of the cytopathologists. Moreover, the proposed algorithms can be adapted to the diagnosis of cancer cells from several other fluid specimens such as peritoneal, synovial and cerebrospinal fluid.

IN THE MEMORY OF ASSOCIATE PROFESSOR DR. SOMSAK CHOOMCHUAY

The late Professor Somsak Choomchuay (1960－2018)

This page is dedicated to the memory of the late professor Somsak Choomchuay who was my advisor. Besides being a great advisor, he was a good friend and a father-like-figure to me. He was providing me with endless encouragement and guiding me through multiple challenges to be expected in the conduct of any research doctorate. He spent countless hours proofreading my research papers, giving me excellent suggestions which always resulted in improved versions of the papers, discussing my research, and providing me with excellent ideas to improve my research. I feel that I have learned a lot from him, and this is not limited to academic matters. My appreciation for his guidance, encouragement, and tremendous support is immeasurable. He will be deeply missed by me.

ACKNOWLEDGEMENT

I am indebted to many people who generously offered advice, encouragement, inspiration, and friendship throughout my doctoral study at King Mongkut's Institute of Technology Ladkrabang (KMITL) and Tokai University.

First and foremost, I would like to express my deepest appreciation to Assoc. Prof. Dr. Somsak Choomchuay (late advisor), Asst. Prof. Dr. Noppodal Maneerat and Prof. Dr. Kazuhiko HAMAMOTO for being the most enthusiastic advisors in all of my studies, for their valuable ideas and criticisms, and for their generous support for many conferences and articles. Without their strong support and encouragement, this thesis could not have been accomplished. I feel that I have learned a lot from them, and this is not limited to academic matters.

I owe my great gratitude to ASEAN University Network/Southeast Asia Engineering Education Development Network (AUN/SEED-Net) and Japan International Cooperation Agency (JICA) for funding together these research years under the Ph.D. studentship at KMITL and Tokai University. My sincere thanks go to the staffs of AUN/SEED-Net and JICA who have always given helpful and kind support.

I greatly appreciate the experts in Department of Pathology, Faculty of Medicine, Srinakharinwirot University, Thailand, for the generous support for the datasets and insightful suggestions. They contributed a lot of time to helping me get to and work with the cytology images; I hope that what I've developed will help them also. I would like to offer my special thanks for the time and energy from Dr. Manasanan Raveesunthornkiat. Thank her for agreeing to be my clinical collaborator, co-authoring a good number of publications, and providing valuable and insightful suggestions which have been the great assistance on my study.

I must express my deep gratitude to many professors from KMITL, especially from Department of Electronics Engineering, for their feedbacks and advices throughout my doctoral study. Special acknowledgements go to my friend, Syna Sreng for helping me out and for all the enlightening discussions. I want to show special thanks to all my friends at KMITL and Tokai University for their encouragement, inspiration, and friendship during my doctoral study. I am thankful to Arunee Pufa who have taken care of me during my stay in KMITL. Also, I would like to thank all my friends, near or far, who always be there to support and motivate me.

My parents deserve the most credits for their unconditional love, support, care and great encouragement over the span of my entire life. I also would like to show my special thanks to my siblings for without failing their continuous support. A lot of people helped and inspired me to accomplish this thesis. Hence, I would like to thank each and every one who support me during my doctoral study.

Khin Yadanar Win

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
CAD	Computer Aided Diagnosis
CCA	Contour Concavity Analysis
CPE	Cytology Pleural Effusion
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
DT	Decision Trees
EBC	Ensemble Bagging Classifier
FOS	First Order Statistics
GA	Genetic Algorithm
GLCM	Gray Level Co-occurrence Matrix
GLRLM	Gray Level Run Length Matrix
KNN	K Nearest Neighborhood
LDA	Linear Discriminant Analysis
LogR	Logistic Regression
MPE	Malignant Pleural Effusion
NB	Naïve Bayes
PSO	Particle Swarm Optimization
RGB	Red Green Blue
SA	Simulated Annealing
SLIC	Simple Linear Iterative Clustering
SVM	Support Vector Machine
WSI	Whole Slide Imaging

CHAPTER 1

INTRODUCTION

1.1 Research Background

Globally, cancer is one of the leading causes of death with high morbidity and mortality rates. It refers to a class of diseases caused by uncontrolled growth of cells. Normal cells perform the function of growth and death. In contrast, cancer cells are no longer respond to many of the signals that control cellular growth and death. Abnormal cells grow infinitely in uncontrolled manner. Some cancers may eventually invade into other parts of the body [1]. Pleural effusion or pulmonary effusion is the pathologic accumulation of fluid in the pleural cavity, between the visceral and parietal layers surrounding the lung, as demonstrated in Figure 1.1. Normally, the pleural space is lined by a thin layer of mesothelial cells and contains about 5-10 ml of clear fluid to lubricate during respiratory movement [2,3]. When the abnormal accumulation of fluid buildup in the pleural cavity, it forms Pleural effusion (PE). PE can be developed by a variety of diseases. In addition to various cancers, it can also be seen in infections and other diseases (Table 1.1). If the effusion is caused by cancer cells, it is called Malignant Pleural Effusion (MPE). MPE is an aggressive cancerous effusion and a sign of advanced stage of cancer. Half of the cancer patients end up developing MPE. MPE

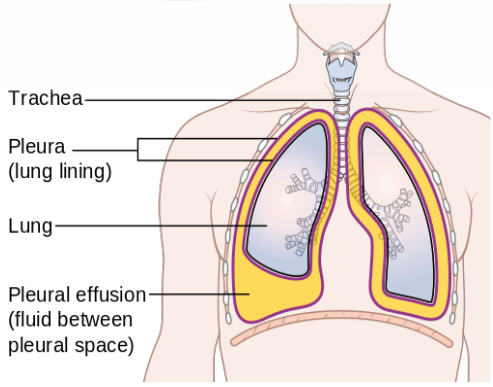


Figure 1.1 The anatomy of pleural effusion (Image credit [3])

accounts for 10% of all PE cases. As seen in Table 1.1, primary cancer, as well as metastases cancers, can cause MPE. Mesothelioma (a primary cancer of the pleura itself) is a rare form of MPE. Metastases cancers from lung, breast, and lymphoma are the main causes of MPE. Cancers from stomach, kidney, ovaries, and colon also may invade to pleural fluid and cause MPE.

Table 1.1 Variety of diseases that cause PE [4-6]

Malignant tumors	Tumors from lung, breast, ovary, kidney and colon; leukemia; lymphoma; Hodgkin's diseases; tumors of the genitourinary tract and gastrointestinal; malignant pleural mesothelioma
Infectious diseases	All pneumonic pleural inflammations, acute empyema, chronic empyema, tuberculosis, parasitic infection
Collagen diseases	Rheumatoid arthritis, systemic lupus erythematosus, Churg–Strauss syndrome
Gastrointestinal disease	Liver cirrhosis, acute pancreatitis, liver abscess, sub phrenic abscess, peritonitis, esophageal perforation
Cardiovascular disease	Congestive heart failure, lung infarction, ruptured thoracic aortic aneurysm, Dressler syndrome
Renal disease	Nephrotic syndrome
Gynecological disease	Meigs syndrome, pleural endometriosis
Iatrogenic disease	Drugs, post-thoracic/abnormal surgery complication, radiation
Other	External injury, spontaneous pneumothorax, benign asbestos pleurisy, sarcoidosis, yellow nail syndrome, pulmonary lymphangioleiomyomatosis

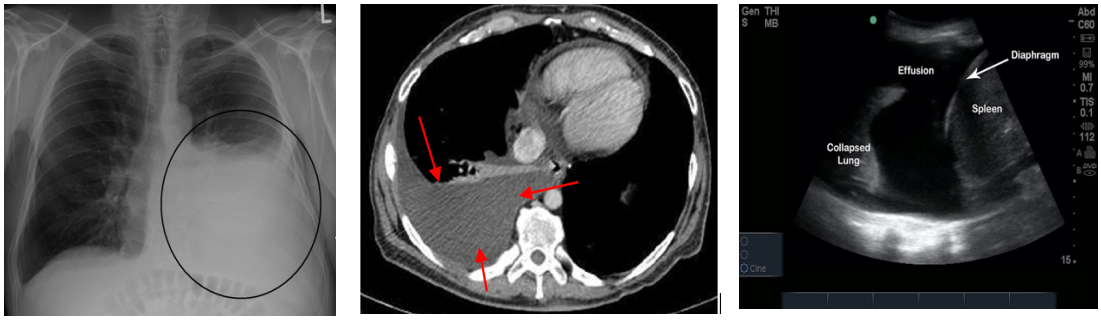
MPE always implies an advanced stage of cancer and reduced survival. The survival rate varies depending on the origins of cancer cells and tumor characteristics. Median survival ranges from 3 to 12 months. MPE caused by lung cancer has the shortest survival, and stomach cancer the longest survival. Depending on the disease and amount of the effusion, the symptoms of PE can vary from being asymptomatic in some people to being very troublesome in others. The common symptoms of MPE are

dyspnea (i.e., shortness of breath), chest pain, cough, pain when taking a deep breath, fever, fatigue and so on [4-6].

1.2 Imaging Techniques for Detection of Pleural Effusion

A physician may suspect the presence of PE based on a person's symptoms and physical examinations such as auscultation (listening to your lungs with a stethoscope), percussion (tapping on the chest), and other maneuvers. If symptoms and signs are suspected to have PE, it requires one or more imaging tests to confirm the presence of PE and the amount of fluid. Common tests used to identify pleural effusions include:

- **Chest X-ray film:** Plain X-ray films of the chest are often the first step in identifying PE. PE appear on chest X-rays as white space at the base of the lung. If PE is likely, additional X-ray films, called decubitus X-ray films, may be taken while a person lies on her side. These radiographs can show if the fluid flows freely within the chest.
- **Computed tomography (CT):** A CT scanner takes multiple X-rays rapidly, and a computer constructs images of the inside of the chest. Compared to chest X-rays, CT scans give more detailed information about PE and other lung abnormalities than chest x-ray. They can detect even small amounts of PE, thereby, playing a significant role in the investigation of intrapulmonary and extrapulmonary lesions [7,8].
- **Ultrasound:** Ultrasound imaging uses sound waves to produce pictures of the inside of the body. Ultrasound does not involve any radiation and can help guide drainage and identify whether pleural effusions are free-flowing [7-8]. Figure 1.2 illustrates the presence of PE on the different imaging tests.



(a) (b) (c)

Figure 1.2 The imaging techniques showing the presence of PE: (a) Chest X-ray, (b) CT scan, and (c) Ultrasound (Image credit [9-11])

1.3 Cytology Examination for Diagnosis of Cancer Cells in Pleural Effusion

Once PE is confirmed on the imaging tests, PE sample is usually taken to determine the causes and relieve the symptoms. In order to determine the malignancy (cancerous) of PE, cytology examination, the study of cellular content in the body fluids, is deemed as the gold standard because it is a simple, cheap, less invasive and highly useful tool. During cytology examination, the samples of PE are first collected in a clinical setting, by a physician, physician assistant, nurse, or dedicated technician. These samples are transported to the cytopathology lab wherein various sample preparation steps are performed prior to analysis including centrifugation to concentrate cells in dilute samples, followed by preparation and staining of slides. Finally, the cytologists or pathologists visually examine the morphology changes and visual abnormalities of every single cell in the glass slides under the microscope to determine the prevalence of malignancy. The procedures for cytology examination are depicted in Figure 1.3 [12-14]. The details of sample collection, sample preparation, and analysis are presented in the following subsections.

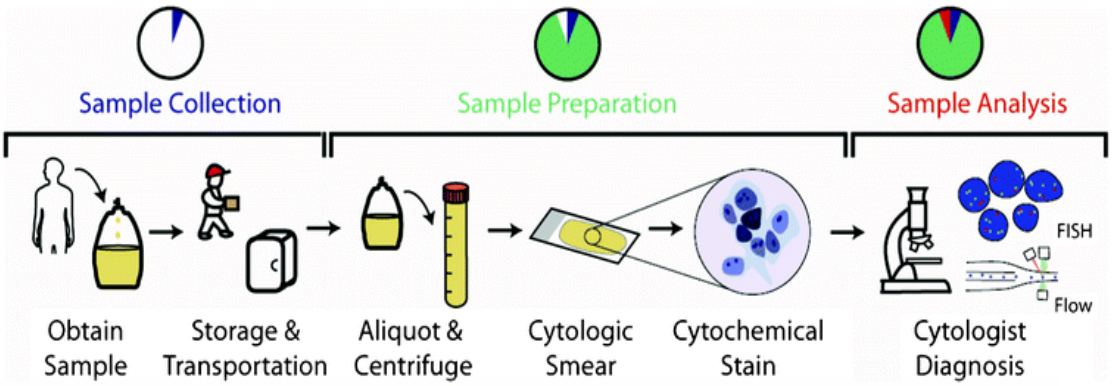
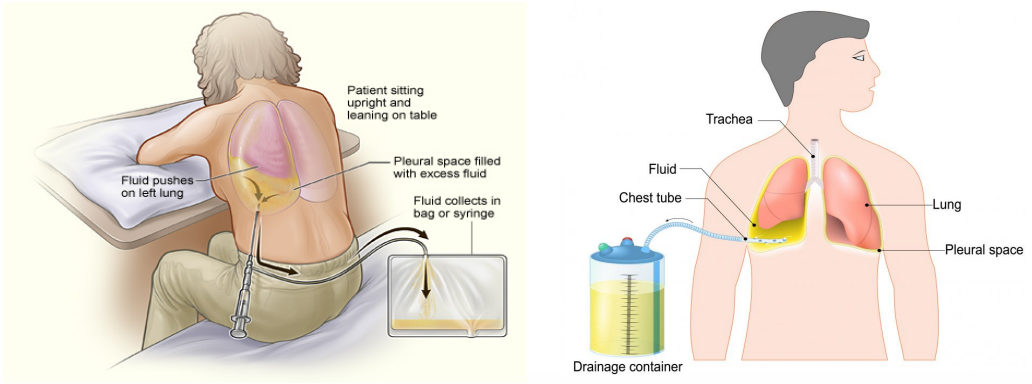


Figure 1.3 The preparation of cytological glass slides for cytology examination (image credit: [14])

1.3.1 PE sample collection

A small amount of fluid is withdrawn for testing, and a large amount can be removed simultaneously to ease the symptoms. To access PE, a few approaches may be taken. In a procedure called thoracentesis, also known as pleural fluid aspiration, a doctor inserts a needle and a catheter between the ribs, into the pleural space, and the fluid is drained into a bag. Locating the fluid before thoracentesis reduces the risk of puncturing the lung, liver, or spleen. In the procedure called chest tube placement, a chest tube is a flexible tube that is placed from outside the body and into the pleural space. The tube may be left in place for varying amounts of time depending on the reason it is placed [6-8,15-16]. Figure 1.4 shows the accessment techniques of PE. Fig 1.4 (a) shows a person having a thoracentesis, in which the person sits upright and leans on a table and excess fluid from the pleural cavity is drained into a bag. Figure 1.4 (b) shows the diagram of a chest tube draining pleural effusion.



(a) Aspiration of PE (thoracentesis) (b) Withdrawing PE with chest tube

Figure 1.4 The withdrawal of PE using thoracentesis and chest cube placement (images credit: [15, 16])

1.3.2 PE sample preparation

Most laboratories prepare cell smears or cell blocks by two or more methods of preparing cytology samples for further diagnosis and analysis. Cell smears or cell blocks are prepared by a cytopathologist using stains. An aliquot of the sample, typically 50 mL in PE specimens, is centrifuged and the sediment is used for preparing direct cell smears or cell blocks. In direct smear, the sediment cells are manually conveyed onto a glass slide and the cytologists review it under the microscope directly. As an alternative, the cytocentrifuge method takes the sediment cells and evenly scatters the cells onto a designated circle on a glass slide in an automatic and reproducible fashion. Cell blocks are formed from cell sediments that are then embedded in paraffin and cut into histological sections. Following cell preparation, the slides are stained with colored dyes to help differentiate cells by color-specific features of cellular morphology. The slides are alcohol-fixed or air-dried to remove water content; this is then followed by a series of washing and staining with stains like Papanicolaou (Pap) or Romanowsky to highlight nuclear and cytoplasmic features [14].

1.3.3 Conventional Cytology Examination

Once the samples are prepared, cytology examination is performed wherein cytopathologists manually examine cytology glass slides (cell smears or blocks) under light microscopy to determine the prevalence of cancer. During the exam, important parameters including cell size, morphology, nuclear to cytoplasmic ratio, and the presence of multi-nucleation of cell aggregates are taken into particular attention. In the case of cancerous cells, cell chromatin is generally more unfolded and will often appear darker. The laboratory diagnosis may be reported with a heading such as `positive for malignant cells', `suspicious', or `negative for malignant cells'. For 'suspicious' and 'positive' cases, cytologists may often request further analysis. PE specimens can also be investigated with flow cytometry, cytogenetic testing using fluorescence in-situ hybridization (FISH), or immunocytochemical analysis [14].

1.3.4 Limitations of Conventional Cytology Examination

Unfortunately, there are a few limitations in conventional cytology examination. Manually analyzing many cytology slides is subjective in nature, tedious, laborious and time intensive since it requires to examine hundreds of thousands of cells. It is also prone to inter- and intra-observer bias (subjective disparity from different cytologists depending on their expertise) that is further exacerbated by the lack of adequate cytopathology experts. These factors motivated us doing the research in the field of computer-aided diagnosis systems.

1.4 Computer-Aided Diagnosis Systems

Thanks to recent advances in digital pathology along with the advent of advanced computer vision algorithms, automated analysis of cytology images can alleviate the aforementioned limitations of manual cytology examination. Computer-

Aided Diagnosis (CAD) systems are becoming the useful and assistance tools to the cytologists in the diagnosis of cancer cells from many cytology specimens. It allows examining the cytology slides in a small amount of time with consistent classification results while catering the vulnerabilities of manual examination. CAD systems would be the great support for cytopathologists, especially when dealing with large number of slides for analysis. In addition,, they are able to provide accurate and productive diagnosis results, reduce the subjective interpretation and the burden of cytopathologists, reduce the cost and time span, and provide the insight and thorough for the research.

1.4.1 Digitalization of Cytological Glass Slides

While developing a CAD system of cancer cells in PE, it is required to digitalize the glass slides into virtual slides. Generally, there are two common digitalization systems: (i) digital cameras mounted on microscopes, and (ii) slides scanners. In the early years, the experts capture the images using the digital cameras mounted on microscopes. In the late 90s, Watzel and Gilbertson developed the world's first whole-slide imaging (WSI) scanner and thus marked the beginning of an era. The present-day WSI scanners enable high throughput (≥35 seconds per slide) slide digitization. This includes loading of the slides on the scanning platform, automated barcode reading, tissue identification, focus, scanning, image compression, generation and updating the digitization information on the laboratory information system [17-19]. Figure 1.5 shows two platforms for capturing the cytology images. Figure 1.5 (a) shows an example of capturing the cytology image of pleural effusion using a digital camera mounted on top of a traditional microscope while Figure 1.5 (b) showing an example of using a digital whole slide scanner.

(a) Olympics microscopic camera (b) Whole slide scanner

Figure 1.5 The digitalization systems of cytology images (a) using camera mounted on Olympics microscope, and (b) using Paranomic whole slide scanner (image credits: [20, 21])

1.4.2 Image Acquisition

In our study, the images are acquired through capturing by cytopathologists using Olympus microscope camera and cropping the interested area from whole slide images which are scanned using Paranomic Flash 250 whole slide scanner. The sample CPE images are given in Figure 1.6. All the images are captured with 40x magnification. They are stored in RGB model with JPEG format. The sizes of the original images were 4050x2050 and 1650x1600 pixels resolutions. Cytopathologists also labeled cancer cells on the cytology pleural effusion (CPE) images. The labeled images are considered as the ground truth to evaluate the diagnostic accuracy.

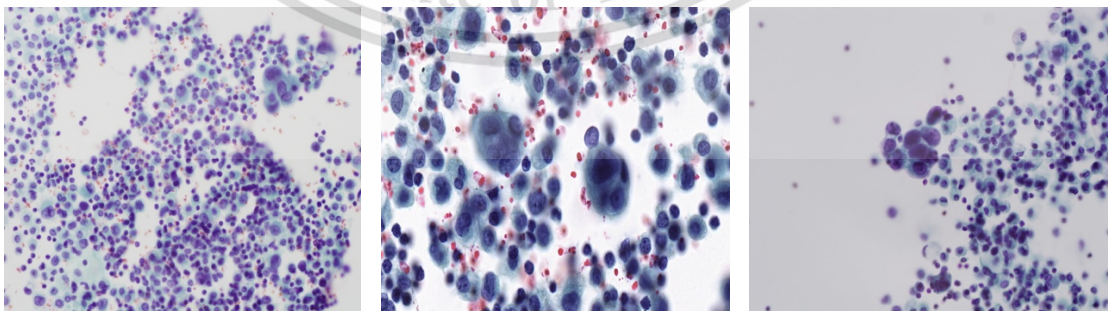


Figure 1.6 The sample CPE images

1.5 Problem Statement and Objectives

As mentioned in Section 1.4, CAD systems promise to bring objectivity and reproducibility using image analysis techniques and has the potential to provide micro and macro prognostic cues, which may be ignored during the visual examination by humans. The main objective of this thesis is to develop the algorithms that can aid towards building the fully CAD systems that detect cancer cells and classify types of cancers from pleural effusion samples. The general framework of a CAD system in this thesis is depicted in Figure 1.7.

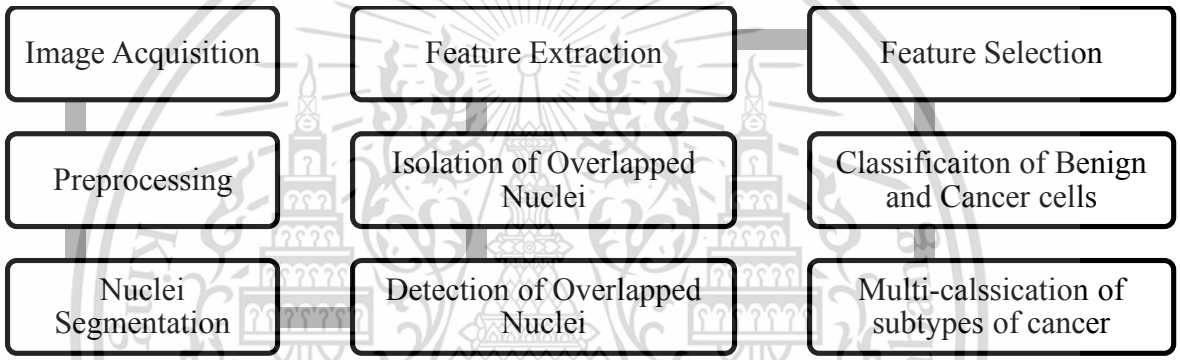


Figure 1.7 The typical architecture of CAD systems of cancer cells

The main focuses are on the accurate classification of cancer and benign cells, and subtypes of cancer using the capabilities of computer vision methods and machine learning methods. To achieve the goal, we aim to develop effective algorithms. One of the generic problems in CPE image analysis is the occurrence of image quality degradation due to the presence of noises and artifacts, uneven staining, and uneven illumination during image acquisition. In order to seek the most effective image preprocessing algorithms, we analyzed several filtering and image enhancement methods, and the most effective ones are employed to enhance the image quality prior to the main analysis. Segmentation of cell nuclei is one of the most challenging tasks in cytology image analysis. Some reasons for this difficulty are: (1) variation due to

staining which may leave some of the nuclei weakly stained. Thus, a CPE image might have a large number of nuclei with broken cell membrane making them difficult to discern from the background texture; (2) diversity in the shape of epithelial cancerous nuclei, whose appearance may vary from a round-shape normal nucleus to an enlarged and irregularly deformed nucleus with scrambled chromatin structures. The difficulty of the problem increases significantly when the cell density of fluid sample is high, resulting in cell overlap and clumping issues. We introduced a novel nuclei segmentation algorithm which hybridize simple linear iterative clustering (SLCI) super pixels and K means clustering method and compared with twelve existing image segmentation methods. Another crucial task in cytology image analysis is the presence of touched or overlapped nuclei. Since the cancer cell and benign cells are differentiated based on certain features, it is important to delineate each cell nucleus precisely. Extraction of the dominant features plays a vital role in biomedical image analysis. We extracted several features and introduce a novel feature selection algorithm to select the discriminant features. Machine learning models are employed as the classifiers to classify between cancerous and benign cells. As the final step, it is beneficial to classify the subtypes of cancer for further diagnosis and treatment. We presented deep learning algorithms to subtype three types of lung cancer from CPE images.

The major aims of the thesis can be summarized as follows:

1) The first task is to develop an accurate nuclei segmentation algorithm
2) The second task is to develop an algorithm for the isolation of the overlapped nuclei into individual ones.
3) When the cell nuclei are precisely delineated from the second task, we further analyze each segmented nucleus for the development of an automated classification system of cancer cells.

4) The fourth task is to automatically classify the subtypes of cancer.

1.6 Main Contributions

The major contributions of this thesis are listed below:

- Development of a novel algorithm for nuclei segmentation by hybridizing SLIC superpixels and K Means clustering methods
- Development of a new framework for detection and classification of overlapping nuclei using the new combination of features and a double-strategy random forest.
- Development of a hybrid feature selection algorithms by hybridizing simulated annealing algorithm and artificial neural network.
- Development of a CAD system of cancer cells on CPE images based on the combination of the above-mentioned developments and ensemble bagging classifier.
- Development of a multi-classification framework of lung cancer types in pleural effusion using deep convolutional neural networks.

1.7 Thesis Structure

This thesis is structured into six chapters. Chapter 1 presents a brief introduction of the area of cytology image analysis, and the main purpose of writing this thesis including problem statement, motivation and objective, and main contributions. The sample preparation and imaging processes are also covered and special attention is given to the computer-aided diagnosis methods. Chapter 2 presents the nuclei segmentation which plays an important role in improving the classification accuracy of cancer cells in Chapter 4. A detailed review of the literature on the topic of nuclei

segmentation in CPE images and comparative studies of twelve image segmentation methods are covered, and particular attention is given to the development of a novel hybrid nuclei segmentation method. In Chapter 3, a novel algorithm for the detection of overlapping nuclei and decomposing them into its constituents is presented. Chapter 4 presents the classification of cancer cells covering feature extraction, feature selection, and classification. Chapter 5 presents the multi-categorical classification of lung cancer types from CPE images using deep convolutional neural networks. Chapter 6 wraps up the results once more, presents the main contributions of the thesis and future directions of the research.

CHPATER 2

NUCLEI SEGMENTATION

2.1 Introduction

Segmentation of cell nuclei is a key task towards developing CAD systems of cytology images. The cell nuclei present significant changes when it is affected by a disease. The morphological changes and visual abnormalities of cell nuclei are mainly associated with the assessment of malignancy. The cell nuclei provide more diagnostic values than other components of the cells. The accurate segmentation of cell nuclei may lead to the good performance of subsequent analysis. Hence, it is crucial that nuclei are accurately extracted from the surrounding background such as cytoplasm, blood cells, and artefacts. However, it is a very challenging problem. Two main factors that make it a challenging problem are: (1) variation due to staining which may leave some of the nuclei weakly stained. Thus, a CPE image might have a large number of nuclei with broken cell membrane making them difficult to discern from the background texture; (2) diversity in the shape of epithelial cancerous nuclei, whose appearance may vary from a round-shape normal nucleus to an enlarged and irregularly deformed nucleus with scrambled chromatin structures. The difficulty of the problem increases significantly when the cell density of the fluid sample is high, resulting in cell overlap and clumping issues. The structure of the cell in CPE images is illustrated in Figure 2.1. Nuclei appear as the dark purple object in CPE images.

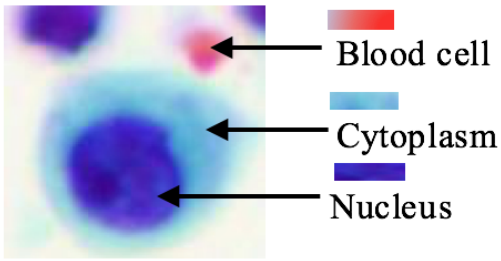


Figure 2.1 The structure of cell in CPE images

The rest of this chapter is organized as follows. Subsection 2.2 reports the related works for the segmentation of cell nuclei in CPE images. Subsection 2.3 presents preprocessing. Subsection 2.4 describes the analysis and comparison of twelve image segmentation methods for extracting the cell nuclei. Subsection 2.5 introduces the novel hybrid nuclei segmentation. Subsection 2.6 presents the post-processing stage. Subsection 2.7 discusses the experimental results. Subsection 2.8 summarizes nuclei segmentation.

2.2 Related Works

Few researchers have studied automated segmentation of cells or nuclei in CPE images. Zhang et al. [22] reported a method to detect malignant cells from CPE images using integrated fuzzy edge detection and Otsu's method. Fuhua et al. [23] presented a method on the basis of the wavelet and morphology transform to detect malignant cells from pleural effusion images. However, a preprocessing stage for removing noise and enhancing contrast is not considered in these methods, thus reducing the accuracy of the detection system. In addition, the methods are not focused on the segmentation process and there is a lack of clear quantitative evaluation of the proposed methods. E. Baykal et al., 2017 [24] have applied the active appearance model to extract the nuclei in CPE images and obtained 98.77% accuracy. The results obtained are compared with those from color thresholding, clustering, and graph-based methods. In [25], they

investigated the detection of cell nuclei using a supervised learning approach. The approach is based on the combination of Haar filter and AdaBoost classifier. Three images with a total of 178 nuclei were used for testing. The True Positive Rate of 89.32% and False Positive Rate of 5.05% were obtained. However, their approaches are designed to segment the image which is assumed to have only a cell. It made hard to reproduce those approaches in real practice since there may be up to million cells in one image. Their framework performed well with independent cell nucleus; however, it has the limitation to segment out the overlapped cell nuclei. Moreover, it required extensive prior knowledge to train the classifier.

Segmentation of cell nuclei, especially the presence of overlapping nuclei, presents many difficulties to traditional segmentation methods. Reliable nuclei segmentation of CPE images is still a challenging task because of the high cell population, their varieties and the presence of overlapping cells. Studies on the automated analysis of cytology CPE images are few because of the lack of reliable cell nuclei segmentation methods. There are still opportunities for further enhancements in the nuclei segmentation of CPE images. Thus, more observations are required to implement and determine the most feasible segmentation method. Most of the nuclei segmentation algorithms often embed pre-processing and post-processing to optimize the segmentation results. Figure 2.2 shows the generalized framework of the nuclei segmentation in this study.

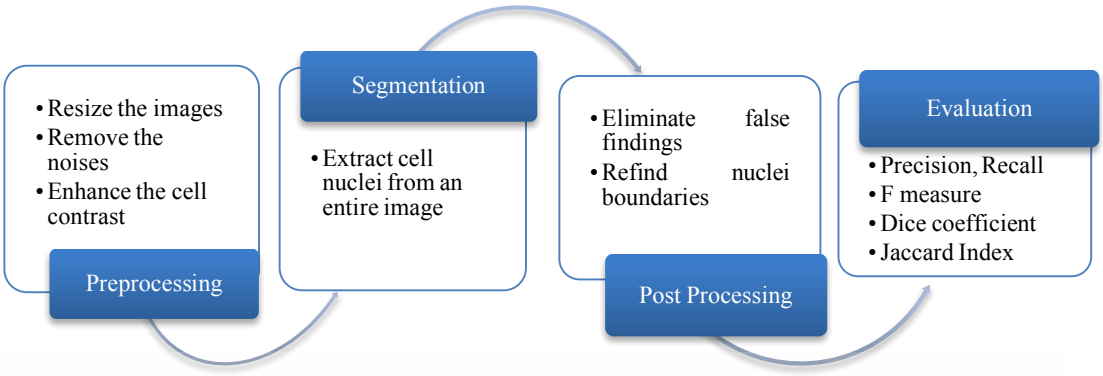


Figure 2.2 Generalized framework of cell nuclei segmentation in CPE images

2.3 Preprocessing

The preprocessing stage is a crucial task to enhance the quality of the image. Firstly, to standardize the sizes of the images and get the low computational load, we resized the original input image into resolutions of 1052 x 1052 pixels. The resized image is then converted into different color spaces based on the utilized segmentation methods. The CPE images might contain debris, dirt, or stained artifacts because of the uneven illumination or dirt on the camera surface resulting from the image acquisition process. Therefore, preprocessing is carried out to reduce the noises and artifacts, and improve cell contrast. Firstly, the filtering methods are applied to suppress the noises. Since there are many filtering methods, we employed five famous filtering methods namely Gaussian filter, Laplacian filter, Wiener filter, median filter, and mean filter and evaluate their performance by computing peak signal-to-noise ratio (PSNR). The PSNR is utilized as a performance measure to evaluate the quality of the filtered image. To compute the PSNR, the first step is to compute the mean square error (MSE) using Eq. (2.1). Then, PSNR is computed through Eq. (2.2). The higher the PSNR, the better is the image quality. Figure 2.3 (a) compares the PSNR results of five filtering methods. Median filter exhibits the highest PSNR. Therefore, we utilized the median filter to reduce the noises. The median filter is a non-linear method to suppress the noises by

windowing the noisy image [26]. Default window size 3x3 is used to remove the small noises.

$$MSE = \frac{\sum_{M,N}[O(m,n) - F(m,n)]^2}{M * N} \tag{2.1}$$

Where O and F represent the original image and filtered image respectively. M and N denotes the number of rows and columns in the input images.

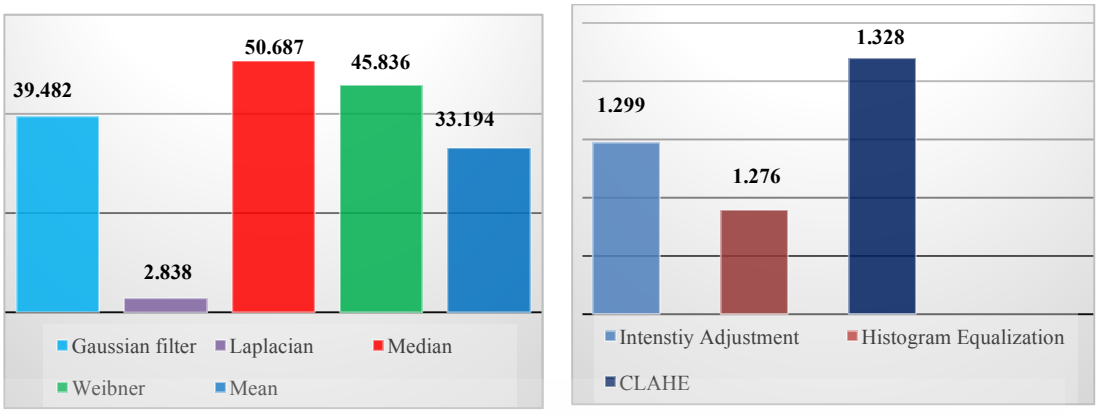
$$PSNR = 10\ log_{10}\left(\frac{R^2}{MSE}\right) \tag{2.2}$$

Where R denotes the maximum fluctuation in the input image.

In order to improve the contrast of the cell, we employed three enhancement methods namely histogram equalization, intensity adjustment and contrast limited adaptive histogram equalization (CLAHE) and evaluated their performance using contrast improvement index (CII) through Eq. (2.3). The CII is used as the performance measure to evaluate the image enhancement techniques in terms of the luminance, contrast, and structure. The higher the CII value, the better the contrast is. Figure 2.3 (b) compares the CII results of three enhancement methods. The CLAHE [27] yielded the highest CII. Hence, CLAHE with 8bit histogram bins is utilized to enhance the contrast of the cell. The visual results after applying CLAHE and median filter are shown in Figure 2.4.

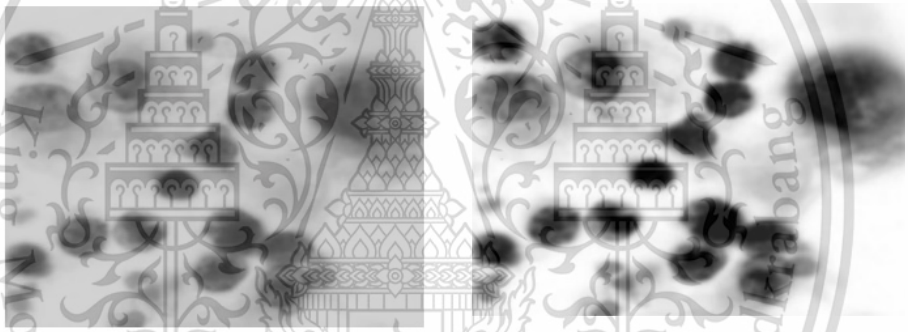
$$CII = \frac{C_E}{C_O} \tag{2.3}$$

Where C_E and C_O represents the average values of the local contrast with 3×3 window in the original image (O) and enhanced image (E).



(a) PSNR of filtering methods (b) CII of enhancement methods

Figure 2.3 Image quality assessment metrics, (a) comparison of filtering methods in terms of peak signal-to-noise ratio (PSNR), and (b) comparison of different contrast enhancement methods in terms of the contrast improvement index (CII).



(a) (b)

Figure 2.4 Preprocessing stage: (a) gray scale image, and (b) preprocessed image after median filter and CLAHE (note that the image was cropped for better visibility).

2.4 Comparative Analysis of Twelve Traditional Image Segmentation Methods for Nuclei Segmentation

In the cytology and histology image analysis, nuclei segmentation often revolves around thresholding techniques, clustering techniques, and active contour techniques. Using a small number of images, we have proposed several alternative nuclei segmentation methods using OTSU thresholding approach, K Mean clustering approach, and supervised pixel classification using ANN [28-30]. Thereafter, we have

collected more images and built the new dataset consisting of 35 CPE images. Using that new dataset, we analyzed twelve segmentation methods: (1) Otsu method, (2) Isodata thresholding method, (3) maximum entropy thresholding method, (4) cross entropy thresholding, (5) minimum error thresholding, (6) fuzzy entropy thresholding method, (7) adaptive thresholding method, (8) K Means clustering, (9) fuzzy c means clustering, (10) mean shift clustering, (11) Chan-Vese level set and (12) graph cut methods to extract the cell nuclei in CPE images. We categorized them into four groups: thresholding, clustering, active contour, and graph-based techniques.

2.4.1 Thresholding Techniques

The thresholding technique is the simplest segmentation method which is based on the gray level image histogram. It seeks the adequate threshold value to differentiate between the foreground (regions of interest) and the background. The threshold value can be global or local. The global thresholding uses a single optimal for the entire image. In contrast, the local thresholding is based on the threshold for each pixel which is computed depending on its local properties. The normalized gray level histogram of the image is used as the input for most of the thresholding techniques.

2.4.1 (a) Otsu Thresholding Method

Otsu's method, which is initially introduced by Nobuyuki Otsu, is one of the global thresholding techniques. Otsu's method aims to select the optimal threshold that minimizes the intra-class variance [31]. It is summarized in Algorithm 1.

2.4.1 (b) Isodata Thresholding

Isodata thresholding method is also one of the global image thresholding techniques [32]. It requires the initial threshold value as the input. The mean intensity

of the image histogram is initiated as the input. The iterative procedures are given in Algorithm 2.

Algorithm 1: Otsu Thresholding
1. Compute the histogram and the probabilities for all intensity levels.
2. Set up the class probability (w_i) and mean (μ_i).
3. Move to all possible maximum intensity of thresholds.
4. Update w_i and μ_i.
5. Select the maximum value among class variances.

Algorithm 2: Isodata Thresholding
1. Initialize the threshold (T) as the mean image intensity
2. Segment the images into two regions (R1 and R2) using T with this formation (R1 <T and R2 ≥T)
3. Find the mean intensity level ($u1$) for R1 and (u2) for R2.
4. Find new threshold value (T= (u1+u2)/2)
5. Loop steps 2 to 4 to find the difference in T.
6. Select T when it is smaller than the predefined parameter.

2.4.1 (c) Maximum Entropy Thresholding

The maximum entropy-thresholding method is another global thresholding method. Similar to Otsu's method, it selects an optimal threshold by maximizing the information measured between the foreground and the background [33]. The normalized histogram of the image is fed as the input parameter. The processing steps of maximum entropy thresholding are summarized in Algorithm 3.

Algorithm 3: Maximum Entropy Thresholding
1. Determine the normalized histogram of the images. 2. Determine the entropy of white and black pixels. 3. Obtain an optimal threshold by maximizing the entropy of white and black pixels.

2.4.1 (d) Cross Entropy Thresholding

The cross-entropy thresholding is one of the entropy thresholding methods. Numerous algorithms have been developed for the cross-entropy thresholding. Here, we focus on the one proposed by Li and Lee (1993), summarized as follows [34]. Similar to maximum entropy thresholding, cross entropy thresholding acquires the histogram of the image as the input parameter. The procedures of cross entropy thresholding are listed in Algorithm 4.

Algorithm 4: Cross entropy thresholding
1. Determine the histogram of the image. 2. Generate the resulting image by setting the threshold value T as the mean image intensity. 3. Compute cross entropy between the original image and the resulting image 4. Select the optimal value by minimizing the cross entropy.

2.4.1 (e) Fuzzy Entropy Thresholding

Fuzzy entropy is defined as the measure of uncertainty of a fuzzy set. It requires two input parameters: (i) histogram of the image to compute the probability distribution and (ii) fuzzy membership function [35]. The procedures of the fuzzy entropy thresholding are summarized in Algorithm 5.

Algorithm 5: Fuzzy Entropy Thresholding
1. Determine the two probability distributions of foreground and background. 2. Convert the image into fuzzy set using membership function. 3. Compute the membership functions for background and foreground using threshold T. 4. Rewrite the fuzzy form of entropic for step 1. 5. Select the optimal threshold by maximizing the fuzzy entropy.

2.4.1 (f) Minimum Error Thresholding

In this method, the image segmentation is based on the average pixel optimization [36]. It requires the normalized histogram of the gray level image as the input parameter. The idea behind this thresholding technique is summarized in Algorithm 6.

Algorithm 6: Minimum error thresholding
1. Determine the histogram that considers the mixture of two normal distributions having respective mean and variance, and respective proportions. 2. Initialize the trial threshold value T for modeling the two resulting pixel populations. 3. Model the two populations using the normal distribution. 4. Set different levels as the threshold value. 5. Compute the fitting criterion for each threshold value. 6. Select the threshold value which minimizes the fitting criterion as the optimal threshold.

2.4.1 (g) Adaptive Thresholding

Adaptive thresholding is the most famous local thresholding method for images with uneven illumination. It aims to select threshold values for each region based on its

local properties. The local window size is empirically set as 12. We summarized the adaptive thresholding procedures in Algorithm 7 [37].

Algorithm 7: Adaptive thresholding
1. Binarize the image with T. 2. Thin the binary image. 3. Remove all branch points. 4. Place the remaining endpoints in line to use as starting point for tracking 5. Track the object with T 6. Set the criteria T=T-1 if the object passed testing; otherwise, return to step 5.

2.4.2 Clustering Techniques

Clustering-based segmentation methods aim to group the collection of pixels into clusters. The pixels in the same cluster are closely related to one another.

2.4.2 (a) K Means Clustering

K-means clustering is one of the clustering methods wherein the data are divided into a specific number of groups by minimizing the within-class variance [38]. It is summarized in Algorithm 8.

Algorithm 8: K Means clustering
1. Select k cluster centers as 2. 2. For each pixel of an image, find its closest center and assign to the closest class 3. Update every center as the mean of its points 4. Repeat until it convergence or when there are no changes during the assignment step, or when the average distortion per point decreases slightly. 5. Reshape the cluster pixels into the image.

2.4.2 (b) Fuzzy c-Means Clustering

The fuzzy c-means clustering is one of the most popular fuzzy clustering methods, wherein the data are partitioned into two or more fuzzy clusters by maximizing the objective function [39]. We summarized the steps involved in this technique in Algorithm 9.

Algorithm 9: Fuzzy c-Means clustering
1. Choose random centroids, at least two.
2. Compute the membership matrix.
3. Calculate the cluster center.
4. Repeat steps 2 and 3 until the minimum objective function value is achieved.

2.4.2 (c) Mean Shift Clustering

Among the clustering-based segmentation methods, the mean shift segmentation is known as an advanced and highly useful technique. In the mean shift, a window is defined for each data point and the mean is subsequently computed. The center of the window is shifted to the mean and the iteration is performed until it converges [40]. Mean shift clustering's procedures are summarized in Algorithm 10.

Algorithm 10: Mean shift clustering
1. Compute features (color, gradients, texture, etc.)
2. Initialize windows at individual pixel locations
3. Perform mean shift for each window until convergence
4. Merge windows that end up near the same "peak" or mode.

2.4.3 Graph based Techniques

Graph-based models consider the image as a weighted graph. Every pixel in the image is considered as a node in the graph. The similarities between the two nodes are stated as edge weights.

2.4.3 (a) Graph based Min Cut Method

A graph cut is a partition of the graph directly or indirectly into two disjoint subsets. The graph is partitioned into clusters using the min cut method. Each cluster is considered as an image segment. The min cut method uses the highly connected subgraphs (HCSs) algorithm to find the clusters [41]. It can be formulated as follows:

$$cut(X,Y) = \sum w(i,j) \tag{2.4}$$

where $i \in X, j \in Y$, and X and Y are two partitioned disjoint sets.

2.4.4 Active Contour Techniques

Active contour models (or snakes) aim to delineate the objects using the energy-minimization function. It is performed by assigning the object boundary as the initial contour, and subsequently evolving the contour to detect the desired object by driving image forces [42].

2.4.4 (a) Active Contour without Edges (Chan–Vese)

Among many active contour methods, the active contour without edges, known as the Chan–Vese method, is widely used in cell segmentation. It helps to detect the objects without a gradient. It has the ability to segment smoothed contour objects by shrinking the contours and works well on convex objects.

2.5 A Novel Hybrid SLIC-K Means Nuclei Segmentation Algorithm

The novel hybrid algorithm is based on the combination of superpixels and K Means clustering method. Firstly, we perform the superpixels segmentation to group the small portion as the pre-segmentation step. Superpixels fragment the image into a set of structurally meaningful segments where the boundaries of each segment take into the consideration of the edge information in the original image. Superpixels have been used as the preprocessing stage in object recognition and medical image segmentation tasks. Among various superpixels segmentation techniques, we have considered Simple Linear Iterative Clustering (SLIC) algorithm because SLIC generates the compact super-pixels with the more regular shape (R. Achanta et al. [43]). By breaking the image into regularly shaped superpixels, it is easier to distinguish between the nuclei and background depending on the superpixels shape. Moreover, SLIC is simple to implement. It requires only the number of desired superpixels as the input parameter and needs low computation time compared to other superpixels techniques [44]. SLIC generates the compact, uniform superpixels by clustering pixels based on their color similarity and proximity. This is done in the combine 5-dimensional space $[labxy]$, where l, a, b is the pixel color vector in LAB color model and pixel position $(x, y\ coordinate)$. SLIC takes as input the desired number of approximately uniform superpixels N. For an image having T pixels, the approximate size of each superpixel is therefore T/N pixels. For roughly equally-sized superpixels there would be a superpixel center at every grid interval S = $\sqrt{T/N}$. The SLIC algorithm is briefly summarized in Algorithm 11.

Algorithm 11: SLIC Superpixels Segmentation

Input: an image $f(x)$, the desired number of superpixels, the compactness of superpixels

Output: a labeled image $u(x)$ where pixels with the same label belong to the same superpixels

1. Initialize cluster centers $C_k = [l_k, a_k, b_k, x_{k,} y_k]^T$ by sampling pixels at regular grid steps S.
2. Perturb cluster centers in an $n \times n$ neighbourhood, to the lowest gradient position.
3. Repeat
4. For each cluster center C_k do
5. Assign the best matching pixels from a 2S × 2S square neighborhood around the cluster center according to the distance measure (Eq. 2.5).
6. End
7. Compute new cluster centers and residual error E {L1 distance between previous centers and recomputed centers} (Eq. 2.6)
8. Enforce connectivity.
9. until $E \leq threshold$
10. Generate the labeled image with superpixels

$$d_{lab} = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2}$$

$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}$$

$$D_s = d_{lab} + \frac{m}{S} d_{xy} \tag{2.5}$$

Where D_S is the sum of the lab distance and the xy plane distance normalized by the

grid interval S, and m is the compactness of a superpixels. The greater m m is, the more compact the cluster and the spatial proximity is more emphasized.

$$G(x,y) = \| I(x+1,y) - I(x-1,y) \|^2 + \| I(x,y+1) - I(x,y-1) \|^2 \quad (2.6)$$

Where $I(x,y)$ is the lab vector corresponding to the pixels at the pixels coordinate (x,y), and $\|.\|$ is the L2 normalization. This take into account both color and intensity information.

Once SLIC generated the superpixels, we converted RGB image into L*a*b* color channel (See Appendix A) and determined the median color feature of each superpixel region in the L*a*b* color space. K Means clustering [45] is utilized to classify the color feature of each compact superpixel into nuclei or non-nuclei, rather than having to perform clustering over the full original image pixels. Pre-segmenting SLIC superpixels before K Means clustering allows us to preserve the natural shape of the cell nuclei since representing the image by SLIC superpixels can give more accurate boundary information than representing the image by pixels. Also, it can reduce the complexity of the algorithm dramatically. This happens because the number of super pixels is much smaller than the number of pixels. Hence, applying K Means clustering on SLIC superpixels, rather than on pixels, can improve the algorithm efficiency and lead to rapid computation. The processing steps of SLIC-K Means based nuclei segmentation method is given in algorithm 12. The visual results of nuclei segmentation on different images are illustrated in Figure 2.5(c) and Figure 2.5(d).

Algorithms 12: Hybrid SLIC- K Means Nuclei Segmentation

Input: Pre-segmented superpixels image

Output: Extracted nuclei image

1. Preprocess the images to enhance the image quality.
2. Convert RGB to L*a*b color model
3. Compute the super pixels over-segmentation using SLIC super pixel method (**Algorithm 11**).
4. Build a cell array of the set of pixels in each superpixels region.
5. Extract the median color features of each region in L*a*b color space.
6. Cluster the color features of superpixels using K Means (k=2)
7. Generate the nuclei segments

2.6 Post-processing

Post processing is an important step to optimize the segmentation results. While most of the segmented regions obtained through the segmentation step will likely correspond to the nuclei regions, there may also be the existence of false findings such as blood cells, artefacts, etc. which must be filtered out. Therefore, it is essential to remove those spurious regions and retain valid nuclei. A series of morphological operations is utilized to remedy the above problems. Firstly, the morphological filtering method is applied to remove the small objects since the artefacts and blood cells are usually smaller than the actual nuclei. The processing of eliminating the spurious objects is given as Algorithm 13.

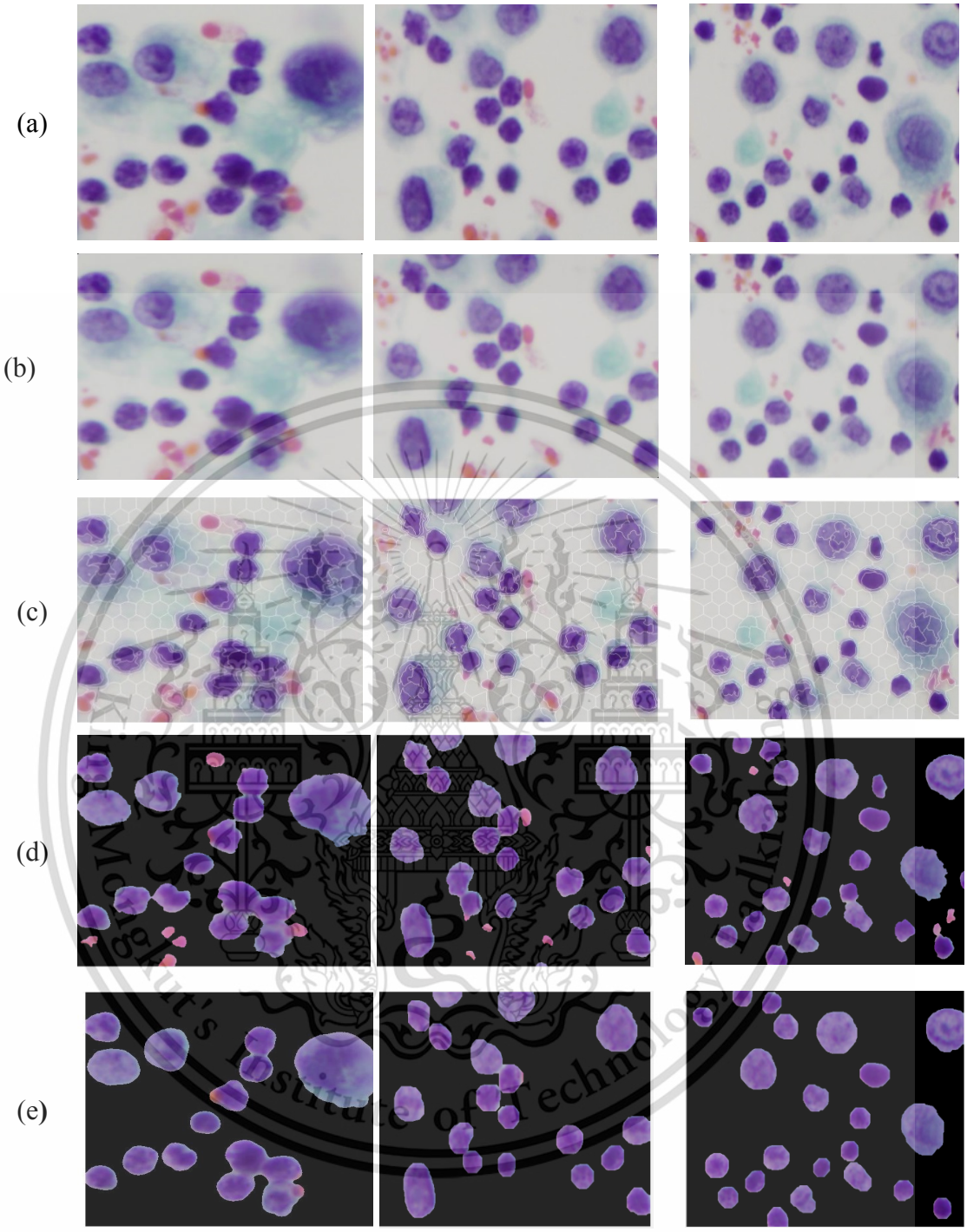


Figure 2.5 Visual results of segmenting cell nuclei from CPE images (a) original image, (b) preprocessed image, (c) superpixels segmentation using SLIC, (d) K Means based unsupervised color segmentation on SLIC superpixels, and (e) post-processed image (refinement of nuclei boundary and elimination of false findings)

Algorithm 13: Removing the false findings

Input: Segmented nuclei regions by SLIC-K Mean algorithm

Output: Actual nuclei regions ($Actual_{nuclei}$)

1. Determine the connected components using 8-connectivity.

2. Count the number of components (N).

3. Compute the area of each component $A_n, n \in N$.

4. Remove small objects using the pre-determined value (P) as follows:

$$\textbf{\textit{For}}\ i = 1{:}N$$
$$If\ \ A_i \geq P$$
$$Nuclei_{Mask} = A_i;$$
$$End$$
$$\textbf{\textit{End}}$$

As described in the above pseudo code, it is required to specify the size of P which is the threshold between the actual nuclei and the spurious regions. The optimal value of P is empirically set as1500 pixels. After removing false findings, we further applied the morphological closing and opening operations for nuclei shape's refinement and simplification.

An important consideration of applying morphological operations is the size and shape of the structuring element (SE). SE identifies the pixels in the image being processed and also designates the neighborhood to be employed in the processing of each pixel. There are two parameters (shape and radius) of SE to be specified. In our algorithm, both opening and closing operations are achieved by using disk shape with the radius 'n' of SE. The radii 'n' of the SE should be specified according to the size of the undesired objects to be removed [46]. However, it is difficult to set the radii 'n' of SE that can work well across all images in the dataset, or, across different nuclei within an image. The optimal radius should be closely related to the size of the false findings

that need to be eliminated. Setting a large structuring element size oversimplifies the image; while using too small SE undersupplies the images (blood cells or noises remain). Hence, we apply a multiscale approach. It means that each image is processed with different radii of SE. For the opening operation, we adapted the range of SE radii to be n {7, 8,….., 15}, which corresponds to the approximately expected range of undesired objects in the pleural effusion cell nuclei. For the closing operations, small SE (half of SE radii in opening operation) sized is adopted. The morphological opening and closing operations are mathematically formulated as follows:

$$Actual_{nuclei} \circ SE = (Actual_{nuclei} \ominus SE) \oplus SE \tag{2.7}$$

$$Opened_{nuclei} \cdot SE = (Opened_{nuclei} \oplus SE) \ominus SE \tag{2.8}$$

Where $\oplus$and $\ominus$ represent the dilation and erosion, respectively.

2.7 Experimental Results and Discussions

2.7.1 Experimental Results

The experiment is carried out using 35 CPE images. To set the gold standard, the ground truth images were prepared with the help of experts from the hospital. First, computer vision researchers manually delineated the cell nuclei. The experts then verified and annotated the cell nuclei. To quantitatively evaluate the segmentation methods, five pixel-based performance metrics namely precision, recall, F measure, Jaccard Index (JI), and Dice Similarity Coefficient (DSC) were computed for each algorithm. The performance measures are computed by matching the pixels in the segmented image and the pixels in the ground truth image. Each connected region in the segmented results is considered as one nucleus while ignoring the number of

nucleus inside the region. Ground truth images are also prepared in the same way. The performance measures can be formulated as follows:

$$\text{Precision(Pre)} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \tag{2.9}$$

$$Recall(Re) = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2.10}$$

$$\text{Fmeasure(Fm)} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.11}$$

$$\text{JI} = \frac{\text{TruePositive}}{TruePositive + FalsePositive + FalseNegative} \tag{2.12}$$

$$\text{DSC} = \frac{2 * \text{TruePositive}}{2 * TruePositive + FalsePositive + FalseNegative} \tag{2.13}$$

The visual results of different segmentation methods are given in Figure 2.6. Figure 2.6 (b) depicts the sample of ground truth image. The quantitative results of those methods are reported in Table 2.1. The compared quantitative results show that the segmentation performances of the Otsu's method, k-means, mean shift, Chan–Vese level set method, graph-based min cut and proposed hybrid SLIC-K Means are excellent, whereas the proposed method achieved the best accuracy compared to the rest. The accuracies meet clinical requirements. For the highlighted methods, we further evaluated the nuclei detection rate (NDR) of the images depending on the recall value. The NDR is considered as true positive when the recall is greater than 60%. The overall NDR of each algorithm is computed and compared, as shown in Figure 2.7 (a). Similar to the NDR, we estimated the abnormal NDR. Figure 2.7 (b) shows the comparison results. The comparison results show that hybrid SLIC-K Means and the mean shift

clustering method exhibited the best performance in terms of the overall NDR and abnormal NDR. To evaluate the time complexity of each method, the computational time of each method is computed and compared, as shown in Figure 2.8. Otsu's method is found to be relatively simple and fast. In contrast, the Chan-Vese method is computationally expensive. The proposed hybrid SLIC- K Means achieved the better accuracy of segmentation and detection of nuclei while requiring the acceptable computational time. Even though the mean shift clustering yielded good segmentation and detection rate, it is computationally expensive in case of many images are required to process.

Table 2.1 Quantitative Experimental Results of Pixels-Based Evaluation

Segmentation technique	Methods	Performance measures				
		Pre	Re	Fm	JI	DSC
Thresholding	Otsu's method	0.91	0.89	0.90	0.89	94%
	Isodata	0.86	0.84	0.85	0.84	91%
	Maximum Entropy	0.84	0.94	0.88	0.94	91%
	Cross Entropy	0.94	0.82	0.87	0.82	90%
	Minimum Error	0.85	0.82	0.83	0.82	89%
	Fuzzy Entropy	0.68	0.67	0.68	0.67	80%
	Adaptive thresholding	0.96	0.50	0.66	0.50	67%
Clustering	K Means	0.90	0.89	0.89	0.89	94%
	Fuzzy c-means	0.94	0.77	0.85	0.77	77%
	Mean Shift	0.93	0.91	0.92	0.91	95%
Active contour	Chan–Vese method	0.89	0.87	0.88	0.87	94%
Graph based method	Graph-based min cut	0.87	0.95	0.91	0.87	93%
Hybrid superpixels	Hybrid SLIC-K Means	0.93	0.92	0.93	0.92	96%

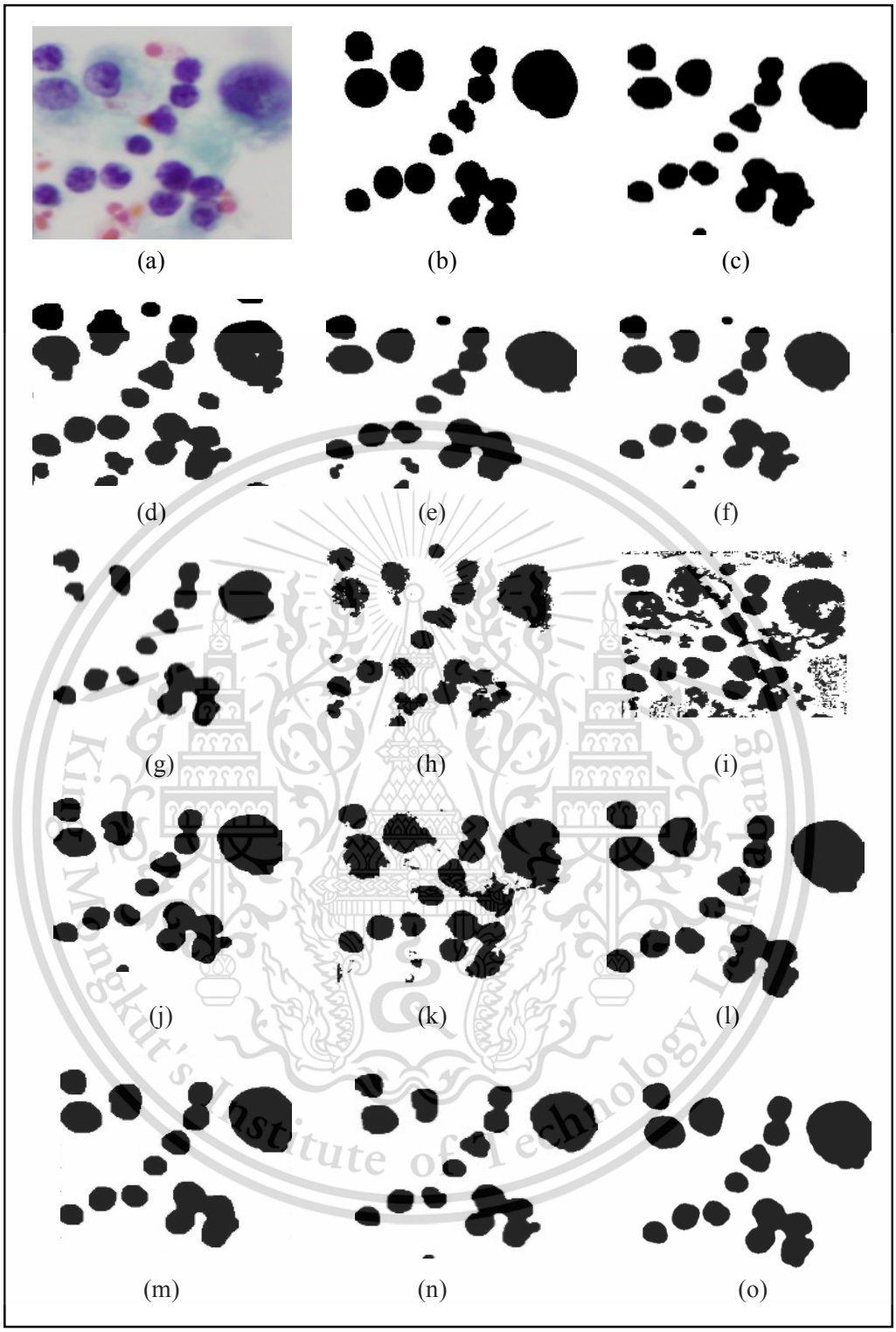
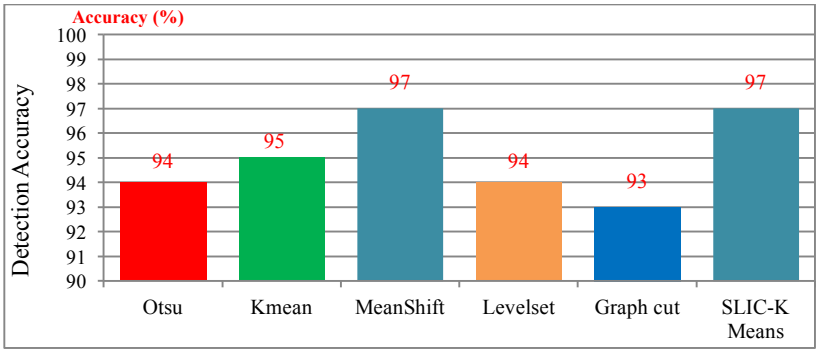
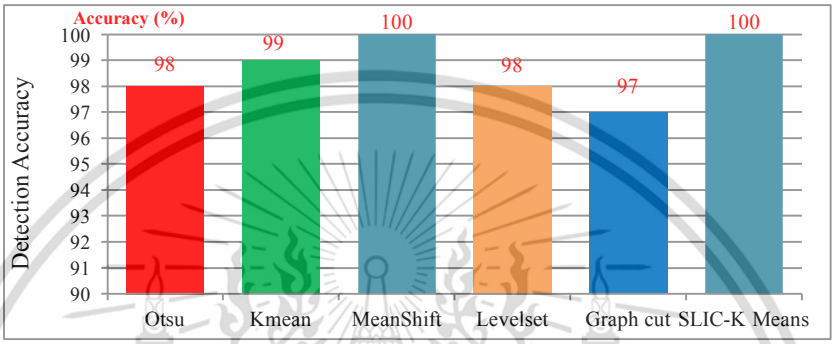


Figure 2.6 Visual results of segmented cell nuclei using twelve segmentation methods and novel hybrid SLIC-K Means algorithm: (a) original image, (b) ground truth image, (c) Otsu's method, (d) Isodata, (e) maximum entropy, (f) cross entropy, (g) minimum error, (h) fuzzy entropy, (i) adaptive thresholding, (j) K Means clustering, (k) Fuzzy c-Means clustering, (l) mean shift clustering, (m) Chan–Vese method (n) hraph-based min cut, and (o) novel hybrid SLIC-K Means (note that: the images were cropped for better visibility here)



(a)



(b)

Figure 2.7 Comparison of nuclei detection rates in terms of the recall value: (a) overall nuclei detection rate, and (b) abnormal cell nuclei detection rate

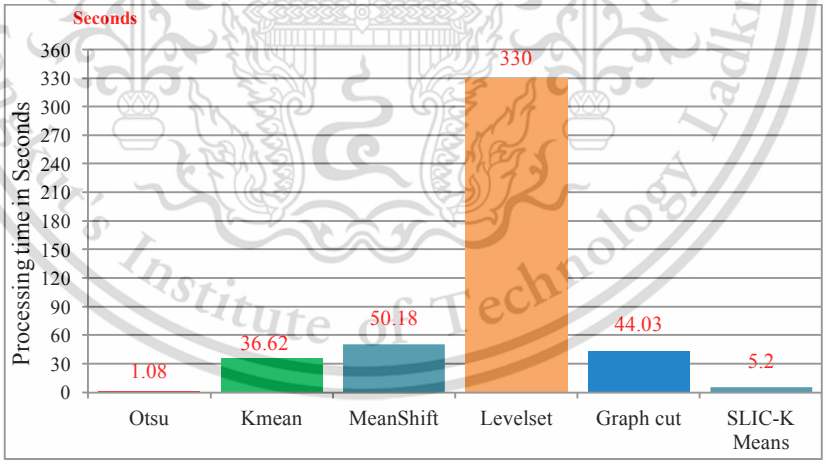


Figure 2.8 Processing time of five highlighted methods

2.7.2 Parameters Tuning and Discussions

The highlighted segmentation methods are discussed along with their adjustable parameters, advantages, and limitations. The experiment results show that the

performances of the segmentation methods strongly depend on the tuning parameters. Therefore, it is required to properly select the most relevant parameters for our applications. We experimentally set and adjusted different parameters in each segmentation method and selected the most effective one for all images.

As Otsu method is a non-parametric method, it is not required to specially assign prior parameters. However, Otsu method is sensitive to outliers. To deal with this issue, the CLAHE and median filter methods are employed to enhance the image quality before applying Otsu method. The cytology pleural effusion images comprise three main parts: cell nuclei, cytoplasm, blood cells, and background. As the color of the nuclei region appears to be dark purple, with other parts appearing lighter in color, the image histogram is assumed to be a bimodal distribution. Thus, Otsu method provides relatively good performance in our application. The Otsu method is relatively simple and the result is promising. Therefore, it can be applied to real applications. However, the performance is degraded when the image contains significant noises because the method is sensitive to noise.

The segmentation result of the K Means clustering method strongly depends on initializing the k clusters. A poor initialization can significantly affect the clustering performance and result in a poor convergence speed. Therefore, we set multiple k clusters for the test and chose the most effective one. The numbers of k clusters were set as 2, 3, 4, and 5. When k is 3, the nuclei are mixed with other image components. Thus, it was difficult to separate only the nuclei regions from the mixed clustered regions. When k is 4, 5, or more, the nuclei are broken into multi-segmented images because of the high variation in the pixel intensity within the nuclei. Hence, the nuclei regions should be obtained from multiple images. When k is 2, the nuclei are segmented in a straightforward manner. In addition, the nuclei appear dark in color with some regions

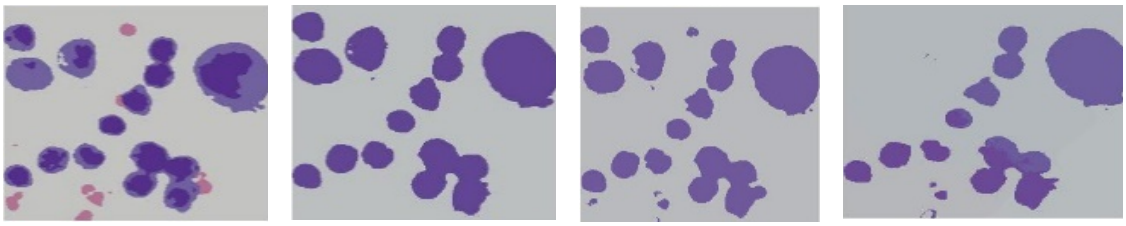
remaining bright. This fact also supports in setting up the value of k as 2 to cluster two groups (dark and bright colors). Figure 2.9 shows the visual segmented cell nuclei with

(a) (b) (c) (d)

Figure 2.9 Image index labeled with different k clusters: (a) k = 2, (b) k = 3, (c) k = 4, and (d) k = 5

different k clusters. Moreover, it is worth noticing that the K Means clustering method performs well for round-shaped objects. Hence, the method is effective in segmenting cell nuclei because the cell nuclei are largely round shaped. In addition, it is simple, fast, and easy to implement. However, the disadvantage is that the K Means clustering is extremely sensitive to the k clusters and performs badly when the clusters are convex shape.

In contrast to the K Means clustering method, the mean shift is a non-parametric clustering method. It is not necessary to define the clusters and restrict the cluster shape. Nevertheless, the clustering result of the mean shift strongly depends on the bandwidth size. It is required to carefully select the most relevant size for particular applications. In our experiments, we experimentally set multiple bandwidth sizes and chose the best one. Figure 2.10 shows the significant differences of the segmentation results in terms of the bandwidth sizes. The experiment results show that a bandwidth size of 0.2 exhibits the best clustering performance, appropriate for cell nuclei segmentation. The advantage of the mean shift is that it is not necessary to initialize the cluster numbers; moreover, the



(a) (b) (c) (d)

Figure 2.10 Different clustering results in terms of different bandwidth (bw) sizes: (a) bw = 0.1, (b) bw = 0.2, (c) bw = 0.3 and (d) bw = 0.4

method need not be robust to outliers. Only the size of the bandwidth is required to be set. The limitation of the mean shift is that it is computationally expensive compared to other clustering methods, as many windows need to be shifted, thus making many computations redundant.

In the Chan–Vese level set method, the boundaries of the regions are used as a mask, which is initial contour the evolution of the segmentation start. To achieve a fast and accurate output, we initially specified the mask that is close to the nuclei regions. The mask either shrinks or expands based on the image features. In addition, it is crucial to specify an appropriate maximum iteration for the contour evolution. The iteration is stopped if the maximum iteration is reached, when the energy remains constant, or when the contour is not moving. The maximum number of iterations affects the largest variation in the segmentation results. Therefore, we experimentally tested with different iterations and chose the most effective one for all images. However, it is the main limitation, as the images contain various types of cells. It is difficult to fix the number of iterations if more images are added into the datasets. If a large number of maximum iterations is set, the computation becomes expensive. In contrast, a small number of iterations lead to under-segmentation, because the iteration is stopped before finishing the contour evolution. Figure 2.11 depicts the segmentation results obtained through different iteration numbers.

(a) (b) (c)

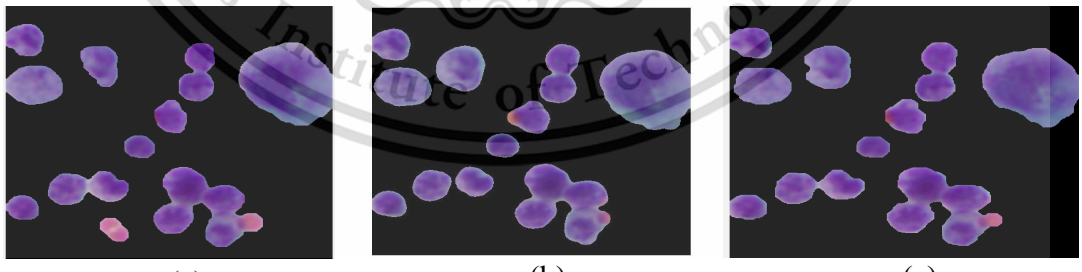
Figure 2.11 Segmentation results obtained using different iterations: (a) iterations = 500, (b) iterations = 300, and (c) iterations =100

In the graph-based min cut segmentation, two parameters need to be defined. The first one is alpha, which is the penalty parameter with respect to the total-variation term. For the case wherein the image-edge weights are incorporated, alpha is given by the constant in all cases. For the case with image-edge weights, alpha is given using the two-pixel wise weight function. The second parameter is the step-size of the augmented Lagrangian method, the optimal range of which is [0.3, 3]. We set it as 0.3, as it is not significantly different for segmentation. A significant variation in the segmentation result is found when setting up different values of alpha. We experimentally tested with different alpha values and chose the best one. Figure 2.12 depicts the visual result of the segmentation in terms of the alpha values. We chose 0.3 as the alpha value in our application. The graph-based min cut method is simple, easy to control, and fast in processing. It returns the clusters as image segments. However, the drawback is that multiple small segments may be separated by cutting small sets of isolated nodes in the graph.

(a) (b) (c) (d)

Figure 2.12 Variation of segmentation results in terms of different alpha values (av): (a) av = 0.3, (b) av = 0.5, (c) av = 10, and (d) av = 15

In our novel hybrid SLIC-K Means based nuclei segmentation, we need to specify two parameters: the number of superpixels for SLIC and k clusters for K Means. The desired number of superpixels is set to 500 through fine-tuning setting. The visual results of different superpixels numbers are given in Figure 2.13. Setting a small number of superpixels will result in poor convergence. In contrast, setting a large number of superpixels will result in heavy computational time. In the proposed method, the number of superpixels is obtained through empirical parameter tuning. It was set as 500. Similar to classical K Means based nuclei segmentation, k is set as 2 because cell nuclei are segmented in a straightforward way when k is 2.

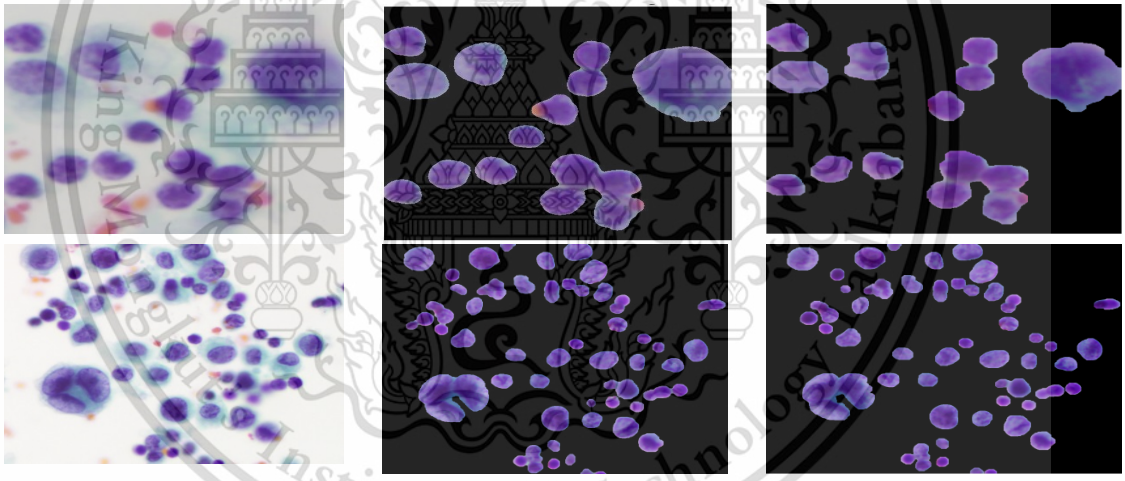


(a) (b) (c)

Figure 2.13 Variation of segmentation results in term of different number of super pixels: (a) superpixels =100, (b) superpixels=500 and (c) superpixels =1000

In SLIC-K Means, SLIC method is firstly performed to pre-segment the image into the small impact superpixels. Then, K Means clustering is carried out to cluster

each super-pixel into two groups by using the extracted features from super-pixels. Features extracted over the uniform and compact SLIC superpixels tend to be more discriminative, helping K Means to produce better segmentation. The good adherence to image boundaries exhibited by SLIC super-pixels result in smoother and more accurate segmentation. Utilizing K Means clustering on the superpixels can shorten the computation because the number of super pixels is lesser than the number of pixels. It scales up linearly in computational cost and memory usage. The proposed segmentation method extracts the cell nuclei at a lower computational cost and preserves the natural shape of the cell nuclei while achieving the accurate segmentation results. The segmentation results obtained by classical K Means and hybrid SLIC-K Means are given in Figure 2.14.



(a) (b) (c)

Figure 2.14 Comparison results of nuclei segmentation methods: (a) original image, (b) proposed novel hybrid algorithm (SLIC-K Mean), and (c) K Means clustering based segmentation

2.8 Summary

This chapter presented a comparative analysis of traditional segmentation methods and a novel hybrid algorithm for segmentation of cell nuclei in CPE images. In the comparative analysis, we evaluated twelve image segmentation methods that are

(1) Otsu method, (2) Isodata thresholding method, (3) maximum entropy thresholding method, (4) cross entropy thresholding, (5) minimum error thresholding, (6) fuzzy entropy thresholding method, (7) adaptive thresholding method, (8) K Means clustering, (9) fuzzy c means clustering, (10) mean shift clustering, (11) Chan-Vese level set and (12) graph cut methods to extract the cell nuclei from the surrounding objects such as cytoplasm, blood cells, and artefacts. Regarding development of a novel hybrid SLIC-K Means nuclei segmentation algorithm, its novelty stems from a SLIC superpixels segmentation method as the pre-segmentation step to minimize the computational time of K mean clustering and to cluster the nuclei regions while preserving their natural shape. The algorithms also embed preprocessing process to enhance the image quality and filter out the small noises before segmentation. They also comprise of the post-processing process to eliminate any false findings and refine the segmented nuclei based on size and shape to ensure only the true nuclei are retained. This chapter presented an empirical evaluation of the performance of the algorithms and discussed the results. The performance of the methods is evaluated mainly based on six pixel-based evaluation metrics. The object-based overall nuclei detection and abnormal nuclei detection are also judged to ensure the effectiveness of each algorithm. In addition, the processing time is also computed to judge the complexity of each algorithm. The experimental results reveal that our proposed hybrid SLIC-K Means algorithm outperformed twelve traditional segmentation algorithms tested in this study by given the average segmentation accuracy 96% with less computation time.

CHAPTER 3

DETECTION OF OVERLAPPING NUCLEI AND DECOMPOSITION INTO ITS CONSTITUENTS

3.1 Introduction

Most of CPE images contain nuclei that overlap to different degrees. Due to the close resemblance between overlapping and cancerous nuclei, misclassification of the overlapped nuclei can affect a CAD system's final decision. For instance, the excessive enlargement of nuclei and their irregular shapes are highly suggestive of malignancy. Due to the excessive enlargement of size and irregular shapes of overlapping nuclei, CAD systems may misclassify them as malignant cells. Moreover, cytologists mostly focused on the nuclei morphology, e.g., size, shape, and density to determine the cell malignancy. Thus, the overlapping nuclei are required to be accurately delineated. Detection and separation of overlapped cell nuclei is a high priority task for the optimal segmentation. Although human experts find little difficulty in differentiating between single and overlapping nuclei, it is still a challenging task for automatic systems. The appearance of overlapping nuclei in CPE images are dark purple regions with a close resemblance among them. Thus, a CAD system may wrongly interpret them as single nuclei. It is difficult to retrieve and quantitatively analyze features such as nucleus morphology and density if cells are touching, overlapping or clustered. Thus, overlapping nuclei should be detected and decomposed into its constituents for retrieving the accurate measurements of each nucleus and avoiding the misclassification of cancerous cells.

3.2 Related Works

Many studies have sought several methods to delineate the interregional contours of overlapping cell nuclei or to isolate them into individual ones. In the literature, the most widely used overlapped objects isolation methods are watershed methods [47,48] and concavity analysis based methods [49,50]. Those methods have contributed a lot to the field of microscopic image analysis. Recently, Kumar et al. proposed a rule-based clump separation technique for decomposing the overlapping nuclei into their constituents [51]. In another study, Wang et al. proposed a bottleneck rule method for separating the overlapping cells [52]. These previous studies exhibit that there has been a great interest in the decomposition of overlapped cells in microscopic image analysis. However, most of those methods attempt to detect the splitting points first and then decide whether to split a connected component at these split points afterward. Detecting splitting points on all nuclei which may include non-overlapping nuclei may cause heavy computational cost and over-splitting. It is our idea that we first judge whether there is a presence of overlapping nuclei through the pre-determination procedure and, if positive, apply suitable splitting algorithms to isolate the overlapping nuclei.

Thus, it is crucial to accurately determine the presence of overlapping nuclei prior to the occurrence of any splitting process. Some studies have presented the methods to differentiate between the overlapping and single nuclei. Tafavogh et al. [53,54] reported a technique to identify the presence of overlapping nuclei on cellular images of neuroblastoma. Mean shift method is utilized to segment the nuclei, and three size-and shape-based features of cells namely (i) area, (ii) diameter equality, and (iii) concavity dominance are extracted to distinguish the overlapped nuclei from the single ones using step-by-step conditional filtering. Abbas et al. [55] presented a framework

for identifying the overlapping nuclei on red blood cellular images before splitting them. First, a thresholding method is applied to binarize the image, then three features, namely (i) convex hull, (ii) area, and (iii) elongation, are extracted. The predetermined value of each feature through parameter-tuning is applied to judge the presence of overlapping nuclei. Wang et al. [52] proposed a framework to identify overlapping nuclei using shape features. Five shape features namely (i) solidity, (ii) convexity, (iii) eccentricity, (iv) area, and (v) variance are used as input to an SVM classifier to differentiate between single and overlapping cell nuclei. The method has given the accuracy of 86%, 90%, 88%, and 88% for oil cells, yeast cells, blood cells, and curvularia cells datasets, respectively. Guven et al. [56] introduced an unsupervised data-clustering method for the detection of overlapping cell nuclei on Pap smear cervical images. The morphological operation is applied to extract the nuclei borders, and the fuzzy clustering method is employed to differentiate between individual and overlapping nuclei using three shape features and two local minima features. The performance of the algorithm is assessed based on 290 nuclei and gives 79.1% F-score, 67.4% recall, and 95.7% precision. The reported techniques in [53-55] are parameter-dependent and limited to objects with a great variation of size and shape. The method reported in [52] used the shape and size features only. In the case of CPE images, the forms of overlapping nuclei vary greatly. Thus, it can be deduced that considering only size and shape features may not be sufficient for classifying between individual and overlapping nuclei. The method presented in [56] is an unsupervised clustering method to determine the presence of overlapping nuclei using a combination of shape and local minima based features and provides acceptable performance. Nevertheless, the method is conceived specifically for cells on the pap smear cervical images. It cannot be taken for granted that this method will provide good results with pleural effusion cells. The

originators of the aforementioned method did not take into consideration the textural pattern difference between single and overlapping nuclei despite the fact that the texture pattern of single and overlapping nuclei varies greatly. Moreover, supervised learning techniques could greatly help to attain a more accurate detection rate [57]. The textural patterns and supervised machine learning methods are mainly considered to detect the overlapped nuclei. After detecting them, we employed the splitting algorithm to decompose the detected overlapped nuclei into their constituents.

3.3 The Proposed Algorithm for Detection of Overlapping Nuclei and Decomposition into Their Constituents

In this study, we present a new integrated algorithm for the accurate delineation of overlapping nuclei. Our proposed method comprises of two stages: (1) determination of the presence of overlapping nuclei and (2) decomposition of the detected overlapped nuclei into their constituents. In the first stage, we propose a novel algorithm to determine the presence of overlapped nuclei. The distinct of our method from the previous studies is that we introduced a new combination of geometrical and textural features, and a double strategy random forest is used as an ensemble feature selector to select the most discriminant features and as an essential classifier to differentiate between single and overlapping nuclei. The detected overlapped regions from the first stage will be fed to the second stage which decomposes them into their constituents with the help of concavity analysis. Figure 3.1 demonstrates the block diagram of the detection of overlapping nuclei and decomposition into their constituents.

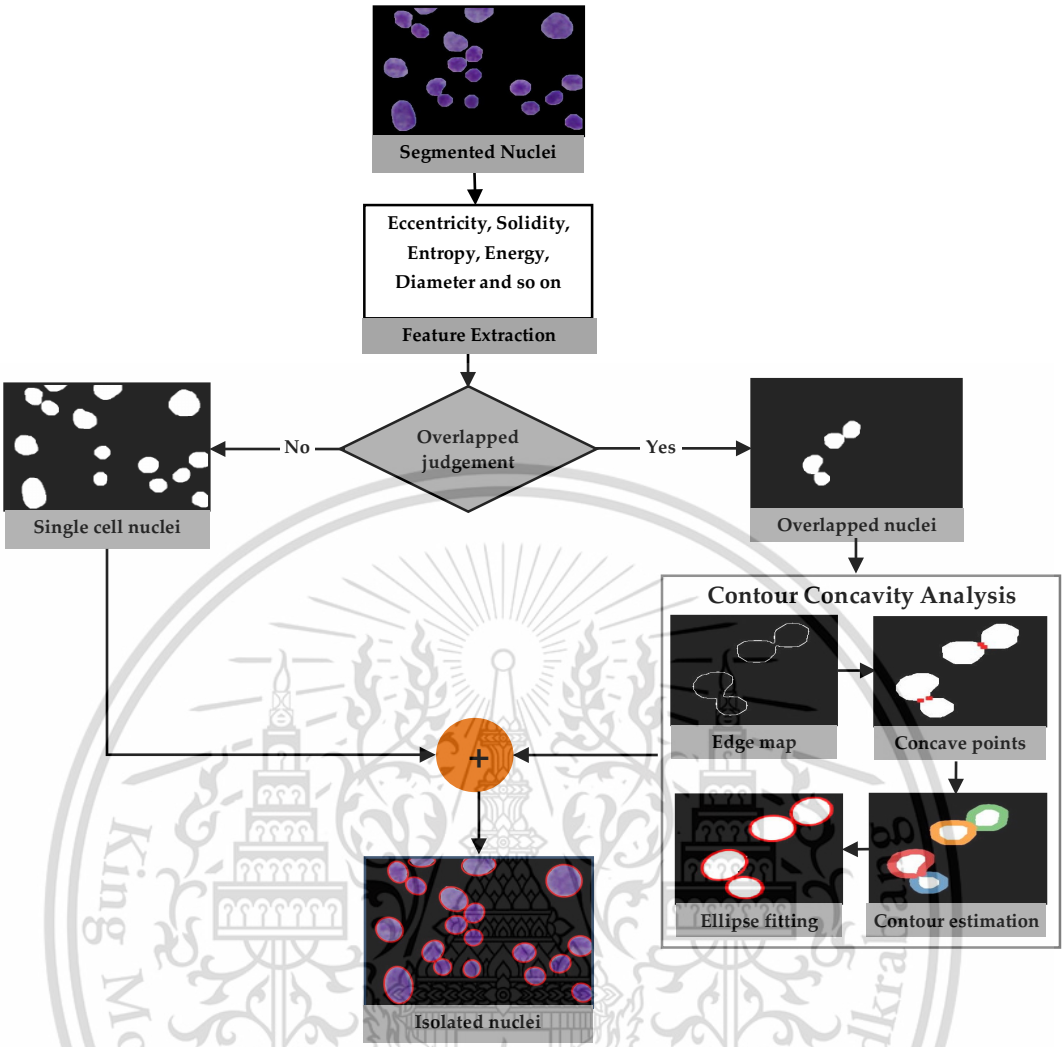


Figure 3.1 The block diagram of detection of overlapping nuclei and decomposition into its constituents

3.4 First Stage: Detection of Overlapping Nuclei

3.4.1 Feature Extraction

Extracting rich and semantically discriminative features from nuclei is of paramount relevance to advancements in the classification of individual and overlapping nuclei. As shown in Figure 2.3, it can be seen that there are different forms of overlapping nuclei such as light touching, multi-nuclei touching, multi-nuclei overlapping, and cohesive tight clusters. Therefore, depending solely on size and shape features may not be sufficient for the robust detection of overlapping nuclei in CPE

images. It is our observation that the textural pattern between single and overlapping nuclei varies greatly. Thus, we considered textural features, as well as geometric features (i.e., size and shape features), and propose a new combination of them [58] to distinguish overlapping nuclei from single ones. A total of 26 features (i.e., 16 geometric and 10 textures) are extracted from each segmented region. They are listed in Tables 3.1 and 3.2.

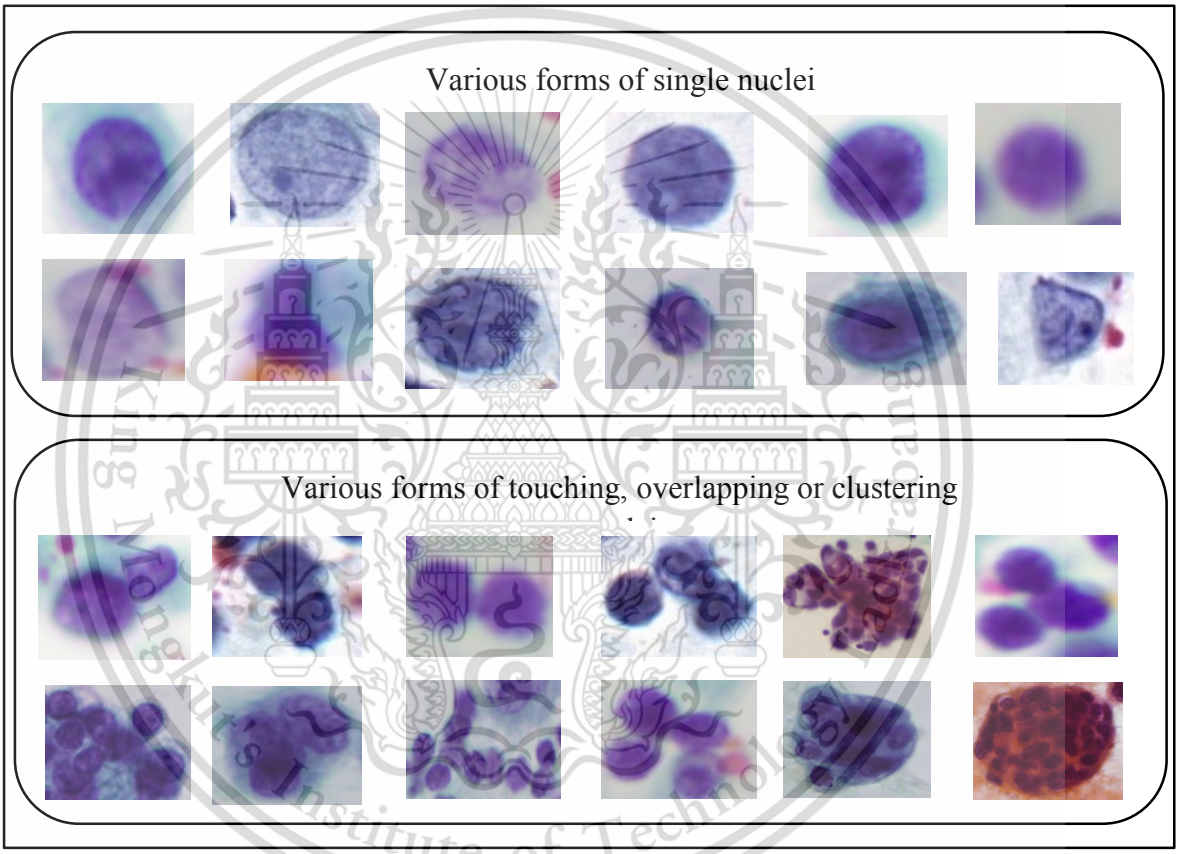


Figure 3.2 Various forms of single and overlapping nuclei in CPE images

Table 3.1. Extracted geometric features.

No.	Feature Name	Description
1.	Area (A)	This is denoted as the actual number of pixels inside the nucleus region.
2.	Perimeter (P)	This is measured by computing the total number of pixels on the nucleus edge.

3.	Roundness	This is defined by $\frac{4\pi \times area}{perimeter^2}$, which represents the similarity between the nucleus region and a circle. It varies between 0 and 1 and a circle's roundness circularity is equal to 1.
4.	Solidity	This specifies the proportion of the pixels in the convex hull that is also in the nucleus region. It is formulated as; $\frac{Area}{ConvexArea}$.
5.	Equivalent Circular Diameter (EDC)	This is defined as the diameter of a circle with the same area as the nucleus region. It is represented using; $\sqrt{\frac{4 \times Area}{pi}}$.
6.	Compactness	This specifies the ratio of area and square of the perimeter. It is computed as $\frac{Area}{perimeter^2}$.
7.	Eccentricity	This represents the eccentricity of the ellipse that has the same second-moments as the nucleus region. Its value is between 0 and 1. A cell whose eccentricity is 0 is a circle, while 1 is a line segment.
8.	Local minima	This is the number of local minimum points in the nucleus region.
9.	Aspect ratio of the nucleus:	This is denoted by the ratio of nucleus width to nucleus height using; $\frac{Width_{nucleus}}{Height_{nucleus}}$.
10.	Major Axis	This represents the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the nucleus region.
11.	Minor Axis	This specifies the length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the nucleus region.
12.	Elongation	This is represented by the ratio between the major and minor axis using; $\frac{majoraxis}{minoraxis}$.
13.	Actual Diameter (AD)	This is represented by the circle's diameter circumscribing the nucleus region. It is formulated as; $\frac{perimeter}{2 \times pi}$.
14.	ECD to AD	It is defined as; $\frac{ECD}{AD}$.
15.	Convex Area	This represents the number of pixels in the convex nucleus.
16.	Maximum distance between local minima	This is measured by computing Euclidean distance between local minima points and assigning the maximum distance as the region's maximum distance property.

Table 3.2. Extracted textural features

No.	Features	Description
1.	Mean	The mean gray values inside each segmented region.
2.	Standard deviation	The deviation of gray values inside each segmented region.
3.	Smoothness	The local variation in radius lengths of each segmented region.
4.	Variance	The variance value of the gray values inside each segmented region.
5.	Skewness	The skewness of gray values of each segmented region.
6.	Kurtosis	The kurtosis of gray values of each segmented region.
7.	Energy	Energy of gray values of each segmented region.
8.	Entropy	Entropy of gray values of each segmented region.
9.	Entropy	Entropy of entropy filtered image.
10.	Entropy	Entropy of standard deviation filtered image.

3.4.2 Classification

There may be noisy and irrelevant features in the initial feature set. Using such features may degrade the performance of a classifier and need intensive computational complexity. In bioinformatics applications, there are two ways to improve classification accuracy. They are: selecting the most relevant features and choosing the best-suited classifier. In this study, we use a double-strategy random forest algorithm to deal with these two issues. The reason for utilizing random forest is that it provides favorable results for unbalanced data classification, and is robust in dealing with noisy data. Our dataset of nuclei was highly unbalanced. The number of overlapping nuclei constitutes only a very small minority of the dataset at 625 (16%) and, in contrast, the number of individual nuclei composes an abundant majority of the dataset at 3275 (84%). Random forest (RF) is one of the most successful ensemble classification models which was proposed by Ho [59,60], and later by Breiman [61]. RF is an ensemble of decision trees which integrates the idea of Ho's "bagging (bootstrap aggregation)" and Breiman's

"random variable selection". The principle of RF is to create multiple decision trees using randomized bootstrapped samples from a learning dataset and randomly selecting a subset for training data. Each decision tree, also known as a Classification And Regression Tree (CART), is grown using randomized bootstrap samples of input data and generates its own classification results. Finally, majority voting is applied to aggregate the predictions of all trees in RF. The schematic diagram of RF classifier is shown in Figure 3.3. Observations that are not contained in training bootstrap samples are "out-of-bag" (OOB), and they are used for predicting errors. Figure 3.4 shows data utilization in constructing RF. RF is widely used as a feature selection algorithm [62,63] and classifier [64–66] in medical diagnosis analyses. In this study, we utilize RF to select the discriminant features and classify between single and overlapped nuclei. First, we applied RF-based ensemble feature selection to sort the importance of features based on OOB error permutation and select the most important features. Then, we train RF ensemble classifier using the selected features and the trained RF classifier is utilized to predict the new data from the testing dataset.

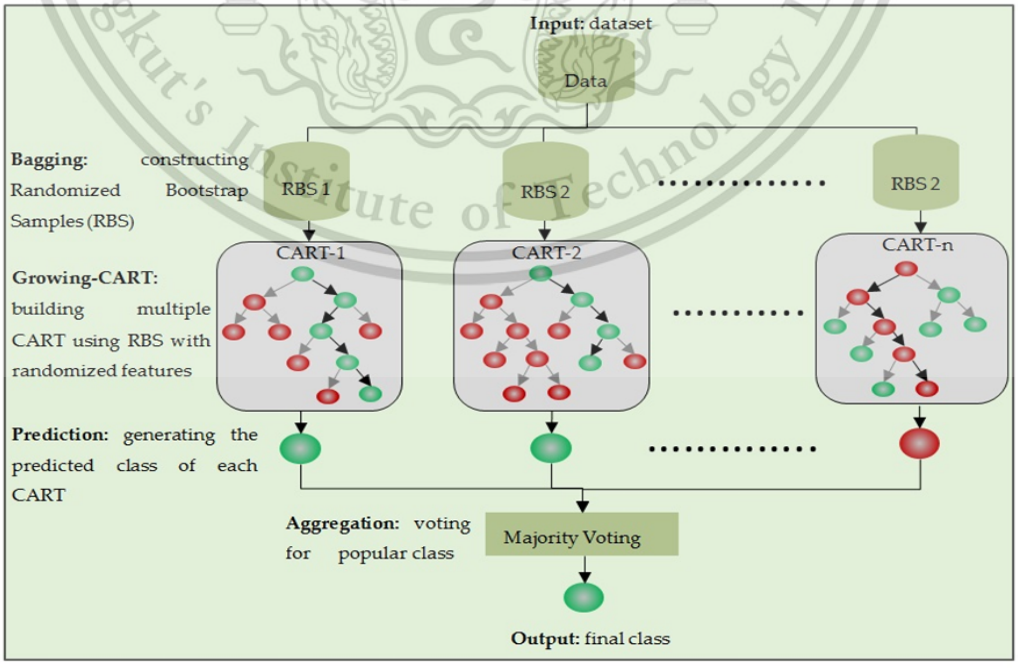


Figure 3.3 The schematic diagram of a RF classifier

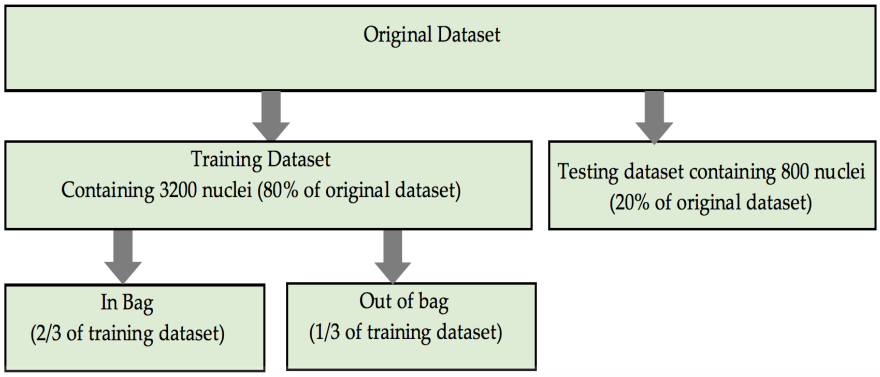


Figure 3.4 Data utilization in constructing a RF classifier

3.5 Second Stage: Decomposition of Detected Overlapping Nuclei into Its Constituents

The detected overlapped nuclei are used as input to the splitting algorithm. We utilized Contour Concavity Analysis (CCA), introduced in [67], to decompose the overlapped nuclei into its constituents. CCA involves contour evidence extraction and contour estimation. The flowchart of CCA based overlapped nuclei splitting method is depicted in Figure 3.5. The contour evidence extraction comprises of two sub-processes: contour segmentation and grouping. In contour segmentation, canny edge method is utilized to extract the edge map. Then, the curvature scale space (CSS) method based on curvature analysis is applied to detect the corner points of the object boundaries. CCS based corner points detection produces the points with the maximum curvature lying on both concave and convex regions of object contours. Since being only interested in the concave points joining the contours of overlapping objects, the detected corner points are examined if they lie on concave regions. Let us denote a detected corner point by p_i, and its two kth adjacent contour points by $p_{i-k}$and p_{i+k}. The corner point p_i is qualified as concave if the line connecting p_{i-k} to p_{i+k} does not

reside inside the object. The obtained concave points are used to split the contours into contour segments.

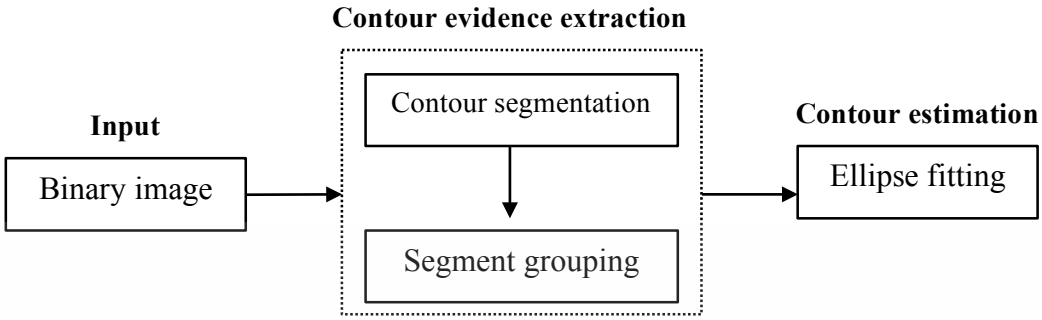


Figure 3.5 The flowchart of CCA for splitting the overlapping nuclei

Due to the overlap between the objects and the irregularities in the object shapes, a single object may produce multiple contour segments. Then, segment grouping is performed to merge all the contour segments belonging to the same object. It is an iterative procedure that repeats over each pair of contour segment, examining if they can be combined. In order to limit the search space, the contour segment under the grouping process is carried out for two neighbor contour segments wherein the Euclidean distance between their center points is less than the predefined threshold value. Then, the grouping process is performed through the process of the ellipse fitting (See Appendix B). It groups contour segments that compose an object with ellipse shape. When the contour evidence is acquired, the contour estimation is carried out using a stable direct least square fitting method. Each processing step of CCA is depicted in Figure 3.6

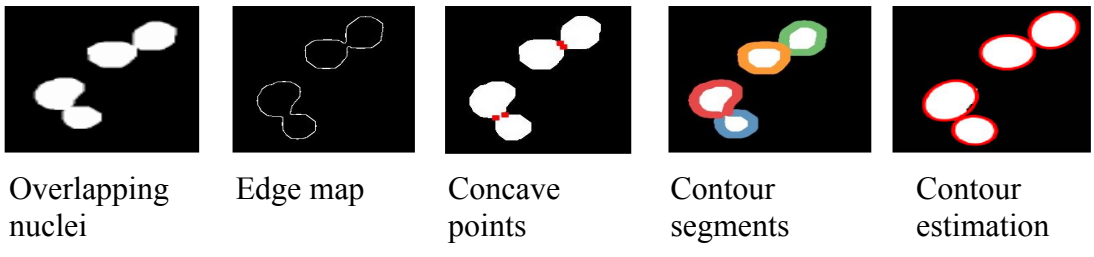


Figure 3.6 The processing steps of CCA for splitting the overlapping nuclei

3.6 Experimental Results and Discussion

Almost all the images in the studied dataset present the overlapped cell nuclei in different degrees. Separating them into individual ones is hence essential. In the literature, almost all the related work directly applied the cell splitting method on the entire segmented image. It means that the splitting method is processed not only on the overlapped region but also on the single cell nuclei regions. Such an attempt can lengthen the computation time. In contrast, we proposed the sequential integration of overlapped nuclei detection and decomposition to identify the overlapped areas and isolated them into individual ones. First, overlapped nuclei are detected with the help of a new combination of geometrical and textural features and a double-strategy random forest and, the detected ones are isolated using the contour concavity analysis. The experiments were carried out in a MATLAB_R2016b environment using an Intel(R) Core (TM) i7 CPU 3.40–3.70 GHz personal computer and Microsoft Windows 7, 64-bit operating system. To obtain a comprehensive discussion, the experimental results are divided into two stages: detection of overlapping nuclei and decomposition them into its constituents.

3.6.1 Detection of Overlapping Nuclei

From the segmented nuclei, we extracted a new combination of 16 geometrical and 10 textural features. The dataset is made up of 4000 nuclei × 26 features dimensional array. It is fragmented into training and testing sets in an 80/20 ratio, as given in Table 3.3. The training set is used for building the learning model and testing set for assessing the performance of the learning model.

Using the training set, a double-strategy RF was applied to select the dominant features and to classify between single and overlapping nuclei using selected high-

ranking features. One of the important parameters that we needed to adjust while constructing RF was the number of decision trees to be grown. The empirical tuning is performed to determine the optimal number of decision trees. Figure 3.7 shows OOB errors using a different number of decision trees. The graph shows that OOB errors are decreased starting from 250 decision trees, and started to stabilize from 300 trees. Therefore, 400 decision trees are grown in constructing RF in order to maintain classification stability and keep the computation cost low.

Table 3.3 Data partition in the detection of overlapping nuclei

Observational Data	Training	Testing	Total
Single Nuclei	2692	683	3375
Overlapped Nuclei	508	117	625
Total	3200	800	4000

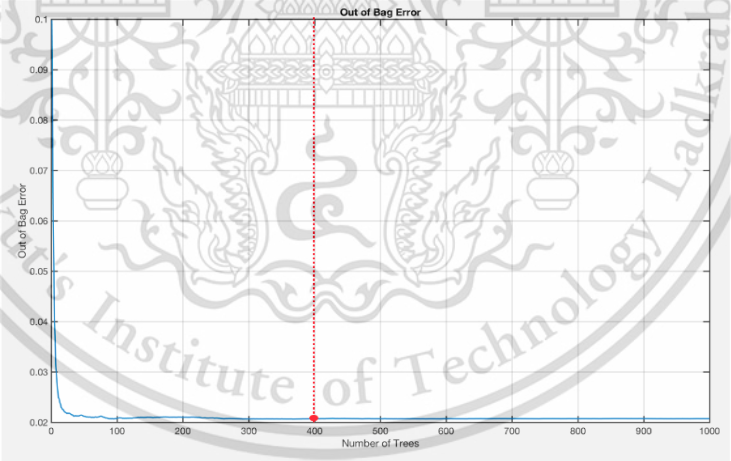


Figure 3.7. Out-of-bag (OOB) error of RF using different number of decision trees

When RF classifier was designed with 400 decision trees, it is used as an ensemble feature selector to rank the features by scoring OOB permutation errors of each feature. The relative importance of features by using ensemble RF feature selector is given in Figure 3.8. To choose the dominant features, the different number of feature in ascending rank order are used to examine their training accuracy as given in Figure

3.9. As seen from the chart, it reveals that the first eight ranked features provided the highest training accuracy. They are (i) energy, (ii) variance, (iii) equivalent circular diameter to diameter, (iv) eccentricity, (v) ratio between area and perimeter, (vi) entropy of local standard deviation filtered image, (vii) actual diameter, and (viii) entropy. Among these high-rank features, we evidently found that textural features are also as significant as the geometric features for identifying the overlapping nuclei. Using the selected features, we train an RF ensemble classifier for classification between single and overlapping nuclei. The trained classifier is validated using the testing dataset. The performance of the proposed algorithm for the detection of overlapping nuclei is assessed on a testing dataset using six performance measures. These measures are sensitivity, specificity, precision, F1 score, accuracy, and geometric mean (G-mean) and are formulated in Equations (3.1) through (3.4). It is worthy to note that sensitivity is also referred to recall.

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \times 100\% \tag{3.1}$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \times 100\% \tag{3.2}$$

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative} \times 100\% \tag{3.3}$$

$$G\ Mean = \left(\sqrt{Sensitivity \times Specificity}\right) \times 100\% \tag{3.4}$$

- *TruePositive* represents the number of correctly detected overlapped nuclei.
- *TrueNegative* denotes the number of correctly detected single nuclei.
- *FalsePositive* represents the number of wrongly detected single nuclei as overlapping nuclei.
- *FalseNegative* represents the number of missing overlapped nuclei by the proposed method.

In order to judge the proposed method graphically, receiver operating characteristics (ROC) is plotted as sensitivity against (1-specificity). From the ROC curve, Area Under ROC (AUROC) is further computed.

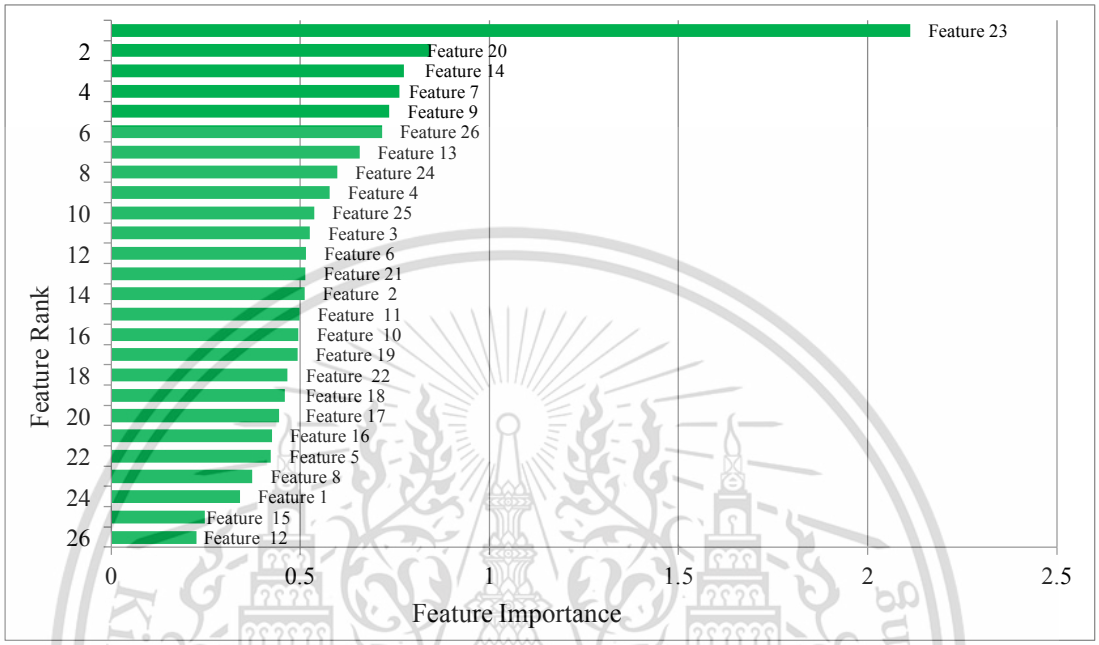


Figure 3.8 Ranking the relative importance of features using RF ensemble feature selection

The classification accuracy of using RF selected features was compared to the accuracy of using all features. Furthermore, we also investigated four alternative classifiers: Naïve Bayes (NB) [69], Support Vector Machine (SVM) [70], K Nearest Neighborhood (KNN) [71], and Decision Tree [72] by pairing with all features and RF's selected features. The classification accuracies of using all features and RF selected features blending with five classifiers are described in Tables 3.4 and 3.5, respectively. The experimental results exhibit that RF selected features are superior to using all features for most classifiers except NB. The results also reveal that the RF ensemble classifier gives preferable accuracy compared to NB, SVM, KNN, and DT classifiers. The synergy between RF selected features and the RF ensemble classifier has given the highest classification accuracy. A ROC curve is further plotted in Figure 3.10 to

evaluate the classifiers graphically. From the curves, it can be evidently seen that RF classifier provides higher accuracy and stability than others. From the ROC curves, we also computed the AUROC of the investigated classifiers. The comparison of AUROC is presented in Figure 3.11. A double-strategy RF (combination of RF ensemble feature selector and RF ensemble classifier) achieved the highest AUC by a given 99.09%.

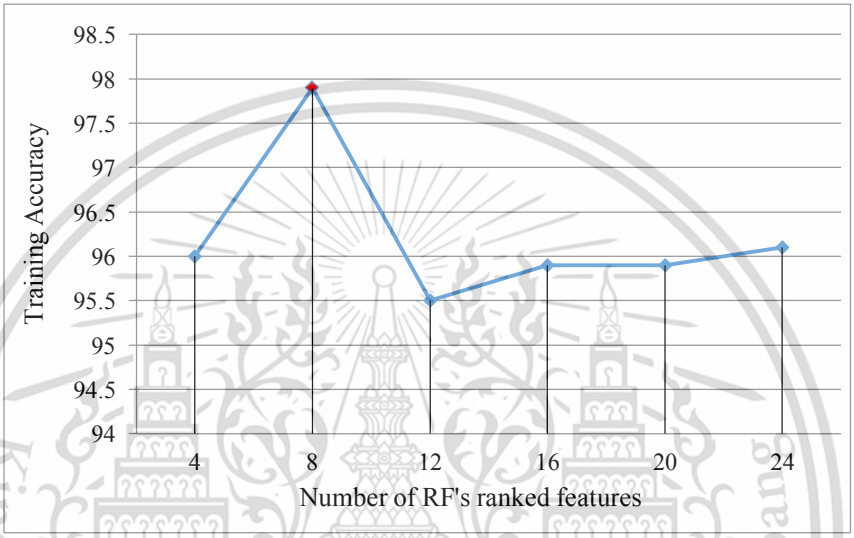


Figure 3.9 Different number of features and their training accuracies using RF

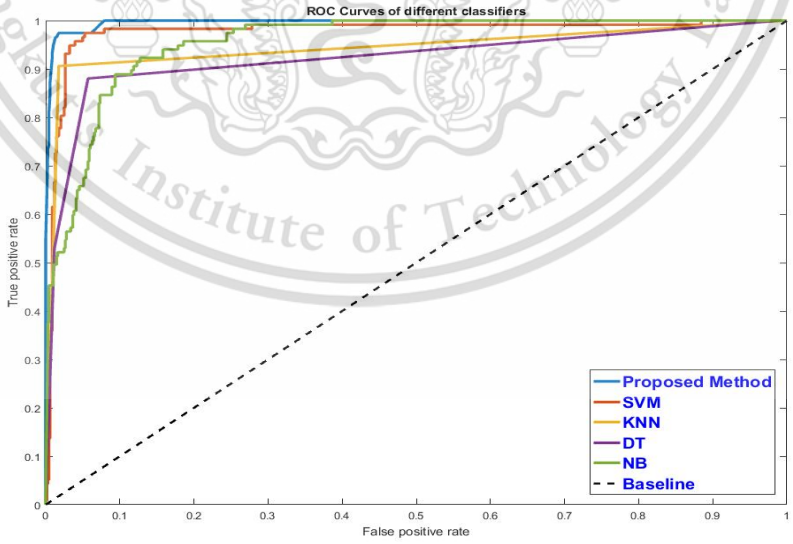


Figure 3.10 ROC curves of different classifiers using RF selected features (proposed method is a double-strategy RF)

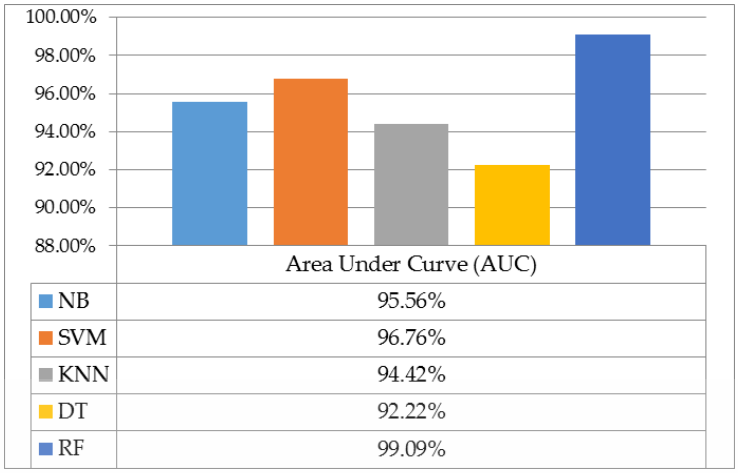


Figure 3.11 AUC of different classifiers using RF selected features

Table 3.4 The classification results of different classifiers using all features

Classifiers	Performance Metrics					
	Sensitivity	Specificity	Precision	F1 Score	Accuracy	G Mean
NB	62.07%	98.68%	88.89%	73.10%	93.38%	78.26%
SVM	78.45%	97.51%	84.26%	81.25%	94.75%	87.46%
KNN	79.31%	97.66%	85.19%	82.14%	95.00%	88.01%
DT	66.67%	97.07%	79.59%	72.56%	92.63%	80.45%
RF	84.48%	97.51%	85.22%	84.85%	95.63%	90.77%

Table 3.5 The classification results of different classifiers using RF's selected features

Classifiers	Performance Measures					
	Sensitivity	Specificity	Precision	F1 Score	Accuracy	G Mean
NB	52.14%	97.51%	78.21%	62.56%	90.88%	71.30%
SVM	93.16%	97.22%	85.16%	88.98%	96.63%	95.17%
KNN	90.60%	98.24%	89.83%	90.21%	97.13%	94.34%
DT	65.52%	98.68%	89.41%	75.62%	93.88%	80.41%
RF	96.58%	98.68%	92.62%	94.56%	98.38%	97.63%

For the comparison, existing studies are tested on the different kind of microscopic images and, in addition, the common dataset is not available. To make a fair and objective comparison, the existing methods are tested on our dataset. It should be noted that all the methods in the comparison were evaluated with the same experiment settings and the same dataset used to test the proposed method. Therefore, the evaluation results were compared fairly without affecting by any other factors. The comparison of existing methods and the proposed method is given in Table 3.6. From the quantitative comparison, it is deduced that the proposed method yields superior accuracy compared to existing studies [52, 56]. It is also reasonable to conclude that the new combination of geometric and textural features is more discriminant than the features used in existing methods for differentiating between single and overlapping nuclei. Figure 3.11 shows the samples of detected overlapping nuclei by the proposed method.

Table 3.6 The performance comparison of the proposed method and existing methods using CPE images dataset

Methodology	**Experimental data**	**Features/Classifiers**	**Classification performance**
H. Wang et al. [52]	4000 nuclei from CPE images	5 size and shape features SVM	F1 score 84.12% Accuracy 95.38% G mean 90.31%
M. Guven et al. [56]	4000 nuclei from CPE images	3 shape and 2 local minima based features Fuzzy c-Means clustering	F1 score 62.15% Accuracy 88.13% G mean 78.23%
Proposed algorithm	4000 nuclei from CPE images	4 shape and 4 textural features Double-strategy RF	F1 score 94.56% Accuracy 98.38% G mean 97.63%

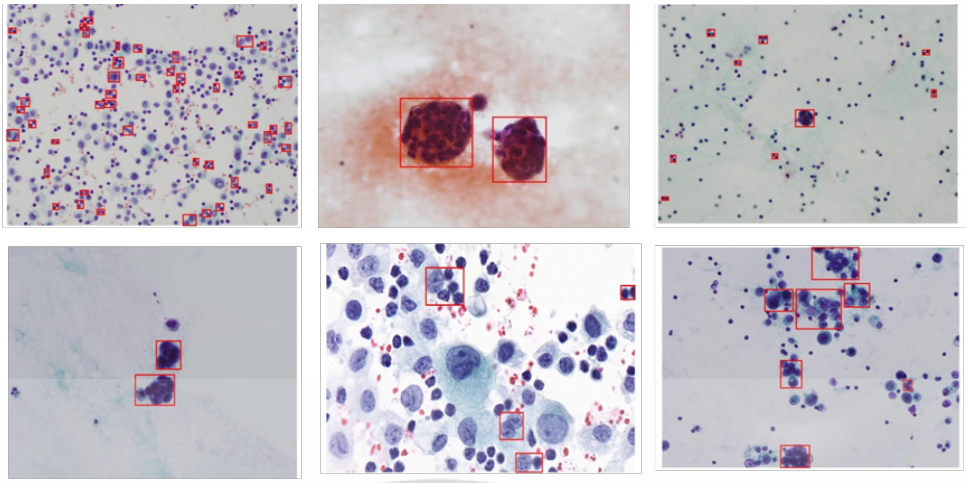


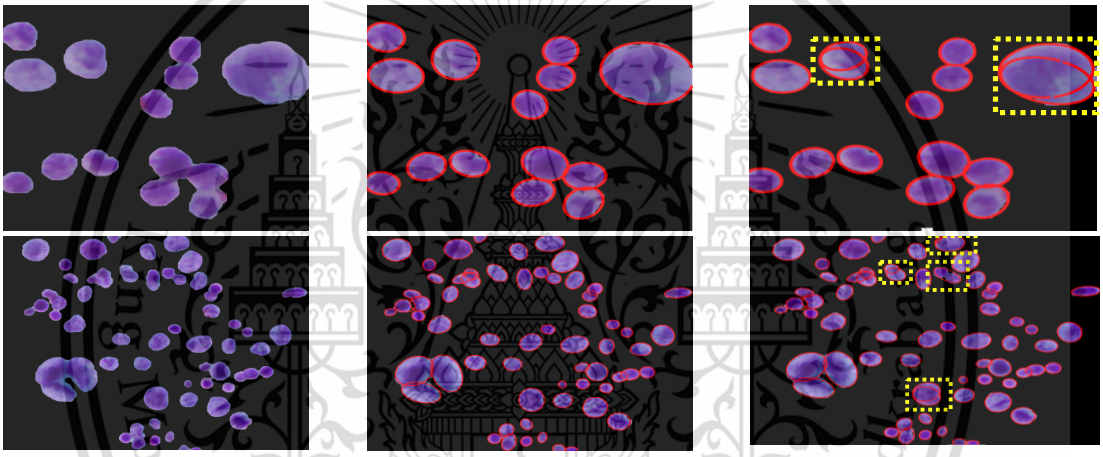
Figure 3.12. The sample images of detected overlapping nuclei by the proposed algorithm

The proposed method is able to accurately detect and differentiate adjacent, overlapping or aggregating nuclei from individual nuclei. Since it is simple and provides high accuracy, it can serve as a new supportive tool in developing new overlapping cell separation algorithms. Moreover, our method has the potential to integrate with existing overlapping-separation methods, such as watershed methods, contour concavity analysis, rule-based methods, etc., to separate overlapping nuclei.

3.6.1 Decomposition of Overlapping Nuclei into Its Constituents

When the overlapped nuclei are detected by our proposed method, they are decomposed into its constituents using the contour concavity analysis (CCA). CCA is one of the most famous and efficient splitting algorithm, which is based on the curvature of the objects. By detecting the overlapping nuclei before applying the splitting method can not only prevent from over- and under-splitting but also shorten the computation time, as tabulated in Table 3.7. Figure 3.13 shows the samples of isolated overlapping cell nuclei. Figure 3. 13 (a) is the segmented nuclei image. Figure 3. 13 (b) represents the results images of proposed splitting methods

(i.e., the combination of our proposed overlapping nuclei detection algorithm and contour concavity analysis) and Figure 3. 13 (c) depicts the result images of classical contour concavity analysis. As shown in Figure 3. 13 (b), employing the splitting method only on the identified overlapped region can prevent the single cell nuclei from over-splitting and overlapped cell nuclei from under-splitting. This happened because the splitting method is focused solely on the overlapped area. The yellow shading box in Figure 3. 13 (c) is illustrated to highlight the over- and under- splitting using the classical concavity analysis based splitting method.



(a) (b) (c)

Figure 3.13 Comparison results of overlapped nuclei splitting methods (a) segmented nuclei (input), (b) proposed splitting method based on the combination of shape analysis and concavity analysis, and (c) contour concavity analysis. (note that the yellow rectangular box indicates the over- and under- splitting)

TABLE 3.7 Comparison of computation complexity of overlapped nuclei splitting methods

Splitting methods	Average processing time
Concavity analysis	10.2 seconds
Integration of detection of overlapped nuclei and concavity analysis	6.8 seconds

3.7 Summary

This chapter presented the detection of overlapping nuclei and decomposition them into their constituents. We introduced the new overlapping nuclei detection algorithm based on a new combination of shape and textural features and double-strategy RF. The proposed algorithm provided an average accuracy of 98.38% for the identification of overlapped nuclei. Then, the proposed algorithm integrates with existing contour concavity analysis to split the overlapping nuclei. Therefore, at first, we detected the overlapped nuclei through the proposed method, and the detected ones are decomposed into their constituents using CCA. The main distinct from the existing methods that we judge whether the split is necessary, if positive, splitting procedure is applied. It can be concluded that accurately detecting overlapping nuclei before decomposing them into their constituent parts can help to reduce the workload of separation methods because these methods need to work only on detected overlapping nuclei instead of on all nuclei.

CHAPTER 4

CLASSIFICATION OF NORMAL AND CANCER CELLS

4.1 Introduction

After the cell nuclei' contours are precisely delineated, it is to establish the classification scheme which differentiates between normal and cancer cells. Classification of normal and cancer cells is the ultimate goal of cytology image analysis in this thesis. From each segmented nucleus, the certain features are extracted to analyze the morphology and visual changes of the cell. Then, feature analysis is carried out to learn the characteristics of the cancer cells in CPE images.

4.2 Related Works

Some researchers have proposed the automated analysis of CPE images to diagnose the cancer cells [22-23,73-74]. In [23] proposed the automated analysis of CPE images to detect the adenocarcinoma cells. Morphological and wavelet features are used as input to BP neural network to differentiate between healthy and adenocarcinoma cells. [73] reported a clinical decision support system (CDSS) for the detection of cancer cells in CPE images. The system is designed using Evolutionary Programming / Evolutionary Strategies (EP/ES), evaluated on the dataset with 928 nuclei and obtained accuracy of 93.4 %. In [74] presented another CDSS to diagnose the malignant cells from CPE images. CellProfiler software is used to segment the cells and measure their features. A total of 393 features are extracted from 500 cells. Then, genetic algorithm (GA) is applied to select the important features. The different number of features such as 10 features, 30 features, 50 features, and 39 features are selected and the discriminant ability was compared. The highest accuracy of 91% is obtained by

using 39 features as input to linear SVM classifier. Despite the success of above-stated methods, there still exist some serious issues that must be resolved. Those methods did not take consider the image quality improvement even though CPE images may contain a great deal of noises and uneven background or illumination. It is a pity that [23, 73] did not address precisely the cells segmentation procedure. Accurate delineation of cells is the most important task of cell analysis and often affects the final decision of the diagnosis results. In [74] used the ready-made software to segment out the cells and measure the features. With these regards, it is reasonable to conclude that their system is not fully automatic and independent. The lack of reliable cell segmentation procedures had limited those methods to be redesigned and/or reproduced for practical use. In addition, those existing methods are missed to address the overlapping cell problem. Precisely segmenting the boundary of the cell nuclei would provide better accuracy of cell analysis. Only [73] and [74] had evaluated the effectiveness of their methodology in the cancer cells classification. Although convincing achievement has been demonstrated in those studies, some good results have been hampered by a small testing dataset. It can be noted; the aforementioned shortcomings do require careful experimental design. Obviously, there is still room for further enhancements in automated cells analysis of CPE images.

4.3 Proposed CAD System for Classification of Normal and Cancer Cells

The block diagram of the complete CAD system is depicted in Figure 4.1. It mainly comprises of preprocessing, nuclei segmentation, post-processing, overlapped nuclei isolation, feature extraction, feature selection, and classification stages. As we have presented from processing to overlapped nuclei segmentation in the previous

chapters, we will present the feature extraction, feature selection, and classification in this chapter.

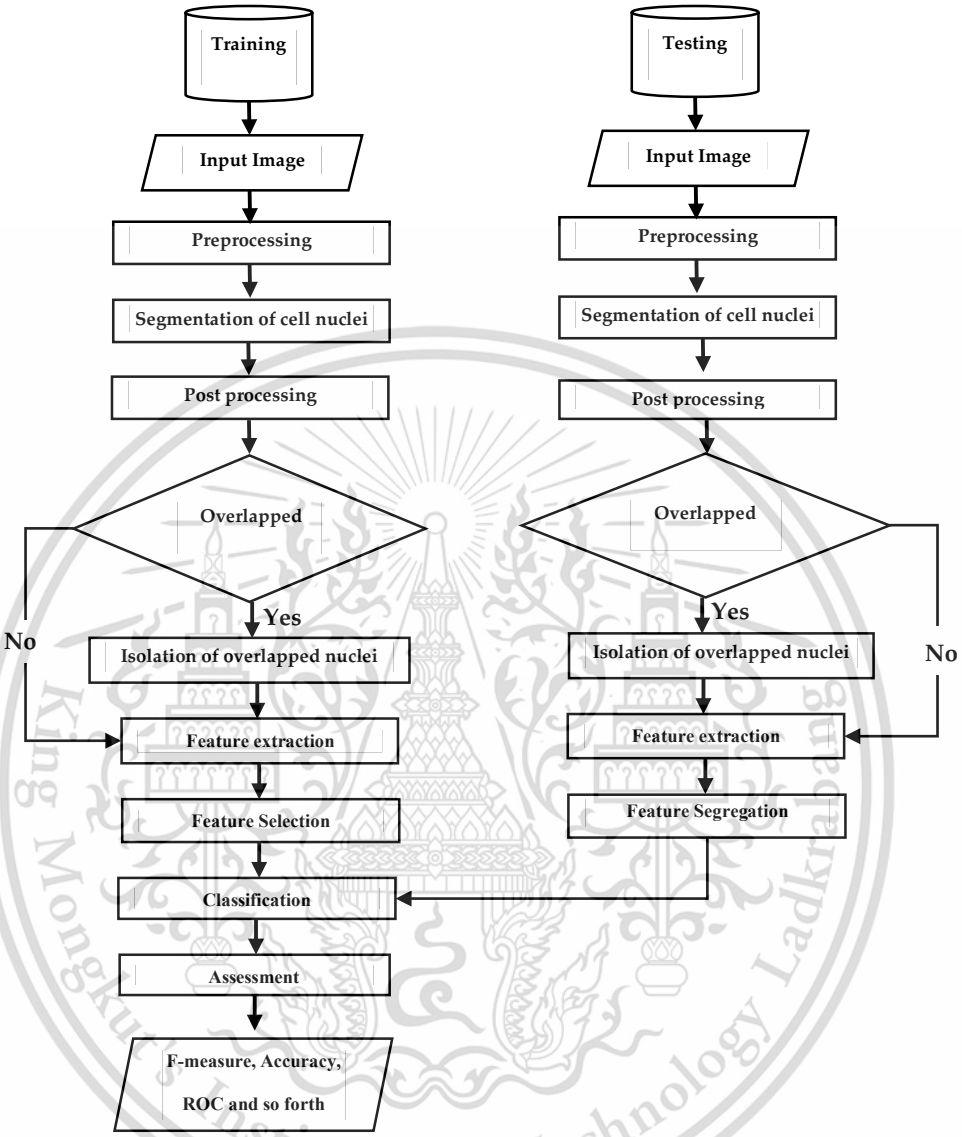


Figure 4.1 System framework of the proposed CAD system

4.4 Feature Extraction

After the cell nuclei are accurately delineated, feature extraction is established to extract the features that reflect the observation of cytologists. In the literature of cytology and histology image analysis, the dominant features for the diagnosis of malignancy used by the cytologists are related to morphometric, colorimetric and textural features [75-79]. Similar to other cytological images, CPE images are also rich

in various features like color, shape, and texture. In this study, 201 features related to the morphometric, colorimetric and textural, are extracted and combined to obtain a robust, information-rich and discerning feature set.

4.4.1 Morphometric Features

There are certain differences in morphology between benign and cancer cell nuclei in CPE images. For instance, excessive growth of cell nuclei size, and a significant variation of cell nuclei size in the image are suggested as malignancy. Moreover, the irregular shape of cell nuclei such as the unsmooth margin of nuclei occurs in malignant cases. Thus, in this study, 14 morphometric features are extracted to evaluate the nucleus size and shape irregularity. The description of these features is given in Table 4.1 and coded as F1-F14.

Table.4.1 List of morphometric features and their associated equations

Code	Feature Name	Equation
F1	Area	$\sum_{i=1}^{n}\sum_{j=1}^{m} S(i,j)$
F2	Perimeter	$Even\ count + \sqrt{2}(odd\ count)$
F3	Roundness, circularity	$\frac{4\pi * Area}{Perimeter^2}$
F4	Solidity	$\frac{Area}{ConvexArea}$
F5	Equivalent circular diameter	$\sqrt{\frac{4 \times Area}{\pi}}$
F6	Compactness	$\frac{Area}{Perimeter^2}$
F7	Eccentricity	$2 * (\sqrt{(\frac{ma}{2})^2 - (\frac{mi}{2})^2})/\ ma$
F8	Diameter	$\frac{Perimeter}{2\,\pi}$
F9	Major axis length (ma)	$\sqrt{(x_1 - x_2)^2 - (y_1 - y_2)^2}$
F10	Minor axis length (mi)	$\sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2}$
F11	Elongation	ma/perimeter
F12	MaxIntensity	max $(pixelValues)$
F13	MinIntensity	min$(pixelValues)$
F14	MeanIntensity	mean$(pixelValues)$

Where $S(i,j)$ is the segmented image of rows i and columns j, ma and mi are major axis and minor axis of the nucleus respectively. x_1, y_1 and x_2, y_2 are the end points of the manor axis and minor axis.

4.4.2 Colorimetric Features

The usage of colorimetric features has tremendously increased in computer vision tasks due to their discriminative ability of different types of objects. The color provides useful information to determine the malignancy. According to the cytological study, if the particular nuclei are affected by the disease, the nucleus region changes in color. For instance, malignant cell nuclei become darker in color. In order to capture color features, mean of R, G, B, H, S and V components are extracted independently from RGB and HSV model. These features are coded as in the range of F15 to F20.

4.4.3 Textural Features

In cytological pleural effusion images, malignant and cancer cell nuclei heavily differ in the distribution of color and chromatin. For instance, the frequent appearance of the distinct mass in the nucleus may be suggested as malignancy. To exploit the color and chromatin distribution, texture features have been widely adopted in the literature. In this study, three statistical textural descriptors: First Order Statistics (FOS), Gray level occurrence matrix (GLCM) and Gray level run length matrix (GLRLM) are employed to extract the textural features.

4.4.3 (a) Color Component based First Order Statistics (CCFOS)

FOS describes the distribution of pixel intensities within the nucleus region [58]. In the literature, the combination of color and FOS features have achieved better accuracy compared to the conventional FOS features [80-81]. Thus, seven FOS features for seven color components (namely gray, R, G, B, H, S, and V from RGB and HSV

model) are extracted for each nucleus. The extracted features are named as color component based FOS (CCFOS) and encoded from F21 to F69. The reason for extracting seven color components is to obtain FOS textures from the view of different color components. Different color components describe the different defined texture as given in Figure 4.2. The detail of these extracted features is given in Table 4.2 and coded from F21-F69.

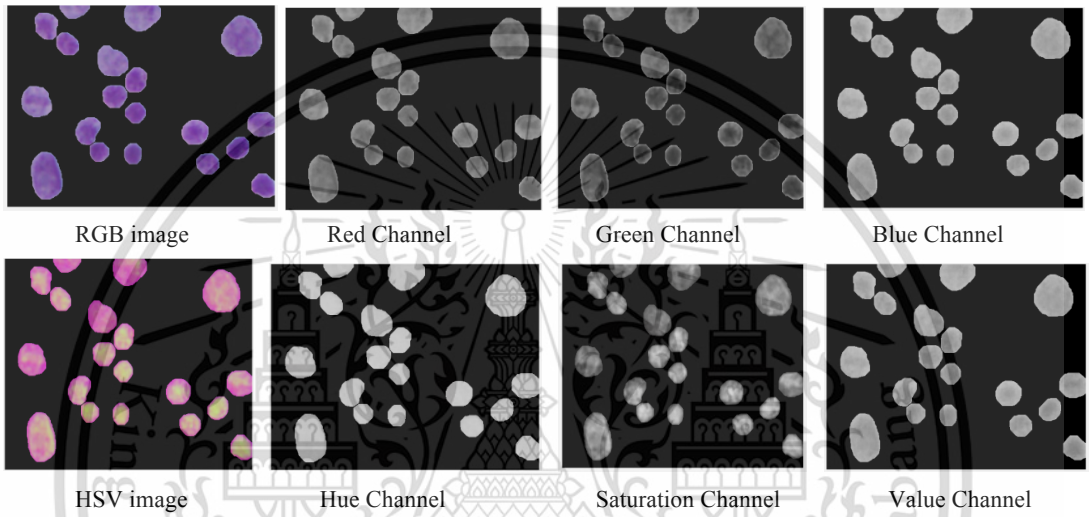


Figure 4.2 Individual color component of RGB and HSV color model in the segmented cell nuclei of CPE image.

Table 4.2 List of CCFOS features and their associated equations

Feature Name	Equation
Mean (μ)	$\sum_{i=0}^{L-1} i\, p(i)$
Standard deviation (σ)	$\sum_{i=0}^{L-1}(i-\mu)^2 \cdot p(i)$
Smoothness	$1-(\frac{1}{1}+\sigma^2)$
Variance	$\sum_{i=0}^{L-1}(i-\mu)^2 p(i)$
Skewness	$\sigma^{-3}\sum_{i=0}^{L-1}(i-\mu)^3 p(i)$
Kurtosis	$\sigma^{-4}\sum_{i=0}^{L-1}(i-\mu)^4 p(i)-3$
Energy	$\sum_{i=0}^{L-1} p\,(i)^2$

Where $p(i)$ is the number of pixels with gray level i, and L represents the number of gray-level bins set for p.

4.4.3 (b) GLCM and GLRLM

FOS captures the features only on individual pixels. It ignores the spatial relationship between the neighborhood pixels. In order to capture texture features that take into account the spatial relationship between neighboring pixels, GLCM [82,83] and GLRLM [84] based higher order statistics features are considered. GLCM represents the distribution of co-occurring intensities ($i\ and\ j$) in a nucleus at a given specific distance (d) and orientation (θ). GLCM can be formulated as Eq. (4.1). When extracting GLCM features, it is required to define three parameters: distance (d) and orientations (θ) that determine the offset and angle between adjacent pixels and the number of gray levels (NG) in the image. In this study, d and NG are set to 1 and 8 respectively. θ is adopted for four orientations $0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$ in order to take into account the rotation of the image. Thus, 22 GLCM features for four different orientations are extracted.

$$G(i,j) = \| \{[(x_1,y_1),(x_2,y_2)]\} \mid x_2 - x_1 = d\cos\theta$$

$$y_2 - y_1 = d\sin\theta, I(x_1,y_1) = i, \sin\theta, I(x_2,y_3) = j \ \| \tag{4.1}$$

Where (x_1,y_1) and (x_2,y_2) are pixels in the segmented nuclei, $I(\cdot)$ is the gray-level of the pixels, and $\|\cdot\|$ is the number of pixels pairs that satisfy the conditions.

GLRLM represents the length of homogeneous runs for each gray level in a definite direction. Similar to GLCM, GLRLM is constructed at four orientations and 8 gray levels. 11 GLRLM features in four different orientations ($0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$) are extracted. Table 4.3 and Table 4.4 describe the lists of GLCM and GLRLM feature and their associated equations. Finally, the feature vector is generated by combining 14 features form morphology, 6 features from color and 181 textural features from CCFOS, GLCM and GLRLM. The list of extracted features is given in Table 4.5. The

class of each nucleus is labeled into either positive or negative class under the guidance of cytologists.

Table 4.3. List of GLCM features and their associated equations

Features	Equations
Autocorrelation	$\sum_i \sum_j (i \cdot j) p(i,j)$
Contrast	$\sum_i \sum_j \lvert i-j \rvert^2 p(i,j)$
Correlation I	$\sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x \sigma_y}$
Correlation II	$\sum_i \sum_j \frac{(i \cdot j)p(i,j)-\mu_x \mu_y}{\sigma_x \sigma_y}$
Cluster Prominence	$\sum_i \sum_j \left(i+j-\mu_x-\mu_y\right)^4 p(i,j)$
Cluster Shade	$\sum_i \sum_j \left(i+j-\mu_x-\mu_y\right)^3 p(i,j)$
Dissimilarity	$\sum_i \sum_j \lvert i-j \rvert \cdot p(i,j)$
Energy	$\sum_i \sum_j p(i,j)^2$
Entropy	$-\sum_i \sum_j p(i,j) \cdot \log\left(p(i,j)\right)$
Homogeneity I	$\sum_i \sum_j \frac{p(i,j)}{1+\lvert i-j \rvert}$
Homogeneity II	$\sum_i \sum_j \frac{p(i,j)}{1+\lvert i-j \rvert^2}$
Maximum Probability	$max_{i,j} p(i,j)$
Sum of square	$\sum_i \sum_j (i-v)^2 p(i,j)$
Sum average	$\sum_{i=2}^{2L} i \cdot p_{x+y}(i)$
Sum energy	$-\sum^{2L} p_{x+y}(i) \cdot \log\left(p_{x+y}(i)\right)$
Sum variance	$\sum_{i=2}^{2L} (i - Sum\ engery)^2 \cdot p_{x+y}(i)$
Difference variance	$\sum_{i=0}^{L-1} i^2 \cdot p_{x-y}(i)$
Difference entropy	$-\sum_{i=0}^{L-1} p_{x-y}(i) \cdot \log\left(p_{x-y}(i)\right)$
Information measure of correlation I	$\frac{HXY-HXY1}{\max(HX,HY)}$
Information measure of correlation II	$(1-\exp[-2(HXY2-HXY)\})^{1/2}$
Inverse Difference Normalized	$\sum_i \sum_j \frac{p(i,j)}{1+\lvert i-j \rvert^2/L}$
Inverse difference moment normalized	$\sum_i \sum_i \frac{p(i,j)}{1+(i-j)^2/L}$

Where $p(i,j)$ is the $(i,j)^{th}$ entry of the co-occurrence propablity matrix, L represents gray level quantization, μ_x, μ_y and σ_x, σ_y are the mean and standard deviation of the p.

Table.4.4 List of GLRLM features and theirs associated equations

Features	Equations
Short run emphasis (SRE)	$\frac{1}{n^r}\sum_{i=1}^{G}\sum_{j=1}^{R}\frac{g(i,j)}{j^2}$
Long run emphasis (LRE)	$\frac{1}{n^r}\sum_{i=1}^{G}\sum_{j=1}^{R} g(i,j) * j^2$
Low gray-level run emphasis (LGRE)	$\frac{1}{n^r}\sum_{i=1}^{G}\sum_{j=1}^{R}\frac{g(i,j)}{i^2}$
High gray-level run emphasis (HGRE)	$\frac{1}{n^r}\sum_{i=1}^{G}\sum_{j=1}^{R} g(i,j) * i^2$
Short run low gray-level emphasis (SRLGE)	$\frac{1}{n^r}\sum_{i=1}^{G}\sum_{j=1}^{R}\frac{g(i,j)}{i^2 * j^2}$
Short run high gray-level emphasis (SRHGE)	$\frac{1}{n^r}\sum_{i=1}^{G}\sum_{j=1}^{R}\frac{g(i,j)*i^2}{j^2}$
Long run Low gray-level emphasis (LRLGE)	$\frac{1}{n^r}\sum_{i=1}^{G}\sum_{j=1}^{R}\frac{g(i,j)*j^2}{i^2}$
Long run high gray-level emphasis (LRHGE)	$\frac{1}{n^r}\sum_{i=1}^{G}\sum_{j=1}^{R} g(i,j) * i^2 * j^2$
Gray level nonuniformity (GNU)	$\frac{1}{n^r}\sum_{i=1}^{G}\left[\sum_{j=1}^{R} g(i,j)\right]^2$
Run length nonuniformity (RNU)	$\frac{1}{n^r}\sum_{j=1}^{MG}\left[\sum_{i=1}^{R} g(i,j)\right]^2$
Run percentage (RP)	n_r/n_p

Where $g(i,j)$ denotes the number of runs of pixels of gray level i and run length j,G is the number of gray levels in the image, R is the number of different run lengths in the image, n_r is the total number of runs, and n_p is the number of pixels in the image.

Table 4.5 List of various features extracted from each nucleus

Name of Feature sets	Number of Features	Ranges
Morphometric Features	14	F1-F14
Colorimetric Features	6	F15-F20
CCFOS (Textural Features)	49	F21-F69
GLCM (Textural Features)	88	F70-F157
GLRLM (Textural Features)	44	F158-201
Combined Feature Set	201	F1-F201

4.5 Feature Selection

The initial feature set contains 201 features related to morphometry, colorimetry, and texture. Directly utilizing all candidate features for classification may cause redundancy and irrelevancy. Redundancy can lengthen the computation time. In turn, irrelevancy may cause poor predictive accuracy. To handle these problems, feature selection is performed in advance of classification. Feature selection is often applied in computer vision when many features get extracted. It improves the prediction performance and generalization capability and provides a faster and more cost-effective model. Feature selection is generally listed into two techniques: filter and wrapper [85]. In filter techniques, the features are chosen depending on their relevance ability with respect to the target. Filter methods are computationally fast and easy to implement. However, there is a possibility that the chosen features might contain redundant information since the selection process is carried out on the statistical measure of each feature. Unlike the filter approach, the wrapper approach depends on the learning methods. It utilizes the estimated accuracy of the learning method as the performance measure to evaluate the usefulness of a feature. As the extension of the wrapper approach, the hybrid approach that combines the metaheuristics methods and supervised learning methods as the integral components of feature selection has been widely utilized in medical image analysis [86-88]. Experiments have found that hybrid methods are more efficient in finding optimal solutions compared to the filter and wrapper methods. The main benefit of the hybrid methods is avoiding being stuck in the local optima. In this study, the novel hybrid feature selection method based on hybridizing simulated annealing, one of the metaheuristics methods, with artificial neural network, one of the popular machine learning methods, is developed to select the most relevant and informative features. The proposed method is known as a hybrid

simulated annealing coupling artificial neural network (SA-ANN) feature selection. The details of SA-ANN are given in the sub-session below.

4.5.1 Hybrid SA-ANN Feature Selection

Simulated Annealing is a global optimization algorithm that is inspired by the natural annealing process in the metallurgy. It models the annealing process of heating material and then gradually cooling by lowering the temperature at the control rate, thus minimizing the system energy [89]. It is typically used to search for the global minimum in a high-dimensional data space. The main advantage of SA is to allow up-hill moves in the iteration to avoid being stuck in a local minimum. SA has been widely used as supervised or unsupervised feature subset selection in data mining techniques, especially in microarray gene classification in biomedical data analysis [90-92]. Inspired by those works, in this study, we develop a novel hybrid feature selection method by hybridizing SA with Artificial Neural Network (ANN). ANN is a machine learning algorithm that mimics the structure of the biological brain. During feature selections via hybrid SA-ANN, the cost value of SA based search space is computed depending on the number of samples correctly predicted by ANN. Firstly, the random initial feature subsets are initially created. These subsets are assessed using a 3-layers ANN trained by Levenberg-Marquard (LM) backpropagation algorithm [93]. The features with the minimal cost are initialized as the best feature set. At each iteration of SA, the neighboring subset is randomly generated by implementing the neighborhood function. Then, similar to the first stage, 3-layers ANN trained by LM backpropagation algorithm is used to evaluate the cost of the neighboring subset. If the neighboring subset has a lower cost than the initial subset, then change the initial subset to neighboring subset. Else, if the neighboring subset has a higher cost, then the individual will move to that subset only if the acceptance probability condition is fulfilled.

Otherwise, the individual remains in the initial subset. By accepting individuals that increase the cost, the algorithm avoids getting stuck by a local minimum in early iterations and explores globally for better solutions. As the algorithm progresses, the temperature is reduced and causing the individual to converge towards the subset with the minimum cost and hence the optimal point. SA-ANN feature selection is summarized in Algorithm 15.

Algorithm 15: Hybrid SA-ANN based feature selection

Input: $Features\_set, MaxIt, Temp, alpha$

Output: S_{best}

$Initial_{subset} \leftarrow CreateInitialSolution(Features\_set)$

$Cost(S_{current}), S_{current} \leftarrow \boldsymbol{CostFunction\_ANN}(Initial_{subset})$

$S_{best} \leftarrow S_{current}$

$\boldsymbol{For}\ (i = 1 :\ MaxIt)$

$New_{subset} \leftarrow CreateNeighbourSolution(S_{current})$

$Cost(S_i), S_i \leftarrow \boldsymbol{CostFunction\_ANN}(New_{subset})$

$\boldsymbol{if}\ (Cost(S_i) \leq Cost(S_{current})$

$S_{current} \leftarrow S_i$

$\boldsymbol{else}$

$DELTA = (Cost(S_i) - Cost(S_{current})) / Cost(S_{current})$

$P = exp(-DELTA/Temp)$

$if\ rand <= P$

$S_{current} \leftarrow S_i$

End

$\boldsymbol{End}$

$if\ (Cost(S_{current}) \leq Cost(S_{best})$

$S_{best} \leftarrow S_{current}$

End

$Temp = Temp * alpha$

$\boldsymbol{End}$

$Return(\ S_{best})$

In Algorithm 15, feature_set, MaxIt, Temp, and alpha are the candidate features, maximum numbers of iteration, initial temperature, and the temperature reduction rate, respectively. S_best is the output that represents the corresponding optimal feature set. The selected features in the optimal feature set are utilized for training and testing the classifier. The code implementation of proposed hybrid SA-ANN feature selection is based on Matlab implementation available in [94] and modified for the necessity. To determine the efficiency of proposed feature selection method, we also examined alternative feature selection approaches: all features approach (i.e. without feature selection), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA).

4.6 Classification

The selected features are utilized as input to the classifier to differentiate the cells between benign and malignant. In cytology and histology image analysis, classification models revolve around Support Vector Machine (SVM), Naïve Bayes (NB), Artificial Neural Network (ANN), K Nearest Neighborhood (KNN), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Decision Tree (DT) and Ensemble Classifier (EC). The selection of a classification model for medical image analysis depends on the type and size of the dataset to be classified. Our dataset of cell nuclei is large and highly unbalanced wherein the class of cancer nuclei is limited while the class of benign nuclei is abundant. Ensemble classification has yielded preferable results for classification of skewed data [95, 96]. Thus, to deal with the unbalanced data distribution, we adopted an ensemble classifier that employs bootstrap aggregation (bagging) decision trees and termed as EBC [97, 98]. The core idea of using EBC is to develop multiple bootstrap data-samples and to build multiple base classifiers for each bootstrapped sample. In this study, one hundred decision trees are used as the base

classifiers. The final prediction of EBC is obtained through major voting. The block diagram of EBC classifier is depicted in Figure 4.3.

Algorithm 16: Ensemble Bagging Classifier

Input: training data S, Weak Learner I, Number of Iteration T, Bootstrap size n

Output: Bagged classifier $H(x) = sign\ (\sum_{t=1}^{T}(h_t(x))where\ \ h_t\ \epsilon\ [-1,1]$ are the induced classifiers

Training:

For $t = 1$ to T do

$S_t \leftarrow$generate bootstrap sample from S

$h_t \leftarrow I(S_t)$

End

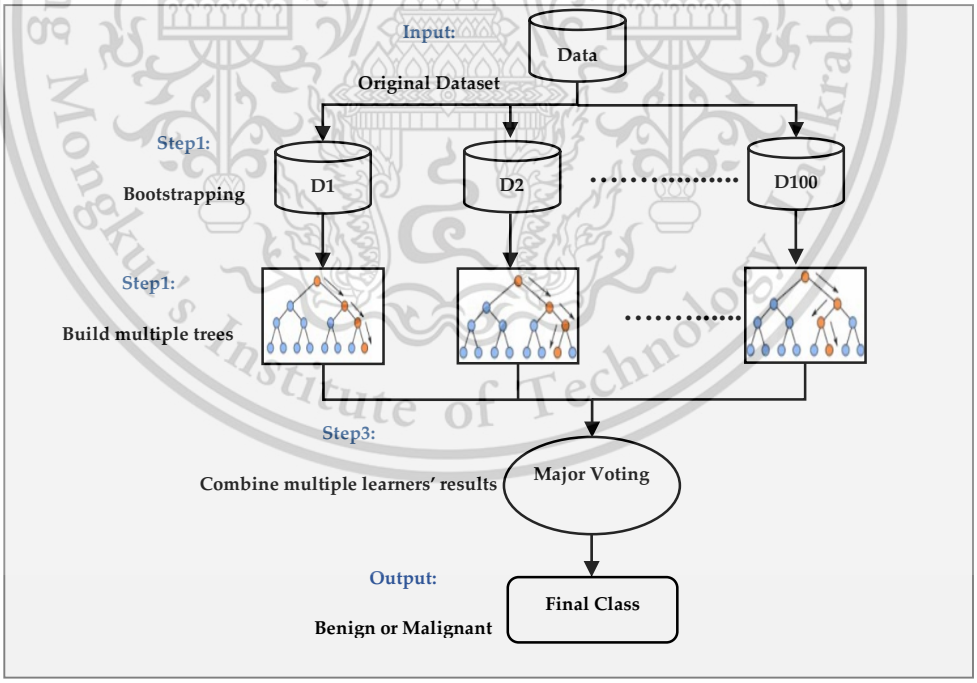


Figure 4.3 Block diagram of ensemble bagging classifier used in this study

4.7 Experimental Results and Discussions

The study was based on 125 CPE images containing around 10500 cells. The studied dataset was randomly partitioned into training and testing sets in an 80-20% ratio. 80% of the images were allocated to the training dataset to train the classifier and 20% to the testing dataset to validate the trained classifier. Training and testing datasets were disjointed (i.e., the same image was not assigned to represent both training and testing datasets). It is noteworthy that all the experiments carried out in this chapter are based on the same experimental setting and environment.

Once the nuclei are accurately delineated (see detail in Chapter 2 and 3), 201 features that represent the morphometric, colorimetric, and textural features are extracted from each nucleus. In order to avoid redundancy and irrelevancy, hybrid SA-ANN feature selection is developed to choose the most discerning and informative features. The promising features that correctly map to the target are identified by supervised ANN and used in the annealing process. The SA-ANN algorithm was iterated 50 times with initial temperature (temp=10) and temperature reduction rate (alpha=0.99). The algorithm was adapted to select the different desired number of features (nf) such as 15, 20, 25, 30, 35 and 40. Based on the experimental results obtained, it is inferred that selecting more than 20 features resulted in slightly decreasing the classification accuracy. Thus, SA-ANN algorithm is fixed to select 20 features out of 201 features. It means that we manually selected the optimal number of features based on the training accuracy. It would be more effective if the optimal number of features are automatically selected. Thus, multi-objective optimization of SA to select both the important features and the optimal number of features are suggested as future work.

The list of 20 selected features and their correlation matrix are described in Table 4.6 and Figure 4.4, respectively. By analyzing the selected features, it reveals that they include one or more representative features from each group of features given in Section III.E. Among 20 selected features, 16 features are textural features. Thus, it is reasonable to conclude that the textural features supply more diagnostic information than other features. Moreover, the correlation matrix demonstrates that the proposed hybrid SA-ANN feature selection selected the most significant features with less redundant information. The selected features are used as input to the classification model to predict the malignancy. Choosing a classification model depends on size and type of the data to be predicted. Our data is highly skewed, wherein the cell nuclei belong to the malignant (positive) is limited, and the cell nuclei belong to benign (negative) are abundant. Thus, we adopted ensemble classification which provides preferable results to the classification of unbalanced data. As mentioned in Section III. F, the dataset is firstly bootstrapped randomly, and 100 decision trees are used as the base classifiers to classify the bagged datasets. The final classification result is obtained through major voting. To evaluate the classification performance, we compared the ground truth and classification results with respect to four performance metrics: sensitivity, specificity, f-measure, accuracy. These four performance measures are formulated in equation 3-8.

Table 4.6 Description of selected features through hybrid SA-ANN

No.	Feature Code	Feature Name	Feature Set
1	F37	Smoothness of B component	CCFOS
2.	F163	Short run high gray-level emphasis	GLRLM0
3.	F 51	Smoothness of S component	CCFOS
4.	F 82	Sum of square	GLCM0
5.	F 96	Cluster Prominence	GLCM45
6.	F 55	Energy of S component	CCFOS
7.	F 146	Homogeneity II	GLCM 135
8.	F 19	Mean color of S component	Colorimetric
9.	F 25	Skewness of R component	CCFOS
10.	F 187	Long run high gray-level emphasis	GLRLM 90
11.	F 88	Information Measure of Correlation	GLCM0
12.	F 132	Difference Entropy	GLCM 90
13.	F 2	Perimeter	Morphometric
14.	F 12	MaxIntensity	Morphometric
15.	F 183	High gray-level run emphasis	GLRLM 90
16.	F4	Solidity	Morphometric
17.	F 70	Autocorrelation	GLCM 0
18.	F 28	Mean from G component	CCFOS
19.	F 168	Run percentage	GLRLM0
20.	F 128	Sum Entropy	GLCM0

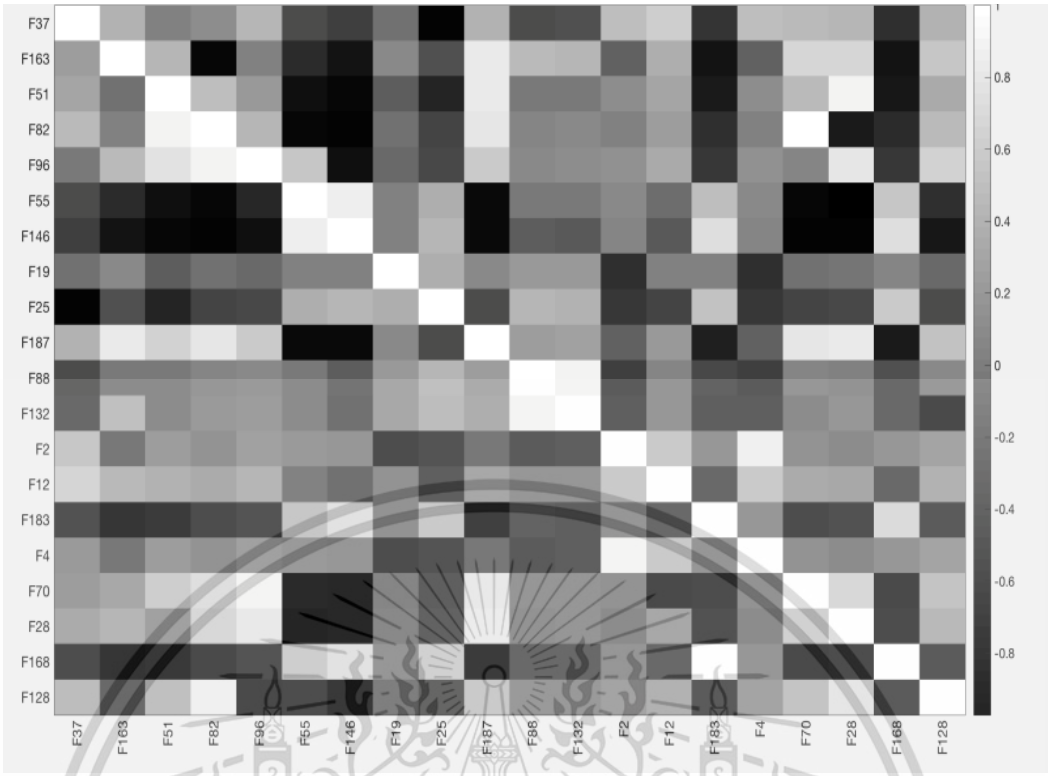


Figure 4.4 Correlation matrix for the selected features using hybrid SA-ANN feature selection. Noted that (Correlation =1 (white) means the highest correlation, -0 (black) no correlation)

To make a fair objective comparison, a common public dataset is required. By far, we are not aware of any common publicly available dataset. Also, the diagnosis schemes of CPE images in the literature are different from the proposed diagnosis scheme. Thus, we build our own experimental setup wherein the impact of using different feature selection methods and different classification models on the classification performance is observed. In the first three experimental scenarios, we compared the classification accuracy achieved with and without feature using the proposed classifier (i.e., EBC). In the first scenario, we compared the results between our proposed SA-ANN approach and all features approach (i.e. without feature selection). Secondly, the result of SA-ANN approach is compared with the results of SA approach. In the third scenario, we established the comparison between SA-ANN

approach and other robust hybrid feature selection methods: PSO-ANN and GA-ANN approach. Furthermore, in the fourth experimental scenario, we employed seven alternative classifiers namely SVM [70,99], ANN [100], NB [69], KNN [71,101], LR [102], LDA [103], and DT [72,104] classifiers and coupled with the feature selection approaches. The result achieved by the proposed synergy between SA-ANN feature selection and EBC is compared with the results obtained through various pairs. Therefore, for each feature selection approach, the experimental results are presented with respect to four performance measures and eight classification models (including EBC). The results from four experimental scenarios are shown in Table 4.7. We clarify that hybrid SA-ANN coupling with ECBDT classifier (shaded in green background) is our proposed method. As reported in Table 11, utilizing the feature selection methods (i.e. SA-ANN, SA, PSO-ANN, GA-ANN or SA) provided better accuracy compared to the all features approaches (i.e. without feature selection) for all classifiers. The results also demonstrate that, with the exception of coupling with SVM, KNN, and LR classifiers, the proposed SA-ANN selection marginally improves the accuracy compared to the SA based approach and yields better accuracy compared to PSO-ANN and GA-ANN approach when coupling with ANN, NB, LD, DT, and proposed EBC classifiers. When coupling with SVM classifier, PSO-ANN approach yields better compared to other selection approaches. Similarly, GA-ANN approach yields better accuracy compared to other feature selection methods, while coupling with KNN classifier. Likewise, SA approach yields better accuracy compared to other feature selection methods, while coupling with LR. The superior feature selection method for each classifier is highlighted in light purple color. It is observed that different classifiers perform differently for different selected features. However, regardless of the feature selection methods utilized, EBC (ensemble classifier) consistently provided better

accuracy compared to other single classifiers. From the experimental results, it is inferred that synergy of hybrid SA-ANN coupling with EBC classifier outperformed other pairs of feature selection approaches and classification models described above in terms of classifying the cells in the CPE images. To get clear comparison results, we further plot the comparison of accuracy and F-measure as illustrated in Figure 4.4 and Figure 4.6, respectively. Moreover, a Receiver Operating Characteristics (ROC) curve for different classifiers coupling with SA-ANN feature selection is depicted in Figure 4.7. It shows that the ROC curve of the proposed method is on the left upper corner and has higher stability of classification rate compared to other comparative methods in the study. Table 4.8 is given for the comparison of performance between the proposed algorithms and previous studies for cancer cell diagnosis in CPE image. The visual results of detected malignant nuclei (both correct and failure cases) are depicted in Figure 4.8. Figure 4.8 (a) is the image with annotated malignant cell nuclei labeled by two experts in which blue and green color represent two experts. Figure 4.8 (b) describes the diagnostic results of the proposed CAD system wherein the red bounding box represent the detected malignant cells. Even though the proposed method yields the promising result, there are still some failures especially when the malignant characteristics of the cell occur in the cytoplasm. Therefore, it remained as future work to detect the malignancy based on the combined analysis of cell nuclei and cytoplasm.

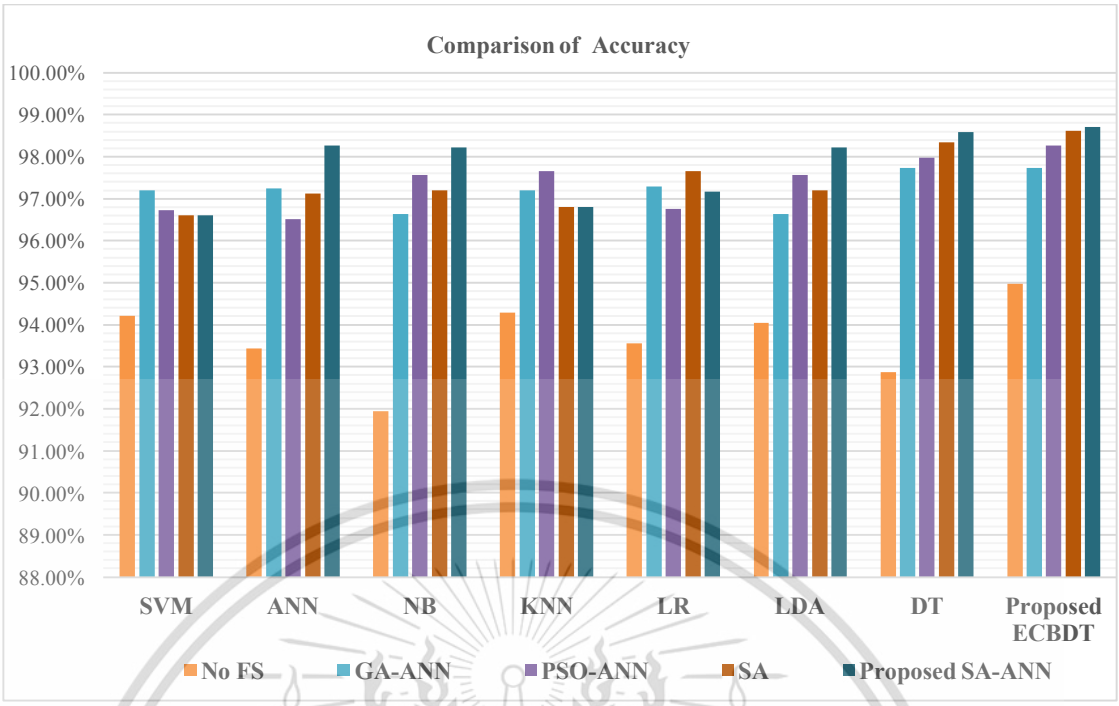


Figure 4.5 Comparison of accuracy using different pairs of feature selection methods and classifiers

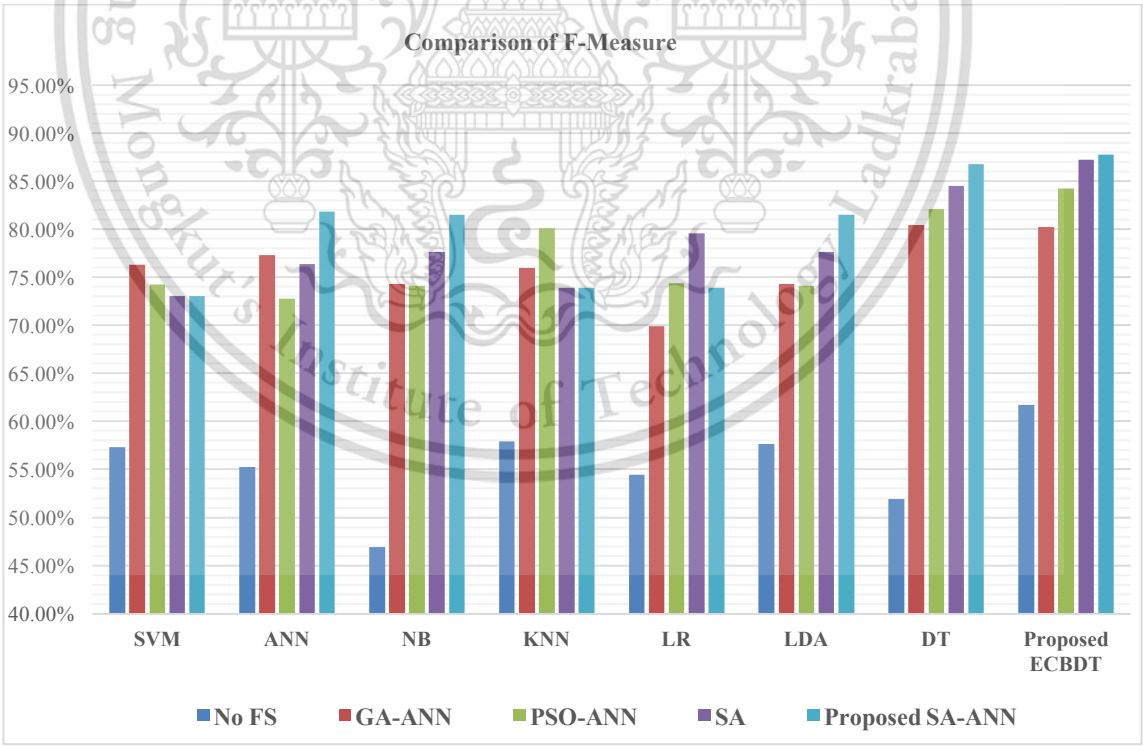


Figure 4.6 Comparison of F-measure using different pairs of feature selection

Table 4.7. Comparison of classification performance achieved by different synergies between feature selection methods and classification models

Classifiers	Performance measures	Feature Selection Approaches				
		All features	PSO-ANN	GA-ANN	SA	Proposed SA-ANN
SVM	Sensitivity	72.18%	73.65%	87.97%	85.71%	85.71%
	Specificity	95.47%	96.64%	97.22%	97.22%	97.22%
	F Measure	57.31%	76.29%	74.29%	73.08%	73.08%
	Accuracy	94.21%	97.21%	96.72%	96.60%	96.60%
ANN	Sensitivity	75.19%	70.91%	86.47%	86.47%	72.93%
	Specificity	94.48%	96.11%	97.09%	97.73%	99.70%
	F Measure	55.25%	77.33%	72.78%	76.41%	81.86%
	Accuracy	93.44%	97.25%	96.52%	97.13%	98.26%
NB	Sensitivity	66.17%	69.16%	64.66%	90.23%	72.93%
	Specificity	93.41%	95.67%	99.44%	97.60%	99.66%
	F Measure	46.93%	74.30%	74.14%	77.67%	81.51%
	Accuracy	91.95%	96.64%	97.57%	97.21%	98.22%
KNN	Sensitivity	72.93%	74.29%	87.97%	84.21%	84.21%
	Specificity	95.51%	96.72%	98.20%	97.52%	97.52%
	F Measure	57.91%	75.96%	80.14%	73.93%	73.93%
	Accuracy	94.29%	97.21%	97.65%	96.80%	96.80%
LR	Sensitivity	71.43%	69.23%	87.22%	84.96%	79.70%
	Specificity	94.82%	96.11%	97.31%	98.37%	98.16%
	F Measure	54.44%	69.96%	74.36%	79.58%	75.18%
	Accuracy	93.57%	97.29%	96.76%	97.65%	97.17%
LDA	Sensitivity	75.19%	69.16%	64.66%	90.23%	72.93%
	Specificity	95.12%	95.67%	99.44%	97.60%	99.66%
	F Measure	57.64%	74.30%	74.14%	77.67%	81.51%
	Accuracy	94.05%	96.64%	97.57%	97.21%	98.22%
DT	Sensitivity	71.43%	71.83%	86.47%	84.21%	86.47%
	Specificity	94.10%	96.32%	98.63%	99.14%	99.27%
	F Measure	51.91%	80.42%	82.14%	84.53%	86.79%
	Accuracy	92.88%	97.73%	97.98%	98.34%	98.58%
Proposed EBC	Sensitivity	74.48%	76.47%	86.47%	87.22%	87.97%
	Specificity	96.11%	97.09%	98.93%	99.27%	99.40%
	F Measure	61.73%	80.28%	84.25%	87.22%	87.79%
	Accuracy	94.98%	97.73%	98.26%	98.62%	98.70%

Table 4.8 Quantitative comparison between the proposed study and the previous studies

Authors	Data	Nuclei extraction	Features/Classifiers	Accuracy (%)
Ref [23]	60 cancers and many normal cells	Not clearly described	• Morphology and wavelet features • BP NN	N/A
Ref [73]	928 nuclei	Not clearly described	• Not clearly described the features • EP/ES	93.4 %
Ref [74]	500 nuclei	Cell profiler software	• 398 features (not described what they are) • Linear SVM	91 %
Proposed study	10500 cells from 125 CPE images	• Hybrid SLIC-K Means • CCA based overlapped nuclei isolation	• 201 morphometric, colorimetric and textural features • hybrid simulated annealing feature selection • ensemble bagging classifier	98.70 %

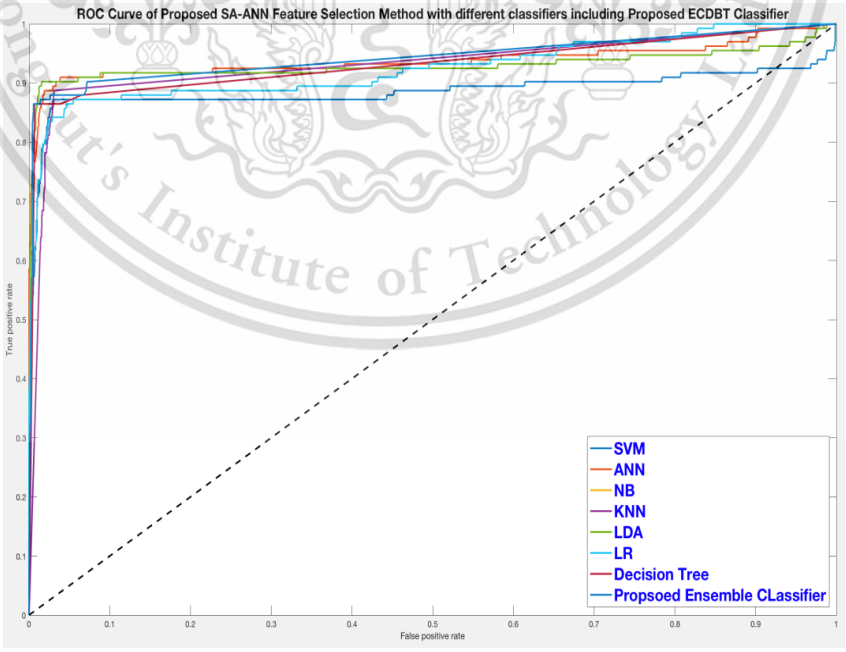
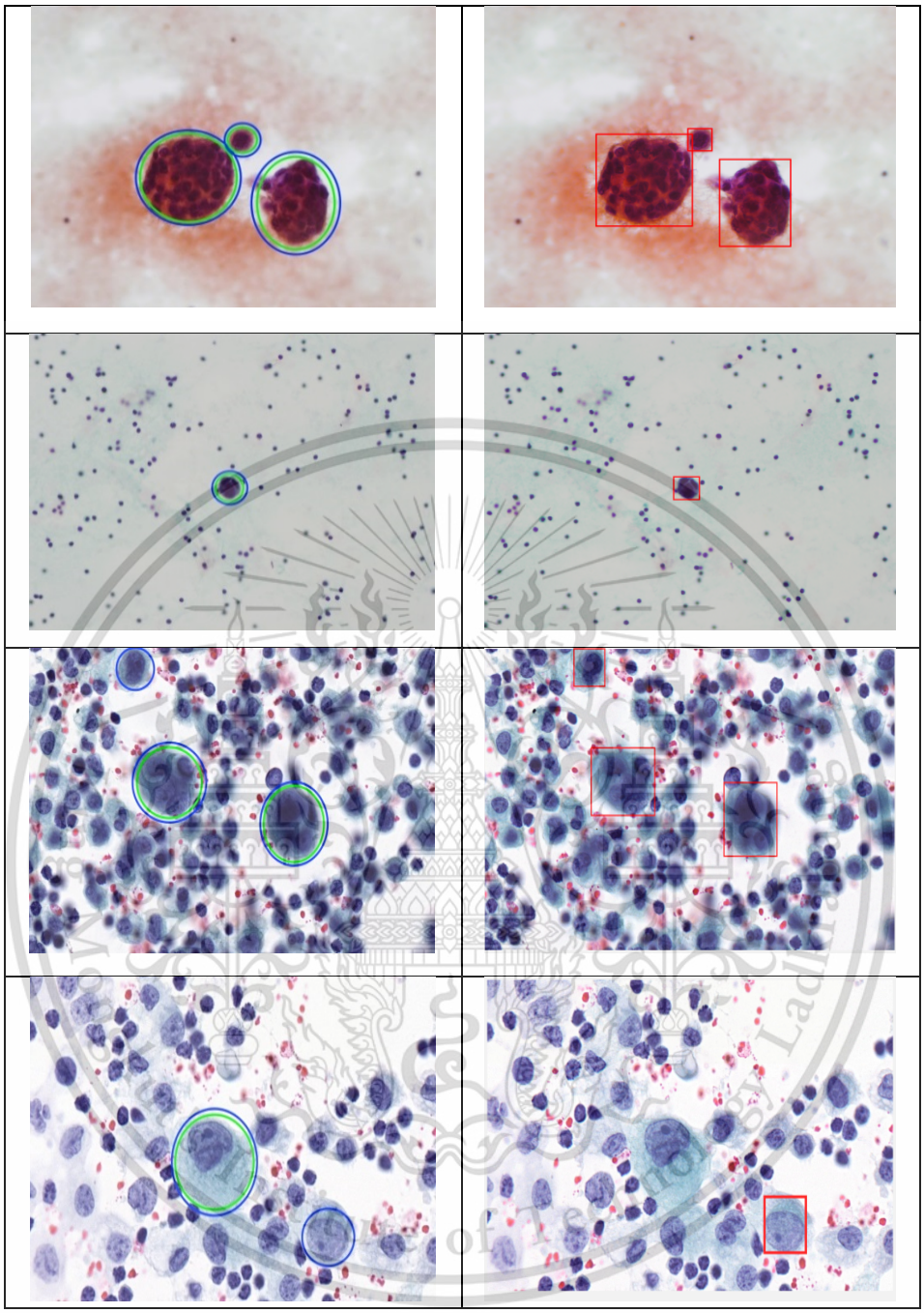


Figure 4.7 ROC Curve for SA-ANN feature selection blending with eight classifiers



(a) (b)

Figure 4.8 Visual demonstration of diagnostic results using the proposed CAD system to detect the malignant cells in CPE images (a) the original image with ground truth malignant cells annotated by two experts (blue and green circles represent two experts), and (b) detected malignant cells by the proposed CAD system

4.8 Summary

This chapter presented the classification between normal and cancer cells from CPE images using optimization and machine learning methods. The novelty of the proposed method is the development of SA-ANN based hybrid feature selection method to select the most discriminant features. Ensemble bagging classifier is used as an ensemble classifier which is able to handle the data-unbalanced problem. To validate the performance of the novel SA-ANN selection method, the other two famous feature selection algorithms, hybrid PSO and hybrid GA are also investigated. To judge the efficiency of the ensemble classifier, we also employed SVM, ANN, NB, LR, LDA, KNN and DT as the classifiers, and paired with feature selection algorithms. The different pairs between feature selection algorithms and classifiers are empirically evaluated with six performance metrics. The experimental results exhibit that the proposed feature selection method outperformed all features approach, SA, hybrid PSO, and hybrid GA. Ensemble bagging classifier provided the preferable accuracy compared to other conducted classifiers. The synergy between novel hybrid SA-ANN feature selection and an ensemble bagging classifier are superior to other pairs in this study by given the accuracy of 98.58%.

CHAPTER 5

MULTI- CLASSIFICATION OF LUNG CACNER SUBTYPES IN PLERUAL EFFUSION USING DEEP LEARNING

5.1 Introduction

Among various cancer causes in MPE, lung cancer is the main cause and has the shortest survival. Around 70% of lung cancer patients end up developing MPE. Due to the global increase of lung cancer mortality and morbidity, the number of malignant pleural effusion caused by lung cancer will be increased. Early diagnosis and classification of lung cancer subtypes can allow the pathologists to prescribe the optimal treatment plan. It is often difficult to precisely differentiate adenocarcinoma and squamous cell carcinoma in terms of their morphological characteristics. There are many varieties of morphologies among these cancer cells. Computer-aided diagnosis (CAD) can potentially be a useful tool for differentiating those types. Among the four major types of carcinoma, large cell carcinoma is the easiest to detect because of its severe atypism. We, therefore, concentrate on the classification of the other three types—adenocarcinoma, squamous cell carcinoma and small cell carcinoma —which are sometimes confused with each other in the cytological specimen.

5.2 Related Works

Most recently, Deep Learning (DL), a state-of-the-art machine learning tool, has been placed in the spotlight in the field of computer vision and, subsequently, in medical image analysis [105]. This is because DL algorithms address the primary limitations of the conventional shallow-structured machine learning tools, such as SVM, KNN, NB, and RF. Deep Belief Net (DBN) [106], Stacked Auto-Encoder (SAE) [107], and Deep Convolutional Neural Network (DCNN) [108] are current deep learning models.

Among these models, DCNNs are widely used in image classification task and characterized by multi-layered interconnected channels, with a high capacity for learning the features and classifiers from data spontaneously given their deep architecture, their capacity to adjust parameters jointly and to classify simultaneously.

DCNNs are well known to give better performance than conventional image classification techniques [105,108]. For example, Alex Krizhevsky et al. [108] won the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) using a deep convolutional neural network (DCNN) to classify high- resolution images. At last, at the ILSVRC 2015, Residual Neural Network (ResNet) by Kaiming He et al. [109] obtained a top-5 error rate of 3.57% which beats human-level performance on ImageNet dataset. DeepFace, the facial recognition system introduced by Facebook, surpassed all face recognition benchmarks in the literature [110]. In addition, many research groups have investigated the application of DCNNs to medical images [111–114]. The most famous medial image analysis using DL is the dermatologist level classification of skin cancer. The deep learning based algorithms can classify skin cancer or benign as accurately as dermatologists. The algorithm can be turned into a mobile app to be used at home or clinic.

Various CAD algorithms have been proposed for the classification of microscopic images using deep learning techniques. For example, Wang et al. combined handcrafted features and deep convolutional neural networks for mitosis detection [115]. Zhang et al. employed convolutional neural networks for classification of cervical cancer in pap smear images [116]. H. Sharma et al. proposed a classification system of gastric carcinoma in whole slide histology images using deep convolutional neural networks [117]. Ertosun et.al proposed an automated system for grading gliomas

using deep learning [118]. E. Arvaniti et al. proposed an automated Gleason grading system for prostate cancer on tissue microarrays dataset using deep convolutional neural networks [119]. To the best of our knowledge, DCNNs have not been applied to cytological images for classification of lung cancer subtypes in pleural effusion. In this study, we developed an automated classification scheme of lung cancer subtypes in CPE images using DCNNs.

5.3 Deep Convolutional Neural Networks for Classification of Lung Cancer Subtypes

DCNNs are constructed by multi-layer interconnected neural networks, wherein powerful low-, intermediate-, and high-level features are hierarchically extracted. A typical CNN framework has two main layers—the convolutional and pooling layers—that, together, are called the convolutional base of the network. Tremendous progress has been made in medical image analysis with DCNNSs, thanks to the availability of large-scale annotated dataset. With the ability of learning highly hierarchical image feature extractors, DCNNs are also expected to solve the multi-categorical lung cancer subtypes classification problems. However, the limited labeled CPE images become a handicap to train a DCNN. To solve this problem, rather than designing and training a convolutional neural network (CNN) from scratch, we employed transfer learning using pre-trained deep CNN models.

Facing the problem of collecting enough training data to rebuild models, transfer learning aims to transfer knowledge from a large dataset known as source domain to a smaller dataset named target domain. Either the feature spaces between domain data are different or the source tasks and the target tasks focus on different topics, boosting the performance of the target task. Transfer learning using CNNs is commonly used in different fields [120, 121]. The workflow of transfer learning for

multi-categorical classification of lung cancer subtypes is depicted in Figure 5.1. During performing transfer learning, we firstly load our training dataset and the pre-trained model which will be adapted to our classification task. Using the training data, we retrained the pre-trained models to learn the specific features from our images and obtained a newly customized DCNNs model. The newly trained DCNNs will be used to predict the new data and its performance is assessed using various performance measures.

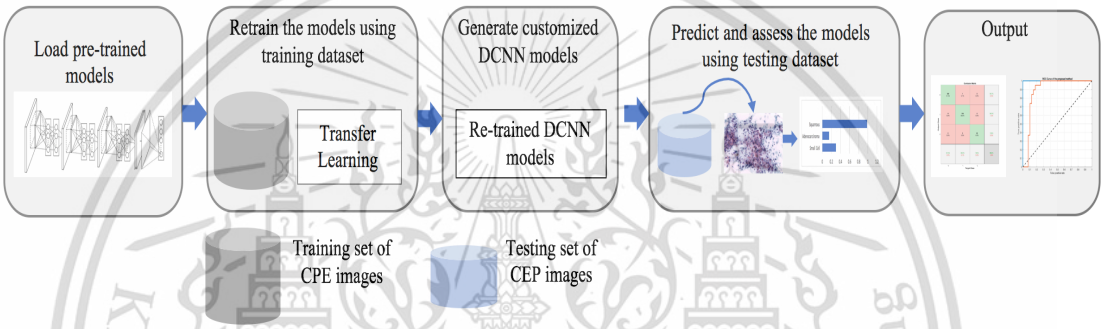


Figure 5.1 Workflow of transfer learning for multi-categorical lung cancer classification on CPE images

Among various pre-trained models, we configured a transfer learning with ResNet50 model and retrained it to be adapted to our task. It is characterized by a very deep network with 50 layers. The ResNet models introduced a deep residual module via the additive transformations to address the gradient vanishing and performance degradation problem of training the deeper networks. In conventional neural networks, each layer feeds into the next layer. In a residual network, each layer feeds into the next layer and directly into the layers about 2–3 hops away. The residual module utilizes a direct path between the input and output and allows fitting each stacked layer to a residual mapping rather than directly fitting to a desired underlying mapping. We denote the desired underlying mapping as $H(x)$ and define the residual function using

$F(x) = H(x) - x$, which can be rearranged into $H(x) = F(x) + x$. Kaiming He et. al claimed that the optimization is much easier on the residual mapping rather than on the original, unreferenced map. Figure 5.2 depicts $F(x) + x$ in feedforward neural networks with "shortcut connections". $F(x)$ and x represents each stacked layer and the identity function(input=output) respectively.

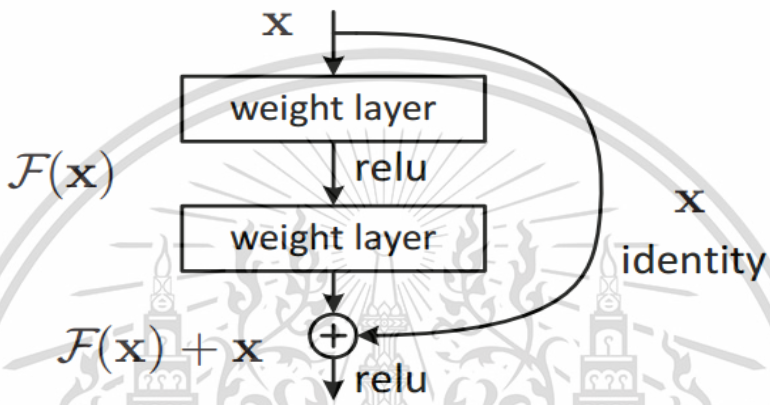


Figure 5.2 The residual module (image credit [120])

ResNet is the winner of ILSVRC-2015 image classification competition, which beats the human-level perform in the large-scale dataset. It is trained on a subset of ImageNet containing a million of images, which can classify a thousand of classes There are many versions of ResNet such as ResNet18, ResNet34, Resnet50, and ResNet101 and ResNet152. We employed ResNet50 in our study because ResNet18 and 34 are quite small networks, and in contrast, ResNet101 and152 are very deep network and takes too much computation time. During transfer learning with ResNet50, we deleted the last soft-max layer along with its weights and retrained it using our dataset to obtain the customized soft-max layer [109,122,123]. The customized soft-max layer will generate the probabilities of three lung cancer subtypes namely adenocarcinoma, squamous cell carcinoma, and small cell carcinoma. The schematic diagram of ResNet50 model which was used in our study is shown in Figure 5.3.

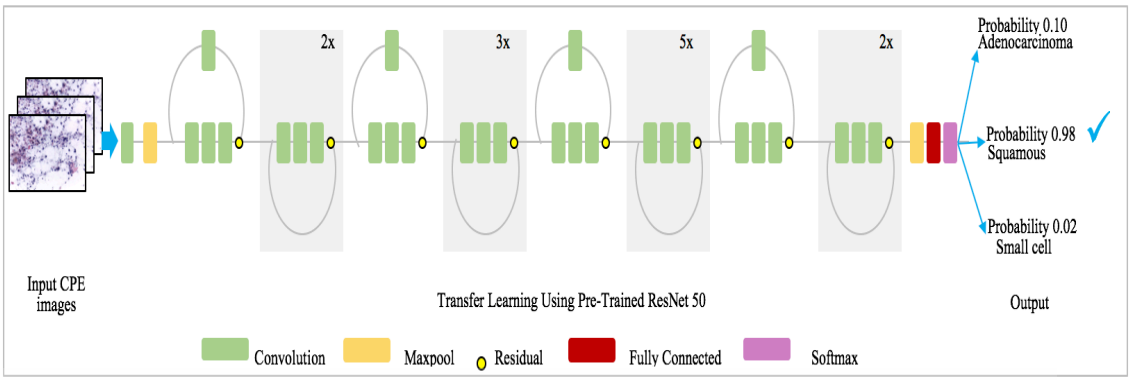


Figure 5.3 Schematic diagram of ResNet50 (compressed view) for lung cancer subtypes classification

5.4 Image Dataset

The images were obtained using the same image acquisition procedure mentioned in the above sections. For the multi-categorical classification task, we collected more CPE images. The new dataset contains 117 adenocarcinomas, 62 squamous cell carcinomas and 13 small cell carcinoma images of pleural effusion. All the images are manually classified by the cytologists to be used as the ground truth data.

Table 5.1 Dataset description and the number of images for each class

	Adenocarcinoma	Squamous cell carcinoma	Small cell carcinoma	Total
Number of images	117	62	13	192

5.5 Experimental Results and Discussions

The experiments are carried out using the same environmental setting as described in the previous chapters. The self-provided preliminary dataset containing 192 images is used in the analysis. 5-fold cross-validation is used to train ResNet50. We also investigated other famous pre-trained DCNN models namely AlexNet

[108,124,125], GoogLeNet [124,126,127], and InceptionV3 [124,128] to classify lung cancer types in CPE images and the obtained results are compared with the results by ResNet50. In addition, we investigated the impact of image preprocessing on the performance of DCNNs. For this reason, we preprocessed the image using a median filter for denoising and histogram equalization for contrast enhancement. The preprocessed images are used as input to DCNN models. Table 5.2 compares the performance of using original and preprocessed images. It evidently shows that using preprocessing image slightly decreased the accuracy. Even though preprocessing is necessary in image analysis using traditional machine learning, it is not strictly necessary to perform an additional processing step in deep learning because DCNN models can learn how to adapt to variation in the data if there is enough data. For the comparison of DCNN models, ResNet50 provided the average accuracy of 97.93%. The stand deviation is about 1.44%, and it means the performance is very stable and the model is not overfitted. From the experimental results, it also shows that ResNet50 provided comparable accuracy with AlexNet. ResNet50 and AlexNet yielded slightly better accuracy than GoogleNet and the significant improvement compared to Inception V3.

Table 5.2 The performance of deep convolutional neural networks with and without preprocessing

Deep CNNs	Accuracy	
	Without image preprocessing (original images are used as input)	With image preprocessing (preprocessed images are used as input)
AlexNet	97.59± 1.5	97.59± 1.97
GoogleNet	95.17± 2.25	93.11± 3.45
Inception V3	88.96± 2.31	86.89± 0.94
ResNet 50	97.93± 1.44	97.59 ± 1.97

It is worth noting that DCNN models are able to interpret cell morphology and placement of cancer cells solely from images without prior knowledge and experience of biology and pathology. To our best knowledge, there has been no work established to classify lung cancer subtypes from pleural effusion using DCNN models. By given 97.93% overall accuracy, it is a very convincing result because classification of lung cancers in cytology slide is a challenging task for cytologists. Hence, the proposed method can be used as an assistance tool to cytologists in subtyping lung cancer from the pleural effusion.

In the future studies, we will collect more images for squamous and small cell carcinoma, and will test the robustness of the examined methods with the large dataset. In this study, the classification of lung cancer subtypes is performed using the image containing multiple cells. However, it is also possible to perform feature analysis and classification by focusing on individual cells. Therefore, we hope to develop a method to comprehensively classify cells and arrays of cells in the future.

5.6 Summary

This chapter presented the classification of lung cancer subtypes in pleural effusion using CPE images and deep learning approach. Three types of lung cancer that is adenocarcinoma, squamous cell carcinoma and small cell carcinoma are classified using DCNNs. Similar to Chapter 4, six performance metrics and confusion matrix are used to evaluate the classification performance of DCNNs. The experimental results exhibit that approximately 97.93 % of images are correctly classified by using ResNet50 model. These results reveal that DCNN is an effective and useful tool for automated classification of lung cancer subtypes in cytology pleural effusion images.

CHAPTER 6

CONCLUSIONS AND FUTURE WORKS

6.1 Conclusions

The overarching goal of this thesis is to investigate the challenges in robust quantitative image analysis techniques in digital cytology images. Since the last decade, a significant amount of research has been done in the field of cytology, focusing on nuclei detection, segmentation, and classification in different image modalities. Nuclei detection, segmentation, and classification are important steps in cancer diagnosis and subtyping. The presence of nuclei and their components are crucial indicators for evaluating the existence of disease and its severity. In this thesis, we have presented quantitative image analysis techniques based on the nuclei using image processing algorithms, machine learning algorithms and deep learning algorithms, which can serve as building blocks for digital cytology-based cancer diagnosis and subtyping. This chapter summarizes and concludes the work presented in this thesis and discusses some future directions.

In Chapter 1, we have introduced the background of the research and the standard diagnostic procedures for the diagnosis of cancer cells in pleural effusion. Among various alternatives for pleural effusion diagnosis, cytology examination in which the pathologists study and interpret each cell in the cytology slides under the microscope is deemed as the gold standard throughout the world. We described the limitations of the manual cytology examination and the benefits of the computerized analysis systems. We also presented various methods for preparing the glass slides from fluid specimens, and digitalization of those slides into virtual slides for quantitative image analysis. We briefly presented a brief description of image acquisition and

preparation of studied datasets in the collaboration with the cytologists at Department of Pathology, Faculty of Medicine, Srinakharinwirot University, Thailand. We also described the motivation of the research and a brief summary of the main contributions of this work. The motivation behind this thesis was to observe the opportunities offered by the quantitative image analysis techniques for the accurate diagnosis of cancer cells and the classification of malignant tumor types and investigate the challenges in developing those techniques.

In routinely stained cytology images, segmentation of cell nuclei is a challenging computer vision problem, due to uneven staining and illumination during image acquisition, high variability in images. In Chapter 2, we presented a comprehensive review of the existing nuclei segmentation methods in CPE image analysis. The review highlights open research areas which are characterized by unique challenges, with few existing studies. Evaluating and comparing the existing studies is objectively impossible to do solely based on their reported results, as they used different (often unbalanced, inconsistent or even medically irrelevant) datasets, evaluation methods, and performance metrics. To get a fair, objective comparison, we undergone a comparative analysis of twelve image segmentation methods in segmenting the cell nuclei. All those methods are evaluated using the same experimental setting, same dataset, and the same evaluation metrics. By benefiting the knowledge and distinguishable features from comparative analysis, we also introduced the novel hybrid algorithm for segmentation of cell nuclei in CPE images. The algorithm is based on the integration of SLIC superpixels algorithms and K Means clustering method. The proposed algorithm yielded good segmentation performance by given 96% accuracy and took less computation time compared to classical K Means based nuclei

segmentation. It is also found out that the proposed novel algorithm outperforms twelve existing nuclei segmentation methods in CPE image analysis.

In Chapter 3, we presented the new detection framework of overlapping nuclei in the CPE images. The framework is based on the new combination of geometrical and texture features of nuclei to analyze the overlapping nuclei. Double-strategy random forest is used as an ensemble feature selector to select the most significant features and as an ensemble classifier to differentiate between single laying nuclei and overlapped nuclei. Then, overlapped nuclei are separated from the single nuclei and CCA is utilized to decompose the detected overlapped nuclei into its constituents. The integrated framework of detection and isolation of overlapping nuclei have reduced the processing time and over-and under-splitting.

In Chapter 4, we presented a novel computer-aided diagnosis system of cancer cells in CPE images. A total of 204 features are extracted from the delineated nuclei using morphometric, colorimetric and textural features in a composite manner. The most discrimination features are selected using a novel hybrid simulated annealing based feature selection algorithm. The selected features are fed as input to the classifiers. Several machine learnings algorithms were investigated for the accurate classification of cancer cells. Ensemble bagging classifier is found to be superior to other machine learning techniques. The achievements also exhibit that the synergy between a hybrid simulated annealing for feature selection and ensemble bagging classifier for classification is jointly powerful for the classification of cancer cells.

In Chapter 5, we presented the multi-classification of lung cancer subtypes using deep learning approach. Three types of lung cancer namely adenocarcinoma, squamous cell carcinoma, and small cell carcinoma are classified using the deep convolutional neural networks through the transfer learning. Four pre-trained DCNN

models namely AlexNet, GoogLeNet, InceptionV3 and ResNet50, are investigated for the multi-classification of lung cancer subtypes and ResNet50 is found to be superior to other pre-trained networks by given 97.93 % classification accuracy.

6.2 Future Directions

The algorithms proposed in this thesis can potentially be served as the assisted tools to the cytologists in the cytology examination for the diagnosis of cancer cells. In this section, we discuss some possible lines of work for extending and improving the performance of those techniques. The most straightforward future development is the integrated analysis of the cell nuclei and cytoplasm in one single framework. The morphology changes of cytoplasmic regions can help to provide a better diagnostic accuracy of malignant pleural effusion. To do so, the accurate segmentation of cytoplasm from the other surrounding objects should be developed. In the current study, the multi-categorical classification of malignant cells was analyzed using the whole image including thousands of cells. Thus, the comprehensive feature analysis and classification of the individual cell for multi-categorical cancer subtypes have been deferred as the future studies.

Appendix A

RGB to CIE-L*a*b color space conversion [129]:

(a) RGB to CIE-XYZ color conversion:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.412 & 0.357 & 0.180 \\ 0.212 & 0.212 & 0.072 \\ 0.019 & 0.119 & 0.950 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad \text{(A.1)}$$

(b) CIE-XYZ to CIE-L*a*b color conversion

$$L^* = 116\left[\delta\left(\frac{Y}{Y_n}\right)\right] - 16 \quad \text{(A.2)}$$

$$a^* = 500\left[\delta\left(\frac{X}{X_n}\right) - \delta\left(\frac{Y}{Y_n}\right)\right] \quad \text{(A.3)}$$

$$b^* = 200\left[\delta\left(\frac{Y}{Y_n}\right) - \delta\left(\frac{Z}{Z_n}\right)\right] \quad \text{(A.4)}$$

Here X_n, Y_n and Z_n are the CIE-XYZ tristimulus values of the reference white point (where subscript n suggests normalized) which can be obtained by setting $R = G = B$ in RGB to XYZ transformation and $n \in \{\frac{X}{X_n}, \frac{Y}{Y_n}, \frac{Z}{Z_n}\}$ where

$$\delta(t) = \begin{cases} t^{1/3} & if\ t > 0.008856, \\ 7.787t + \frac{4}{29} & otherwise \end{cases} \quad \text{(A.5)}$$

Appendix B

The contour segment grouping is carried out through the process of ellipse fitting. Given a pair of contour segments s_i and s_j, and a function measuring the goodness of ellipse fitting, the segment s_i is grouped to s_j if the goodness of ellipse fitted to the joint segments is higher compared to the goodness of ellipses fitted to each individual contour segments separately.

The goodness of fit is described as average distance deviation (ADD) which measures the discrepancy between the fitted curve and the candidate contour points. The lower value of ADD indicates higher goodness of fit and, therefore the joint rule to perform segment grouping in terms of ADD is defined as:

$$ADD_{s_i \cup s_j} \leq ADD_{s_i} \tag{B.1}$$

$$ADD_{s_i \cup s_j} \leq ADD_{s_j} \tag{B.2}$$

where the definition of ADD is as follows:

Given contour segment s_i consisting of n points, $s_i = \{ p_k (x_k, y_k)\}_{k=1}^{n}$ and corresponding fitted ellipse points, $s_{f,i} = \{p_{f,k} (x_{f,k}, y_{f,k})\}_{k=1}^{n}$, ADD_{s_i} is defined as follow:

$$ADD_{s_i} = \frac{1}{n} \sum_{k=1}^{n} \sqrt{(x_k - x_{f,k})^2 + (y_k - y_{f,k})^2} \tag{B.3}$$

Within the transformed coordinate system, $\begin{bmatrix} x'_k \\ y'_k \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_k - x_{eo} \\ y_k - y_{eo} \end{bmatrix}$, the formula of the ADD can be simplified to

$$ADD_{s_i} = \frac{1}{n} \left[\sum_{k=1}^{n} \sqrt{(x'^2_k + y'^2_k)} \left(1 - \frac{1}{|D_k|}\right) \right] \tag{B.4}$$

Where D_k is given by

$$D_k^2 = \frac{x_k'^2}{a^2} + \frac{y_k'^2}{b^2} \quad \text{(B.5)}$$

and $a, b, (x_{eo}, y_{eo})$ and θ are the ellipse parameters, the semi-major axis length, the semi-minor axis length, the ellipse center point, and the ellipse orientation angle with respect to x axis, respectively. The plain ADD criterion for segment grouping often leads to undesired results if the contour points do not strictly fit to the ellipse model. In order to address this issue, additional rules are needed. (see the detail in [67]).

BIBLIOGRAPHY

(1) M. Hejmadi "Introduction to cancer biology", Holstebro, Denmark: Ventus Publishing, 2010.

(2) V.S. Karkhanis, and J.M. Joshi, Pleural effusion: diagnosis, treatment, and management. *Open access emergency medicine: OAEM*, *4*, p.31, 2012.

(3) R. Myatt, "Diagnosis and management of patients with pleural effusions", *Nursing Standard*, *28*(41), 2014.

(4) T. Sato, "Differential diagnosis of pleural effusions", *JAPAN MEDICAL ASSOCIATION JOURNAL*, *49*(9/10), p.315, 2006.

(5) M.E. Roberts, E. Neville, R.G. Berrisford, G. Antunes, and N.J. Ali, "Management of a malignant pleural effusion: British Thoracic Society pleural disease guideline 2010", *Thorax*, *65*(Suppl 2), pp.ii32-ii40, 2010.

(6) "Malignant Pleural Effusion", American Thoracic Soceity, www.thoracic.org.

(7) M.J. Na, "Diagnostic tools of pleural effusion", *Tuberculosis and respiratory diseases*, *76*(5), pp.199-210, 2014.

(8) N.R. Desai, and H.J. Lee, "Diagnosis and management of malignant pleural effusions: state of the art in 2017", *Journal of thoracic disease*, *9*(Suppl 10), p.S1111, 2017.

(9) "Pleural Effusion", https://en.wikipedia.org/wiki/Pleural_effusion (Assessed on 12th Oct 2018).

(10) K. Alagha, C. Tummino, T. Sofalvi, and P. Chanez, "Iatrogenic eosinophilic pleural effusion", *European Respiratory Review*, *20*(120), pp.118-120, 2011.

(11) Y. Wimalasena, L. Kocierz, D. Strong, J. Watterson, and B. Burns, "Lung ultrasound: a useful tool in the assessment of the dyspnoeic patient in the emergency department. Fact or fiction?", Emerg Med J, 35(4), pp.258-266., 2018.

(12) R. Zablockis, and R., Nargela, "Diagnostic value of pleural fluid cytologic examination", *Medicina (Kaunas, Lithuania)*, *38*(12), pp.1171-1178, 2002.

(13) R. Kushwaha, P. Shashikala, S. Hiremath, and H.G. Basavaraj, Cells in pleural fluid and their value in differential diagnosis. *Journal of Cytology*, *25*(4), p.138, 2008.

(14) A.J. Mach, O.B. Adeyiga, and D. Di Carlo, “Microfluidic sample preparation for diagnostic cytopathology”, *Lab on a Chip*, *13*(6), pp.1011-1026, 2013.

(15) “Thoracentesis”, National Health Institute.

(16) “Chest tube insertion: Procedure, complications, and removal”, https://www.medicalnewstoday.com/articles/322161.php.

(17) A.M. Khan, “Algorithms for breast cancer grading in digital histopathology images”, (Doctoral dissertation, University of Warwick), 2014.

(18) Arthur W Wetzel, R Gilbertson II John, Jeffrey A Beckstead, Patricia A Feineigle, Christopher R Hauser, and Frank A Palmieri Jr. System for creating microscopic digital montage images, December 26 2006. US Patent 7,155,049.

(19) F. Ghaznavi, A. Evans, A. Madabhushi, and Feldman, “Digital imaging in pathology: whole-slide imaging and beyond”, *Annual Review of Pathology: Mechanisms of Disease*, *8*, pp.331-359, 2013.

(20) OLYMPUS. URL: http://lri.se. (Accessed on 12th Oct 2018)

(21) 3DHISTECH. 3DHISTECH. URL http://www.3dhistech.com/. (Accessed on 12th Oct 2018).

(22) L. Zhang, Q. Wang and J. Qi, Research based on fuzzy algorithm of cancer cells in pleural fluid microscopic images recognition. In *Intelligent Information Hiding and Multimedia Signal Processing, 2006. IIH-MSP'06. International Conference on* (pp. 211-214). IEEE. December, 2006.

(23) F. Chen, J. Xie, H. Zhang, and D. Xia, A technique based on wavelet and morphology transform to recognize the cancer cell in pleural effusion. In *Medical Imaging and Augmented Reality, 2001. Proceedings. International Workshop on* (pp. 199-203). IEEE. 2001.

(24) E. Baykal, H. Dogan, M. Ekinci, M.E. Ercin, and S. Ersoz, Automated Cell Nuclei Segmentation in Pleural Effusion Cytology Using Active Appearance Model. In *International Conference on Computer Analysis of Images and Patterns* (pp. 59-69). Springer, Cham. 2017, August.

(25) E. Baykal, H. Doğan, M. Ekinci, M.E. Ercin, and Ş. Ersöz, Automated nuclei detection in serous effusion cytology based on machine learning. In *Signal Processing and Communications Applications Conference (SIU), 2017 25th* (pp. 1-4). IEEE. May, 2017.

(26) Jae S. Lim, “Two-Dimensional Signal and Image Processing,” Englewood Cliffs, NJ, Prentice Hall, 1990, pp. 469-476.

(27) K. Zuiderveld, “Contrast Limited Adaptive Histograph Equalization”, Graphic Gems IV. San Diego: Academic Press Professional, 1994. 474–485, (1994).

(28) K.Y. Win, S. Choomchuay, and K. Hamamoto, Automated segmentation and isolation of touching cell nuclei in cytopathology smear images of pleural effusion using distance transform watershed method. In *Second International Workshop on Pattern Recognition* (Vol. 10443, p. 104430Q). International Society for Optics and Photonics. , June, 2017.

(29) K.Y. Win, S. Choomchuay, and K. Hamamoto, K mean clustering based automated segmentation of overlapping cell nuclei in pleural effusion cytology images. In *Advanced Technologies for Communications (ATC), 2017 International Conference on* (pp. 265-269). IEEE. October, 2017.

(30) K.Y. Win, S. Choomchuay, and K. Hamamoto and M. Raveesunthornkiat, Artificial neural network based nuclei segmentation on cytology pleural effusion images. In *Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2017 International Conference on* (pp. 245-249). IEEE. 2017, November.

(31) Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". IEEE Trans. Sys., Man., Cyber. 9 (1): 62– 66.

(32) Velasco, F.R., 1979. Thresholding using the ISODATA clustering algorithm (No. CSC-TR-751). MARYLAND UNIV COLLEGE PARK COMPUTER SCIENCE CENTER.

(33) Kapur, J.N., Sahoo, P.K. and Wong, A.K., 1985. A new method for gray-level picture thresholding using the entropy of the histogram. Computer vision, graphics, and image processing, 29(3), pp.273-285.

(34) Li, C.H. and Lee, C.K., 1993. Minimum cross entropy thresholding. Pattern recognition, 26(4), pp.617-625.

(35) Tobias, O.J. and Seara, R., 2002. Image segmentation by histogram thresholding using fuzzy sets. IEEE transactions on Image Processing, 11(12), pp.1457-1465.

(36) Kittler, J. and Illingworth, J., 1986. Minimum error thresholding. Pattern recognition, 19(1), pp.41-47.

(37) Bradley, D. and Roth, G., 2007. Adaptive thresholding using the integral image. Journal of Graphics Tools, 12(2), pp.13-21.

(38) Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), pp.881-892.

(39) Bezdek, J.C., Ehrlich, R. and Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), pp.191-203.

(40) Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE transactions on pattern analysis and machine intelligence, 17(8), pp.790-799.

(41) Chan, T.F. and Vese, L.A., 2001. Active contours without edges. IEEE Transactions on image processing, 10(2), pp.266-277.

(42) Vicente, S., Kolmogorov, V. and Rother, C., 2008, June. Graph cut based image segmentation with connectivity priors. In Computer vision and pattern recognition, 2008. CVPR 2008. IEEE conference on (pp. 1-8). IEEE.

(43) R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, *Slic superpixels* (No. EPFL-REPORT-149300). 2010.

(44) R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, *34*(11), pp.2274-2282. 2012.

(45) N. Dhanachandra, K. Manglem, and Y.J. Chanu, Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, *54*(2015), pp.764-771. 2015.

(46) Digital Image Processing (Third Edition) by Rafael C. Gonzalez and Richard E. Woods, ISBN 978-93-325-7032- 0(2008).

(47) N. Malpica, C. Ortiz de Solórzano, and J.J. Vaquero et al., "Applying watershed algorithms to the segmentation of clustered nuclei", *Cytometry*, 28(4):289–297, 1997.

(48) X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy", *IEEE Transactions on Circuits and Systems I: Regular Papers*, *53*(11), pp.2405-2414, 2006.

(49) T.T.E. Yeo, X.C. Jin, S.H. Ong, and R. Sinniah, "Clump splitting through concavity analysis", Pattern Recognition Letters, 15(10), pp.1013-1018, 1994.

(50) X. Bai, C. Sun, and F. Zhou, "Splitting touching cells based on concave points and ellipse fitting", *Pattern recognition*, 42(11), pp.2434-2446, 2009.

(51) S. Kumar, S.H. Ong, and S. Ranganath et al., "A rule-based approach for robust clump splitting", *Pattern Recognition***,** 39 (6) 1088–1098, 2006.

(52) H. Wang, H. Zhang, and N. Ray, "Clump splitting via bottleneck detection and shape classification. Pattern Recognition", 45(7), pp.2780-2787, 2012.

(53) S. Tafavogh, D.R. Catchpoole, and P.J. Kennedy, "Non-parametric and integrated framework for segmenting and counting neuroblastic cells within neuroblastoma tumor images", Medical & biological engineering & computing, 51(6), pp.645-655, 2013.

(54) S. Tafavogh, D.R. Catchpoole, and P.J. Kennedy, "Cellular quantitative analysis of neuroblastoma tumor and splitting overlapping cells", BMC bioinformatics, 15(1), p.272, 2014.

(55) N. Abbas, A.H. Abdullah, Z. Mohamad, and A. Altameem, "Clustered red blood cell splitting via boundary analysis in microscopic thin blood smear digital images", *Int. J. Technol*, 3, pp.306-317, 2015.

(56) M. Guven, and C. Cengizler, "Data cluster analysis-based classification of overlapping nuclei in Pap smear samples", Biomedical engineering online, 13(1), p.159, 2014.

(57) L. Guerra, L.M. McGarry, V. Robles, C. Bielza, P. Larranaga, and R. Yuste, "Comparison between supervised and unsupervised classifications of neuronal cell types: a case study", *Developmental neurobiology*, *71*(1), pp.71-82, 2011.

(58) G.N. Srinivasan, and G. Shobha, "Statistical texture analysis", *In Proceedings of world academy of science, engineering and technology* (Vol. 36, pp. 1264-1269), December, 2008.

(59) H.T. Kam, "Random decision forest" *In Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition*, Montreal, Canada, August (pp. 14-18), August, 1995.

(60) T.K. Ho, "The random subspace method for constructing decision forests", *IEEE transactions on pattern analysis and machine intelligence*, 20(8), pp.832-844, 1998.

(61) L. Breiman, "Random forests", *Machine learning*, 45(1), pp.5-32, 2001.

(62) M.B. Kursa, "Robustness of Random Forest-based gene selection methods", BMC bioinformatics, 15(1), p.8, 2014.

(63) R. Genuer, J.M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests", *Pattern Recognition Letters*, 31(14), pp.2225-2236, 2010.

(64) R. Díaz-Uriarte, and S.A. De Andres, "Gene selection and classification of microarray data using random forest", *BMC bioinformatics*, 7(1), p.3, 2006.

(65) G. Suna, S. Lia, Y.Caoa, and F. Lang, "Cervical cancer diagnosis based on random forest", *Int. J. Performabil. Eng*, 13, pp.446-457, 2017.

(66) V. Krishnaiah, D.G. Narsimha, and D.N.S. Chandra, "Diagnosis of lung cancer prediction system using data mining classification techniques", *International Journal of Computer Science and Information Technologies*, *4*(1), pp.39-45, 2013.

(67) S. Zafari, T. Eerola, J. Sampo, H. Kälviäinen, and H. Haario, December. Segmentation of partially overlapping nanoparticles using concave points. In *International Symposium on Visual Computing* (pp. 187-197). Springer, Cham. 2015.

(68) T. Fawcett, "An introduction to ROC analysis", *Pattern recognition letters*, 27(8), pp.861-874, 2006.

(69) H. Zhang, "The optimality of naive Bayes", *AA*, *1*(2), p.3, 2004.

(70) A. Shmilovici, "Support vector machines", In *Data mining and knowledge discovery handbook* (pp. 231-247). Springer, Boston, MA, 2009.

(71) O. Sutton, "Introduction to k nearest neighbor classification and condensed nearest neighbour data reduction", *University lectures, University of Leicester*, 2012.

(72) L. Rokach, and O.Z. Maimon, "Data mining with decision trees: theory and applications (Vol. 69)", *World scientific*, 2008.

(73) D. Bassen, S. Nayak, X.C. Li, M. Sam, J. Sidhu, M.F.Nelson, and W.H. Land Jr, "Clinical Decision Support System (CDSS) for the Classification of Atypical Cells in Pleural Effusions", Procedia Computer Science, 20, pp.379-384, (2013).

(74) T.J. Vargason, J. Cohn, D. Rios, O. Schultz, J. Cleary, D. Lau, W. Land, J.D. Schaffer, Y. Li, C.A. Chou, and S.A. Syouri, "A Clinical Decision Support System for Malignant Pleural Effusion Analysis," (2016), Justice & Well-Being Studies Faculty Scholarship. 1.

(75) K. Rodenacker, and E. Bengtsson, A feature set for cytometry on digitized microscopic images. *Analytical Cellular Pathology*, *25*(1), pp.1-36. 2003.

(76) P. Wang, X. Hu, Y. Li, Q. Liu, and X. Zhu, Automatic cell nuclei segmentation and classification of breast cancer histopathology images. *Signal Processing*, *122*, pp.1-13. 2016.

(77) P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak, Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *IEEE Transactions on Medical Imaging*, *32*(12), pp.2169-2178. 2013.

(78) M.A. Devi, S. Ravi, J. Vaishnavi, and S. Punitha, Classification of cervical cancer using artificial neural networks. *Procedia Computer Science*, *89*, pp.465-472. 2016.

(79) J. Su, X. Xu, Y. He, and J. Song, Automatic detection of cervical cancer cells by a two-level cascade classification system. *Analytical Cellular Pathology*, *2016*.

(80) S. Rathore, M. Hussain, M.A. Iftikhar, and A. Jalil, Ensemble classification of colon biopsy images based on information rich hybrid features. *Computers in Biology and Medicine*, *47*, pp.76-92. 2014.

(81) A. Sengur, Color texture classification using wavelet transform and neural network ensembles, Arabian J. Sci. Eng. 34 (2009) 491–502.

(82) R.M Haralick, and K. Shanmugam, Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), pp.610-621. 1973.

(83) W. Gómez, W.C.A. Pereira and A.F.C. Infantosi, Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound. *IEEE transactions on medical imaging*, *31*(10), pp.1889-1899. 2012.

(84) B.V. Dasarathy, and E.B. Holder, Image characterizations based on joint gray level—run length distributions. *Pattern Recognition Letters*, *12*(8), pp.497-502. 1991.

(85) I. Guyon and A. Elisseeff, An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), pp.1157-1182. 2003.

(86) R.K. Sivagaminathan, and S. Ramakrishnan, A hybrid approach for feature subset selection using neural networks and ant colony optimization. Expert systems with applications, 33(1), pp.49-60, 2007.

(87) D.L. Tong, and A.C. Schierz, Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data. Artificial intelligence in medicine, 53(1), pp.47-56, 2011.

(88) E. Alba, J. Garcia-Nieto, L. Jourdan, and E.G. Talbi, Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In Evolutionary Computation, 2007. CEC 2007. IEEE Congress on (pp. 284-290). IEEE, September, 2007.

(89) S. Kirkpatrick, C. D. Gelatt, and M. P . V ecchi, "Optimization by Simulated Annealing," *Science,* vol. 220, pp. 671-680, May 13, 1983 1983.

(90) S.W. Lin, Z.J. Lee, S.C. Chen, and T.Y. Tseng, Parameter determination of support vector machine and feature selection using simulated annealing approach. *Applied soft computing*, *8*(4), pp.1505-1512. 2008.

(91) M. Filippone, F. Masulli, and S. Rovetta, Simulated annealing for supervised gene selection. *Soft Computing*, *15*(8), pp.1471-1482. 2011.

(92) F. González, and L.A. Belanche, Feature selection for microarray gene expression data using simulated annealing guided by the multivariate joint entropy. *arXiv preprint arXiv:1302.1733*. 2013.

(93) B.M. Wilamowski, and H. Yu, Improved computation for Levenberg–Marquardt training. *IEEE transactions on neural networks*, *21*(6), pp.930-937. 2010.

(94) Yarpiz, http://yarpiz.com.

(95) S. Huda, J. Yearwood, H.F. Jelinek, M.M. Hassan, G. Fortino, and M. Buckland, A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. *IEEE Access*, *4*, pp.9145-9154. 2016.

(96) S. Nagi, and D.K. Bhattacharyya, Classification of microarray cancer data using ensemble approach. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *2*(3), pp.159-173. 2013.

(97) T.G. Dietterich, Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer, Berlin, Heidelberg. 2000, June.

(98) A.M. Prasad, L.R. Iverson, and A. Liaw, Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9(2), pp.181-199, 2006.

(99) C. Cortes, and V. Vapnik, Support-vector networks. Machine learning, 20(3), pp.273-297. 1995.

(100) M.T. Hagan, H.B. Demuth, and M.H. Beale, Neural network design (Vol. 20). Boston: Pws Pub. 1996.

(101) T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.

(102) C. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.

(103) G. McLachlan, Discriminant analysis and statistical pattern recognition (Vol. 544). John Wiley & Sons. 2004.

(104) L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Monterey, CA: Wadsworth Brooks/Cole Advanced Books Software, 1984.

(105) Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

(106) G.E. Hinton, S. Osindero, Y.-W Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput*. 2006, *18*, 1527–1554.

(107) P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion", *J. Mach. Learn. Res*. 2010, *11*, 3371–3408.

(108) A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1106–1114, 2012.

(109) K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778), 2016.

(110) Y. Taigman, M. Yang, M.A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708), 2014.

(111) A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, *542*(7639), p.115, 2017.

(112) W. Li, P. Cao, D. Zhao, and J. Wang, "Pulmonary nodule classification with

deep convolutional neural networks on computed tomography images," *Computational and mathematical methods in medicine*, *2016.*

(113) J.Y. Choi, T.K. Yoo, J.G. Seo, J. Kwak, T.T. Um, and T.H. Rim, "Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database", *PloS one*, *12*(11), p.e0187336, 2017.

(114) D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with Deep Learning," *Scientific reports*, *8*(1), p.4165, 2018.

(115) H. Wang, A. Cruz-Roa, A. Basavanhally et al., "Mitosis detection in breast cancer pathology images by combining hand- crafted and convolutional neural network features," *Journal of Medical Imaging*, vol. 1, no. 3, p. 034003, 2014.

(116) L. Zhang, L. Lu, I. Nogues, R.M. Summers, S. Liu, and J. Yao, "DeepPap: Deep convolutional networks for cervical cell classification", *IEEE journal of biomedical and health informatics*, *21*(6), pp.1633-1643, 2017.

(117) H. Sharma, N. Zerbe, I. Klempert, O. Hellwich, and P. Hufnagl, "Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology," *Computerized Medical Imaging and Graphics*, *61*, pp.2-13, 2017.

(118) M.G. Ertosun, and D.L. Rubin, "Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks," In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 1899). American Medical Informatics Association, 2015.

(119) E. Arvaniti, K.S. Fricker, M. Moret, N.J. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P.J. Wild, J.H. Rueschoff, and M. Claassen, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning", *bioRxiv*, p.280024, 2018.

(120) S.J. Pan, and Q. Yang, "A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*," *22*(10), pp.1345-1359, 2010.

(121) L. Torrey, and J. Shavlik, "Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques",* (pp. 242-264). IGI Global, 2010.

(122) T. Akiba, S. Suzuki, and K. Fukuda, "Extremely large minibatch sgd: Training

resnet-50 on imagenet in 15 minutes", *arXiv preprint arXiv:1711.04325,* 2017

(123) S. Targ, D. Almeida, and K. Lyman, "Resnet in Resnet: generalizing residual architectures", *arXiv preprint arXiv:1603.08029*, 2016.

(124) ImageNet. http://www.image-net.org.

(125) Y.Miki, C.Muramatsu, T.Hayashietal.,"Classification of teeth in cone-beam CT using deep convolutional neural network," *Computers in Biology and Medicine*, vol. 80, pp. 24–29, 2017.

(126) C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9), 2015.

(127) BVLC GoogLeNet Model.
https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet.

(128) C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826). 2016.

(129) R.C. Gonzalez, R.E. Woods, Digital Image Processing, Pearson Education Inc., 2008.

AUTHOR BIOGRAPHY

Author: Ms. Khin Yadanar Win
Degree: Doctor of Engineering
Date: 18th February 2019
Date of Birth: 11th June 1992
Place of Birth: Yangon, Myanmar

Undergraduate and Graduate Education

Doctor of Engineering (Integrated Program of Master and Ph.D.) in Electrical Engineering
Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, 2019
Major: Electrical Engineering

Bachelor Degree of Honors in Computer Science
Department of Computer Studies
Dagon University, Yangon, Myanmar, 2013
Major: Computer Science

PRESENTATIONS, PUBLICATIONS AND AWARDS

International conference papers, international articles and awards included in the thesis resulting from the doctoral research program are listed as below:

International Conference Papers

(1) **K.Y. Win**, and S. Choomchuay, "Detection of Optic Disc and Exudates in Retinal Images", *In AUN/SEED Conf. for Computer and Information Engineering*, Yangon, Myanmar, October, 2016.

(2) **K.Y. Win**, and S. Choomchuay, "Automated detection of exudates using histogram analysis for Digital Retinal Images," Proc. of Int. Symposium on Intelligent Signal Processing and Communication Systems-2016 (ISPACS-2016), Phuket, Thailand, 1-4 December 2014.

(3) **K.Y. Win**, and S. Choomchuay, "Automated segmentation of cell nuclei in cytology pleural fluid images using OTSU thresholding," 2nd International Conference on Digital Arts, Multimedia and Technology (ICDAMT-2017), Chiang Mai, Thailand, February 17-20, 2017.

(4) **K.Y. Win**, S. Choomchuay, and K. Hamamoto, "Automated segmentation and isolation of touching cell nuclei in cytopathology smear images of pleural effusion using distance transform watershed method," Proc. SPIE 10443, *Second International Workshop on Pattern Recognition*, Singapore, 104430Q (June 19, 2017)

(5) **K.Y. Win**, S. Choomchuay, and K. Hamamoto, "Features Extraction from Segmented Cell Nuclei of Cytological Smear Images for Detection of Cancer Cells in Pleural Fluid," Proc. of *29th International Technical Conference on Circuit/Systems Computers and Communications* (ITC-CSCC 2017), Pusan, Korea, July 2-5, 2017, pp.1043-1046.

(6) **K.Y. Win**, S. Choomchuay, and K. Hamamoto, "K-Mean Clustering Based Automated Segmentation of Overlapping Cell Nuclei in Pleural Effusion Cytology Images," Proc. of International Conference on Advanced Technologies for Communications (ATC 2017), Quynhon, Vietnam, October 18-20, 2017.

(7) **K.Y. Win**, S. Choomchuay, and K. Hamamoto, "Artificial Neural Network

Based Nuclei Segmentation on Cytology Pleural Effusion Images," International Conference on Intelligent Informatics and BioMedical Sciences (ICIIBMS 2017), Okinawa, Japan, November 24-26, 2017.

(8) **K.Y. Win**, S. Choomchuay, N. Maneerat, K. Hamamoto and S. Sreng, "Suitable Machine Learning Methods for Malignant Mesothelioma Diagnosis", the 2018 Biomedical Engineering International Conference (BMEiCON 2018), Chaing Mai, Thailand, November 21-24, 2018.

(9) S. Syna, N. Maneerat, K. Hamamoto, R. Panjaphongse and **K.Y. Win**, "Classification of Cotton Wool Spots Using Principal Components Analysis and Support Vector Machine", the 2018 Biomedical Engineering International Conference (BMEiCON 2018), Chaing Mai, Thailand, November 21-24, 2018.

International Journal Papers

(1) **K.Y. Win**, S. Choomchuay, K. Hamamoto, and M. Raveesunthornkiat, "Supervised Learning-based Nuclei Segmentation on Cytology Pleural Effusion Images with Artificial Neuron Network," *Applied Science and Computer Science Publications*, (ASCSP), Vol.3, Issue 3, December 2017, pp. 104-110.

(2) **K.Y. Win**, S. Choomchuay, K. Hamamoto, and M. Raveesunthornkiat, "Comparative Study on Automated Cell Nuclei Segmentation Methods for Cytology Pleural Effusion Images", Journal of Healthcare Engineering, vol. 2018, Article ID 9240389, 14 pages, 2018. https://doi.org/10.1155/2018/9240389.

(3) **K.Y. Win**, S. Choomchuay, K. Hamamoto, and M. Raveesunthornkiat "Detection and Classification of Overlapping Cell Nuclei in Cytology Effusion Images Using a Double-Strategy Random Forest," Advanced Intelligent Imaging, Adanced Intelligent Systems, *Ap. S.,* **2018**, *8*, 1608. https://doi.org/10.3390/app8091608.

(4) **K.Y. Win**, S. Choomchuay, K. Hamamoto, and M. Raveesunthornkiat, "Computer Aided Diagnosis System for Detection of Cancer Cells on Cytological Pleural Effusion Images," BioMed Research International, vol. 2018, Article ID 6456724, 21 pages, 2018. https://doi.org/10.1155/2018/6456724.

(5) **K.Y. Win**, S. Choomchuay, K. Hamamoto, N. Maneerat, M. Raveesunthornkiat and S. Sreng, "Nuclei Textural Pattern Classification of Cytological Pleural Effusion Images Using Integrated mRMR Feature Selection and Support Vector Machine", Submitted for the peer-review in international journal.

Awards

(1) Best Presentation Award

Second International Workshop on Pattern Recognition (IWPR), Singapore, 2017.

(2) Student Paper Award

International Conference on Intelligent Informatics and Biomedical Sciences, (ICIIBMS), Okinawa, Japan, 2017.

(3) Best Paper Award

The Journal of Bioinformatics and Neurosciences (JBINS), Applied Science and Computer Science Publications, 2017.