การจำแนกลักษณะเนื้อแท้รอยโรคปอดจากภาพเอ็นโดบรองเคียลที่บันทึกด้วยคลื่นเสียงความถี่สูงโดยใช้การวิเคราะห์ลักษณะเนื้อแท้

นางสาวบรรพตรี คมขำ

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ ที่ส่งผ่านทางบัณฑิตวิทยาลัย
The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository (CUIR)
are the thesis authors' files submitted through the University Graduate School.

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาคณิตศาสตร์ประยุกต์และวิทยาการคณนา ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

PULMONARY LESION TEXTURE CLASSIFICATION FROM ENDOBRONCHIAL ULTRASONOGRAM USING TEXTURE ANALYSIS

Miss Banphatree Khomkham

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Applied Mathematics and Computational Science
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2017

Copyright of Chulalongkorn University

Thesis Title	PULMONARY LESION TEXTURE CLASSIFICATION FROM ENDOBRONCHIALULTRASONOGRAM USING TEXTURE ANALYSIS
By	Miss Banphatree Khomkham
Field of Study	Applied Mathematics and Computational Science
Thesis Advisor	Associate Professor Rajalida Lipikorn, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

....Dean of the Faculty of Science

(Professor Polkit Sangvanich, Ph.D.)

THESIS COMMITTEE

....Chairman

(Assistant Professor Khamron Mekchay, Ph.D.)

....Thesis Advisor

(Associate Professor Rajalida Lipikorn, Ph.D.)

....Examiner

(Associate Professor Nagul Cooharojananone, Ph.D.)

....External Examiner

(Suriya Natsupakpong, Ph.D.)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บรรพตรี คมขำ : การจำแนกลักษณะเนื้อแท้รอยโรคปอดจากภาพเอ็นโดบรองเคียลที่บันทึกด้วยคลื่นเสียงความถี่สูงโดยใช้การวิเคราะห์ลักษณะเนื้อแท้ (PULMONARY LESION TEXTURE CLASSIFICATION FROM ENDOBRONCHIALULTRASONOGRAM USING TEXTURE ANALYSIS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร. รัชลิดา ลิปิกรณ์, 71 หน้า.

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีในการช่วยแยกลักษณะรอยโรคปอดจากภาพบันทึกด้วยคลื่นเสียงความถี่สูง (Endobronchial Ultrasound—EBUS) ตามการศึกษาข้อมูลทางการแพทย์ พบว่าลักษณะของความเรียบหรือขรุขระ สามารถบ่งบอกว่าภาพนั้นเป็นมะเร็งหรือเนื้องอกอย่างมีนัยสำคัญ ในการศึกษานี้ได้แบ่งกลุ่มลักษณะเด่นที่ใช้ในการคัดแยกออกเป็น 3 กลุ่ม กลุ่มที่ 1 เป็นการใช้ลักษณะเด่นแบบมาตรฐานจำนวน 22 ลักษณะ กลุ่มที่ 2 เป็นลักษณะเด่นที่นำเสนอโดยการสกัดลักษณะเด่นมาจากผลบวกของเมทริกซ์บนและล่างของเมทริกซ์การเกิดร่วมกันของค่าระดับสีเทาแบบถ่วงน้ำหนักจำนวน 12 ลักษณะ และกลุ่มที่ 3 เป็นการรวมลักษณะเด่นในกลุ่ม 1 และกลุ่ม 2 เข้าด้วยกัน ซึ่งการจำแนกชนิดของเนื้องอกในงานนี้จะเลือกลักษณะเด่นที่ดีที่สุดโดยใช้วิธีการเลือกลักษณะเด่น 3 วิธี ประกอบด้วย วิธีเลือกไปข้างหน้า วิธีเลือกถอยหลัง และ วิธีเลือกเชิงพันธุกรรม ร่วมกับ ตัวจำแนก 8 วิธี ขั้นตอนโดยรวมที่ใช้ในงานนี้ประกอบด้วย กระบวนการจัดการภาพเบื้องต้น การเลื่อนหน้าต่าง การสกัดลักษณะเด่น การคัดเลือกลักษณะเด่น และการจำแนกประเภทของเนื้องอก ภาพเนื้องอกที่ใช้ในงานนี้มาจากผู้ป่วยจำนวน 89 ราย ที่ได้รับการยืนยันโดยแพทย์ผู้เชี่ยวชาญว่าเป็นมะเร็งจำนวน 55 ราย และเนื้องอกจำนวน 34 ราย เมื่อนำผลการจำแนกเนื้องอกที่ได้จากวิธีที่นำเสนอพบว่าการจำแนกให้ความถูกต้องสูงสุดเมื่อใช้ลักษณะเด่นในกลุ่มที่ 3 กับวิธีเลือกเชิงพันธุกรรมและซัพพอร์ตเวกเตอร์แมชชีน ที่ให้อัตราความถูกต้องอยู่ที่ 86.517%

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์	ลายมือชื่อนิสิต	....
		ลายมือชื่อ อ.ที่ปรึกษาหลัก	....
สาขาวิชา	คณิตศาสตร์ประยุกต์และวิทยาการคณนา		
ปีการศึกษา	2560		

5772035923 : MAJOR APPLIED MATHEMATICS AND COMPUTATIONAL SCIENCE
KEYWORDS: GLCM / PULMONARY LESION CLASSIFICATION / SUPPORT VECTOR MACHINE (SVM) / GENETIC SELECTION / LUNG CANCER

BANPHATREE KHOMKHAM: PULMONARY LESION TEXTURE CLASSIFICATION FROM ENDOBRONCHIALULTRASONOGRAM USING TEXTURE ANALYSIS. ADVISOR: ASSOC. PROF. RAJALIDA LIPIKORN, Ph.D., 71 pp.

This research aims to develop a method to help distinguish the appearance of pulmonary lesions from a high-frequency sound (Endobronchial Ultrasound—EBUS) image. According to medical information, the appearance of smooth or rough texture of a lesion can significantly indicate that it is malignant or benign. In this study, the features that are used in the classification are divided into 3 groups: group 1 consists of 22 standard features, group 2 consists of the proposed features extracted from the weighted sum of the upper and lower GLCM which consists 12 features, and group 3 is the combination of group 1 and group 2. Not all the features in each group are used in the classification, only the best features are selected from each groups using three feature selection techniques: forward selection, backward selection, and genetic selection. After the best features are selected, they are entered into eight different classifiers for the classification. The overall process of the classification consists of preprocessing, window slicing, feature extraction, feature selection, and classification. The sample input consists of 89 lesion images where 55 of them are identified by the doctor as malignant and 34 of them are identified as benign. The classification results show that the highest accuracy rate of 86.517% can be obtained by using features from group 3 with genetic selection and support vector machine.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Department: Mathematics and Computer Science
Field of Study: Applied Mathematics and Computational Science
Academic Year: 2017

Student's Signature ....
Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to Assoc. Prof. Rajalida Lipikorn, Ph.D., my research advisor, for her valuable guidance, kind supervision, patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis.

Beside my advisor, I would like to thank my thesis committees, Asst. Prof. Khomron Mekchay, Ph.D., Assoc. Prof. Nagul Cooharojananone, Ph.D. and Suriya Natsupakpong, Ph.D., for the valuable assistance and helpful suggestions.

I would like to express my great appreciation to Col. Anan Wattanathum, MD., and Jutamas Dechsanga, MD. from Pulmonary and Critical Care Division, Department of Medicine, Phramongkutklao Hospital who supported me for clinical advice and research data.

I also would like to express my great appreciation to all my teacher who has given all knowledge to me from the past until now. I am immensely obliged to my friends in ACMS and MIMIT lab for their elevating inspiration and encouraging guidance.

Additionally, this research was supported by the Development and Promotion of Science and Technology Talent Project (DPST), under the administration of the Institute for the Promotion of Teaching Science and Technology (IPST), Thailand.

Finally, I would like to thank my family for their encouragement and support throughout my study.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

CONTENTS

Page

THAI ABSTRACT iv

ENGLISH ABSTRACT v

ACKNOWLEDGEMENTS vi

CONTENTS vii

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER 1 INTRODUCTION 1

1.1. Research objectives 3

1.2. Thesis overview 3

CHAPTER 2 LITERATURE REVIEW AND BACKGROUND KNOWLEDGE 4

2.1. Literature review 4

2.1.1. The characteristics of echoes of malignant 4

2.1.2. Texture analysis 6

2.2. Background knowledge 8

2.2.1. EBUS images 8

2.2.2. Digital images 8

2.2.3. Grayscale images 9

2.2.4. RGB images 9

2.2.5. RGB to gray-scale conversion 10

2.2.6. Texture feature 10

2.2.7. Feature selection 14

2.2.8. Classification 15

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Page

2.2.9. Performance measurement 16

CHAPTER 3 METHODOLOGY 19

3.1. Input data 19

3.2. Preprocessing 20

3.2.1. File format conversion 20

3.2.2. Frame selection 20

3.2.3. Boundary detection 21

3.2.4. Region of interest selection 21

3.3. Best window selection 22

3.4. Feature extraction 23

3.5. Feature selection 26

3.6. Classification 29

3.7. The proposed method 34

CHAPTER 4 EXPERIMENTAL RESULTS 38

4.1. The classification results of using standard features with genetic selection and the eight classifiers. 39

4.2. The classification results of using the proposed features with genetic selection and the eight classifiers. 40

4.3. The classification results of using both standard features and the proposed features with genetic selection and the eight classifiers. 41

4.4. The comparison of classification results using confusion matrix 42

CHAPTER 5 DISCUSSION AND CONCLUSION 44

REFERENCES 47

APPENDIX 50

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Page

APPENDIX A Best Frame with window of interest 51

APPENDIX B Feature extraction value 63

APPENDIX C Confusion matrix of all classifier 68

VITA 71

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

LIST OF TABLES

Table 2.1 .Confusion matrix.......... 16

Table 3.1. The parameters of forward selection.......... 27

Table 3.2. The parameters of backward selection 28

Table 3.3 .The parameters of genetic selection 29

Table 3.4 .The parameters of decision tree classifier 31

Table 3.5 .The parameters of neural network classifier 31

Table 3.6 .The parameters of linear regression classifier 32

Table 3.7 .The parameters of logistic regression classifier.......... 32

Table 3.8 .The parameters of support vector machine classifier.......... 34

Table 4.1 .The classification results of using standard features with the eight classifiers.......... 40

Table 4.2 .The classification results of using the proposed features with the eight classifiers.......... 41

Table 4.3 .The classification results of using both standard features and the proposed features with the eight classifiers. 42

Table 4.4 .The confusion matrix of classification results from using standard features with genetic selection and support vector machine 43

Table 4.5 .The confusion matrix of classification results from using the proposed features with genetic selection and support vector machine 43

Table 4.6 .The confusion matrix of classification results from using the proposed feature and standard features with genetic selection with support vector machine.. 43

Table 5.1 .The classification results of three group feature extraction with genetic selection with support vector machine classifier 45

Table 5.2 .Area under the curve.......... 46

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

LIST OF FIGURES

Figure 2.1 .Samples of the endobronchial ultrasound images 6

Figure 2.2 .Sample of EBUS image 8

Figure 2.3 .An 8-bit grayscale Image 9

Figure 2.4 .A 24-bit RGB Image 10

Figure 2.5 .The 4 directions of GLCM 11

Figure 2.6 .Gray-level co-occurrence matrix 11

Figure 2.7 .The process for creating a classification model 15

Figure 2.8 .An example of ROC curve 18

Figure 3.1 .The proposed method 19

Figure 3.2 .An example of frame selection 20

Figure 3.3 .Drawing boundary of the tumor 21

Figure 3.4 .Window sliding in the region between the tumor and the ring 22

Figure 3.5 .An upper and lower triangular GLCMs 25

Figure 3.6 .An upper and lower triangular GLCMs 25

Figure 3.7 .Diagram of the proposed process 30

Figure 3.8 .An example of frame with Metadata 35

Figure 3.9 .The frame after removing metadata 35

Figure 3.10 .Determining boundary of the tumor 36

Figure 3.11 .The intersection area between the tumor and the ring area 36

Figure 3.12 .Window slicing in the region between tumor and ring 37

Figure 3.13 .The intersection area between the tumor and the ring area 37

Figure 4.1 .The ratio of input data 38

Figure 5.1 .ROC curve 45

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

CHAPTER 1

INTRODUCTION

One of the best wishes that we all want to have is “Good health” which becomes the well-known proverb that we have always heard; i.e., “Health is wealth”. Despite the unavoidable facts that lives must experience birth, old-age, illness, and death but we all try to be healthy by taking good care of ourselves. According to the Ministry of Public Health of Thailand, cancer (19%) is the leading cause of death, followed by ischemic heart disease (12%) and strokes (10%) [1]. Among various types of cancers, lung cancer is one of the leading causes of cancer death worldwide. Each year, the world statistics reveal that more than 1.6 million deaths are from lung cancer which is more than colon and prostate cancers combined together [2].

Lung Cancer is one of the cells in the lung that is abnormal and rapidly grows until it becomes a tumor and may spread to other parts of the body. The cause of lung cancers depends on many factors such as smoking, environment, genetics and others. Lung cancers can be divided into two types. Type 1 is a non-small cell which is often found and it grows slowly. There are three kinds of type 1 cancers which are squamous cell carcinoma, adenocarcinoma, and large cell carcinoma. Type 2 is a small cell carcinoma which is rarely found and it grows rapidly. In order to cure of cancer, the doctor will decide what to be done, for example, surgery and radiotherapy in combination with chemotherapy can be chosen as the treatment to stop the growth of cancer cells of type 1 while the doctor may decide to use chemotherapy with surgery and provide radiation therapy, even if it does not detect any spread. Although type 2 is dangerous, it is not often found. Hence there are many studies about type 1 but not that many studies about type 2.

In the past, in order to diagnose the cause of tumor in the lung, a doctor needed to remove tumor from a patient and send it to the laboratory. If it was type 1 cancer, the treatment which may include another surgery was needed to be performed on a patient. However, a surgery is a major procedure that requires anesthesia. Patients need a long period of time to recover. If it was not type 1 cancer, it may be cured

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

without a surgery that is using only drug therapy because a surgery usually causes the unnecessary loss of mass in the lung. Thus, if a doctor can diagnose a tumor before a surgery, a doctor can decide whether a surgery is necessary. If a surgery is necessary, it could be performed only once, and the size of a cancer to be removed can be determined before a surgery. Therefore, knowing the tumor type is very important. Later, there exist many different techniques to generate detailed images of organs, soft tissues, bone and other internal body parts; for example, Computed Tomography (CT) scan, X-ray scan, and Magnetic Resonance Imaging (MRI) scan. The most up-to-date procedure for identifying texture of tissue and biopsy is called Endobronchial Ultrasonography (EBUS).

EBUS is a method which uses a small camera and high-frequency sound waves or ultrasound to generate an image. A doctor can observe the abnormalities within the bronchus clearly. This tool allows doctors to monitor abnormalities within the respiratory system by inserting the camera into the central tracheal cavity. While the ultrasound camera is attached to the end of the line, internal and external images of the bronchus are shown to the doctor. If a doctor finds a lump or other abnormality in the lymph nodes, a doctor can use a small needle attached to the end of the camera to penetrate the bronchial wall to absorb tissue or cut off the abnormal area immediately. The video which was recorded help the doctor to diagnose the type of tumor by considering characteristic of echoic.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

From studies of the characteristics of echoic of EBUS [3, 4], it is found that the texture of tumor or echoic has relationship with the types of tumor. Consider only texture of a tumor, the heterogeneous pattern tends to be malignant whereas the homogenous pattern tends to be benign. Heterogeneity is the characteristic of tumor texture with non-smoothed intensities. The question is how the measurement of smoothness can be determined. This measurement is very specific because different cases use different criteria. Hence, the texture analysis technique can help to solve this problem. Texture Analysis is used for many classifying tasks such as in medical fields and others. The examples of applied tasks are classification of the texture of wood [5], classification of ultrasonic liver images [6], and classification of skin cancer [7].

In this thesis, we propose a new feature and feature extraction called the weight sum lower and upper gray-level co-occurrence matrix which can be used to determine heterogeneous and homogeneous patterns of lung tumor texture. This proposed feature is used to build a classification model. Moreover, the efficacy of the new feature is compared with the standard features.

1.1. Research objectives

1. To propose how to select the window of interest.
2. To define the new feature for measuring homogeneity of a tumor based on the identity of malignant and benign.
3. To use the most suitable features to classify a tumor.

1.2. Thesis overview

The content of this thesis is divided into 5 chapters, beginning with the introduction in Chapter 1.

Chapter 2 presents all background knowledge.

Chapter 3 presents the methodology, how to select the window of interest for feature extraction and the proposed method.

Chapter 4 presents the classification results from using the proposed feature compared to the results from using the standard features.

Chapter 5 concludes the results and provides discussion.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

CHAPTER 2

LITERATURE REVIEW AND BACKGROUND KNOWLEDGE

2.1. Literature review

2.1.1. The characteristics of echoes of malignant

In medical field, there are many studies on lung cancer which tried to find the pattern of malignancy.

In 2002, Koriaki kurimoto et al. [3] developed a classification system for classifying benign and malignant via EBUS by comparing a pulmonary lesion and histology of a tumor using retrospective review. As the results, they divided the characteristics of tumor into three major classes. Class 1 is a homogeneous pattern, class 2 is hyperechoic dots and linear arcs pattern, and class 3 is a heterogeneous pattern. They found that 92% of class 1 are benign whereas 99% of Class 2 and Class 3 are malignant.

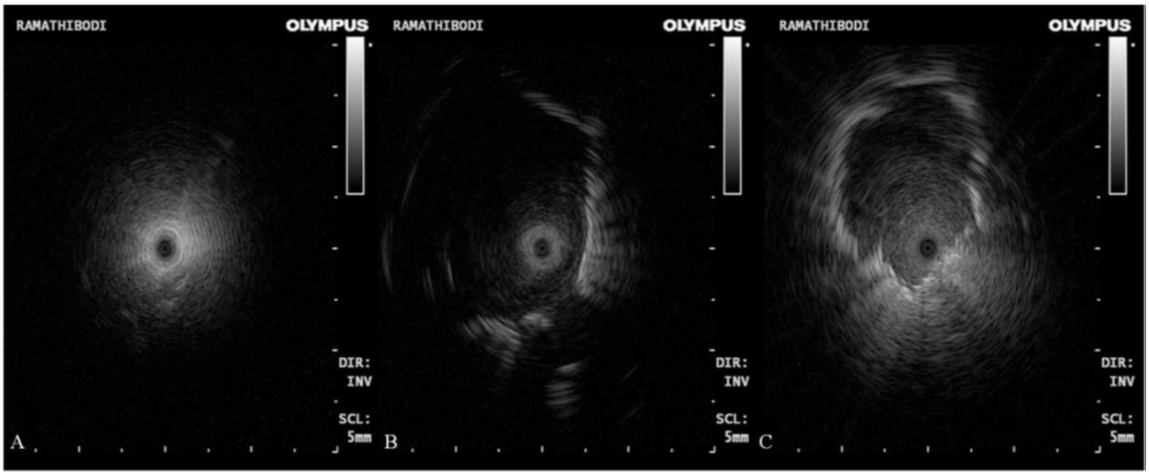
In 2006, Tung-Ying Chao et al. [8] created a common method to distinguish between neoplasm and nonneoplastic peripheral pulmonary lesions based on EBUS images. The study sample consisted of 151 patients. Twenty patients had already been diagnosed as having (1) continuous hyperechoic margin, (2) homogeneous, or heterogeneous internal echoes, (3) hyperechoic dots, and (4) concentric circles along the echo probe. Other 131 patients were diagnosed as the fifth case. The results reveal that 94.7% of homogenous internal echoes and 87.5% of concentric circles have nonneoplastic lesions. While continuous hyperechoic margins and hyperechoic dots did not yield a significant difference ($p = 0.090$ and $p = 0.079$, respectively).

In 2007, Chih-Hsi Kuo et al. [4] evaluated the EBUS according to three characteristic echoic features: continuous margin, absence of linear-discrete air bronchogram and heterogeneous pattern by observing 224 EBUS images of patients who had bronchoscopy. The results show that these three characteristics can be used to classify malignant from benign due to the negative predictive value is high to 93.7% for malignant tumors with none of these three characteristic echoic features and the positive predictive value is 89.2% for the malignant tumor with any two of the three-characteristic echoic features.

In 2009, Chien-Hao Lie et al. [9] studied characteristics of lesion from EBUS images to classify between neoplastic and non-neoplastic diseases. The study sample consisted of 2140 patients who were referred for bronchoscopic examination. Three image patterns of EBUS images, namely, hypoechoic areas, anechoic areas, and luminant areas around the probe were observed from initial forty patients. The results reveal that 85.7% of anechoic areas are neoplasms and 79.2% of lesions without luminant areas are non-neoplastic disease. In addition, both luminant and anechoic areas were significantly different between neoplastic and non-neoplastic categories. This study is not complicated and reducing time spent by using EBUS image patterns for diagnoses.

In 2015, Kei Morikawa Lie et al. [10] decided whether histogram information collected from EBUS-GS images can contribute to the determination of lung cancer. Histogram-based analyses were used to classify lung cancer and inflammatory diseases. In this research, median histogram height, width, height/width ratio and standard deviation were significantly different between lung cancer and benign lesion. The results show that standard deviation is the most effective feature to help diagnose lung cancer via EBUS images.

From the literature survey, many researchers found that heterogeneity and homogeneity are the characteristics which can be used to indicate whether a lesion is malignant or benign. As shown in Figure 2.1(a), an EBUS image of homogenous lesion with no boundary is a normal lesion, Figure 2.1(b) shows an EBUS image of granulomatous inflammation patient who has homogenous lesion with clear boundary that is benign, and Figure 2.1(c) shows an EBUS image of adenocarcinoma having heterogeneous lesion with continuous boundary that is malignant.



(a) (b) (c)

Figure 2.1. Samples of the endobronchial ultrasound images

(a) Homogenous lesion with no boundary which is normal

(b) Homogenous lesion with clear boundary which is benign

(c) Heterogeneous lesion with continuous boundary which is malignant

2.1.2. Texture analysis

In the computational field, texture analysis is the measurement of some features of the texture and can be used to identify the properties of texture. Many methods are used in texture analysis such as:

Gray-level co-occurrence matrix (GLCM) which is one of the most widely used methods for texture analysis in many applications such as in medical , industral, meterail ,and others, was first proposed by Haralick et al. [11] in 1973.

In 1973, Robert M. Haralick et al. [11] proposed a set of twenty-eight textural features based on gray level scale which needed uncomplicated computation but efficient, such as angular second moment, contrast, correlation, sum of squares, inverse difference moment, sum average, sum variance, entropy, and so on. Each feature represents different characteristic, for example mean represents the overall image, entropy represents the irregularity of the intensity and so on. The well- known properties of GLCM are energy, entropy, contrast, homogeneity, and correlation.These features are widely used in texture analysis. Nonetheless, standard features do not work for all types of images. Each type of image is unique. Finding the characteristics and creating associated features are the most important part of image classification. The limitation of their research is that each feature can be applied to any applications, thus it is too general and does not work for some special cases. For example, the more specific features from GLCM have been proposed by Walker and Zainudin [13, 14].

In 1992, Chung-Ming Wu et al. [6] used texture features to classify the ultrasonic liver images. The texture features were applied such as the spatial gray-level dependence matrices, the Fourier power spectrum, the gray-level difference statistics, and the Laws' texture energy measures. The study sample consisted of 90 samples which were divided into 30 samples of normal liver, 30 samples of hepatoma, and 30 samples of cirrhosis. The Bayes classifier and the Hotelling trace criteria were used to calculate the effect of features. The results reveal that the accuracy is not good enough. The process took long time and gave low accuracy rate. Thus, they presented the multiresolution fractal feature set to solve this problem. It was found that the multiresolution fractal feature is a great tool for extracting ultrasonic liver images.

In 2017, Mohamed Abdel-Nasser et al. [12] proposed the super-resolution technique to adjust ultrasound images of breast tumor before extracting five textural features: gray level co-occurrence matrix features, local binary patterns, phase congruency-based local binary pattern, histogram of oriented gradients and pattern lacunarity spectrum. This technique improves the performance of tumor classification by giving 0.99 of the area under curve. It is important to note that removal of any artifacts and noise can improve the performance.

This thesis presents the new feature and an alternative algorithm used to classify peripheral lesion of EBUS image whether it is benign or malignant based on homogeneous and heterogeneous pattern of internal echoes of an EBUS image. Texture analysis technique is applied to extract the information of image and classification model is used to find the most effective classification process.

2.2. Background knowledge

In this section, we present the principle knowledges which are used in our research work.

2.2.1. EBUS images

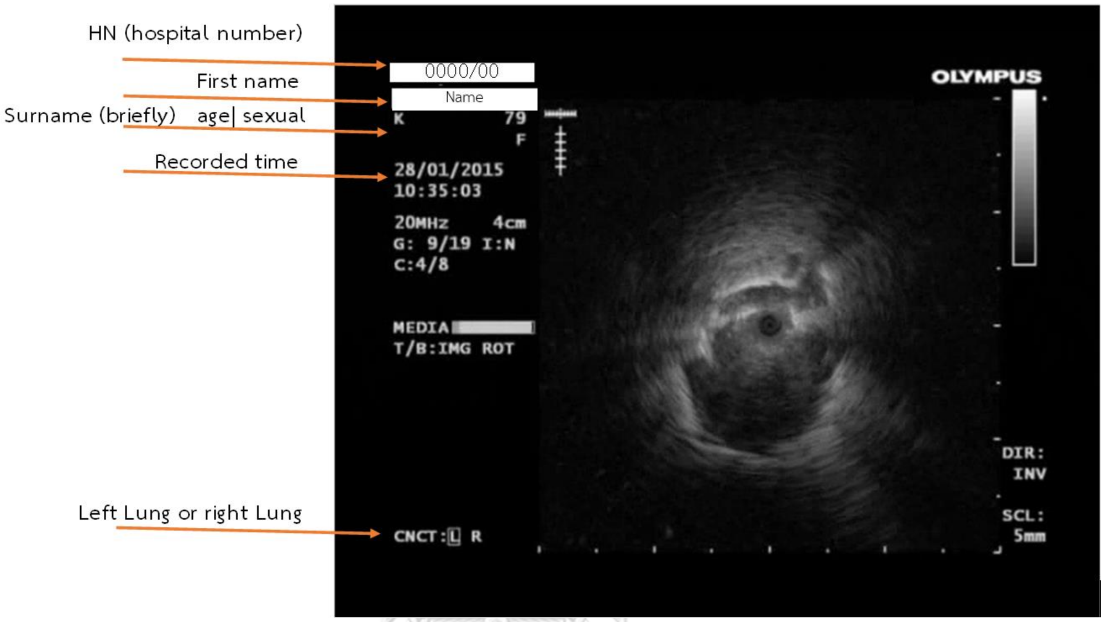


Figure 2.2. Sample of EBUS image

EBUS images can be extracted from EBUS video which was recorded via Endobronchial ultrasonography. An EBUS image is a 24-bit RGB image. Each EBUS image has details of a patient who had undergone Endobronchial ultrasonography, such as hospital number, first name, last name, age, gender, recorded time, the position of tumor in lung, the rang of frequency, and zooming distance as show in Figure 2.2.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

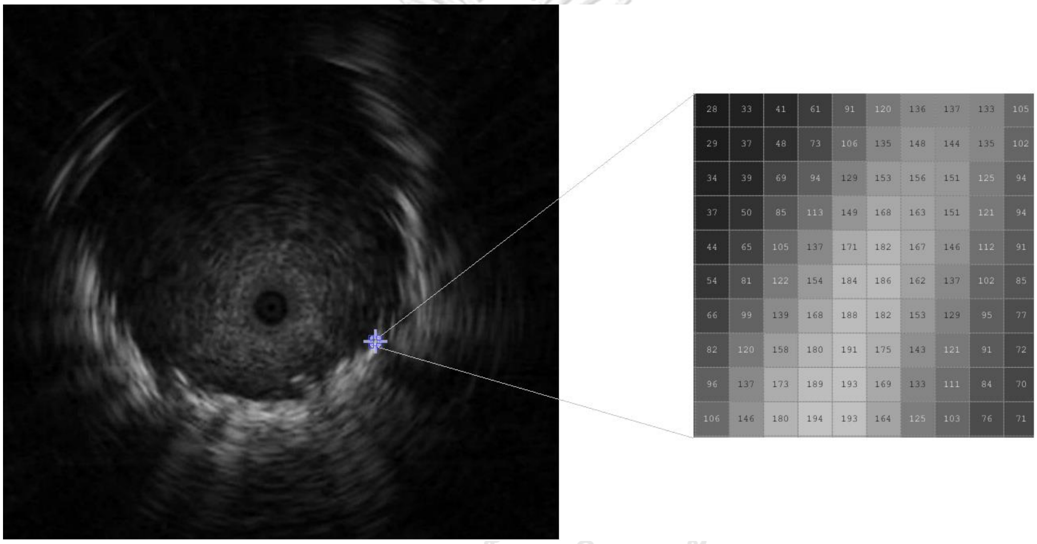
2.2.2. Digital images

In digital image, the coordinates of each pixel is represented by (x, y), where x represents row and y represents column. The origin point of an image is located at the upper left corner of an image. Let $\mathbf{I}$ be a matrix which represents an image of size $m \times n$. $I(x, y)$ is the element of $\mathbf{I}$ which represents the intensity of an image at position (x, y).

$$\mathbf{I} = \begin{bmatrix} I(0,0) & I(0,1) & \cdots & I(0,n-1) \\ I(1,0) & I(1,1) & \cdots & I(1,n-1) \\ \vdots & \vdots & \vdots & \vdots \\ I(m-1,0) & I(m-1,1) & \cdots & I(m-1,n-1) \end{bmatrix}$$

2.2.3. Grayscale images

Each pixel of a grayscale image contains a gray scale level or intensity. In an 8-bit image, the intensities begin from 0 (black) to 255 (white). Generally, the most widely used grayscale images (8-bits) have 256 levels of gray scales as shown in Figure 2.3.



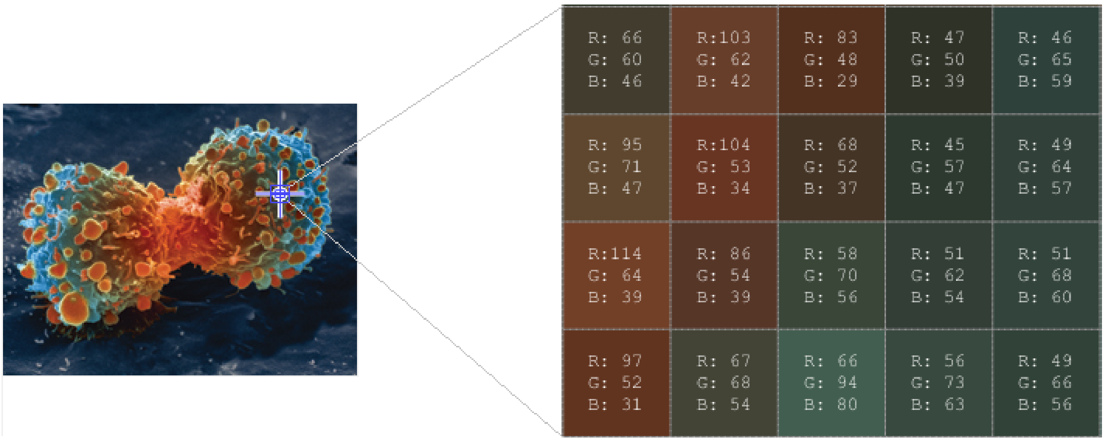
28	33	41	61	91	120	136	137	133	105
29	37	48	73	106	135	148	144	135	102
34	39	69	94	129	153	156	151	125	94
37	50	85	113	149	168	163	151	121	94
44	65	105	137	171	182	167	146	112	91
54	81	122	154	184	186	162	137	102	85
66	99	139	168	188	182	153	129	95	77
82	120	158	180	191	175	143	121	91	72
96	137	173	189	193	169	133	111	84	70
106	146	180	194	193	164	125	103	76	71

Figure 2.3. An 8-bit grayscale Image

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

2.2.4. RGB images

An RGB image is represented by a ratio of red, green, and blue colors. For example, a 24-bit RGB image uses 8 bits for each color as shown in Figure 2.4.



R: 66 G: 60 B: 46	R:103 G: 62 B: 42	R: 83 G: 48 B: 29	R: 47 G: 50 B: 39	R: 46 G: 65 B: 59
R: 95 G: 71 B: 47	R:104 G: 53 B: 34	R: 68 G: 52 B: 37	R: 45 G: 57 B: 47	R: 49 G: 64 B: 57
R:114 G: 64 B: 39	R: 86 G: 54 B: 39	R: 58 G: 70 B: 56	R: 51 G: 62 B: 54	R: 51 G: 68 B: 60
R: 97 G: 52 B: 31	R: 67 G: 68 B: 54	R: 66 G: 94 B: 80	R: 56 G: 73 B: 63	R: 49 G: 66 B: 56

Figure 2.4. A 24-bit RGB Image

(Cited: https://simple.wikipedia.org/wiki/Lung_cancer, April 13, 2018)

2.2.5. RGB to gray-scale conversion

Converting RGB value to grayscale value can be performed by using a weighted sum of the $R, G,$ and B values as expressed in Eq. (1):

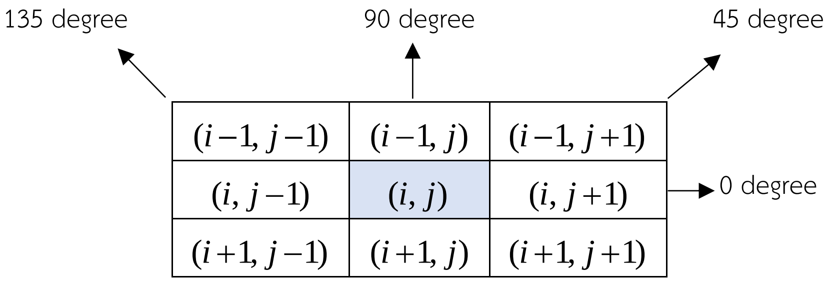
$$Grayscale = (0.2989 \times R) + (0.5870 \times G) + (0.1140 \times B) \tag{1}$$

2.2.6. Texture feature

The texture of an image is one of the most important characteristic to identify the type of object or the interesting area in an image such as medical imaging, satellite imaging, landscape imaging and so on. The textures of these images describe the characteristics of the images and can be used to interpret the content in an image. The characteristics of an image can be used to classify medical images such as EBUS images, CT images, mammogram, and ultrasound images. One of the most important characteristic of an image is the pattern of intensity distribution of an image, the changes of intensity levels of an image can be used in image classification.

In a grayscale image, the pattern of intensity distribution can be generated by using a Gray-Level Co-occurrence Matrix (GLCM). GLCM is a matrix whose elements represent the frequency of a pair neighbor intensities with interest direction. The formulas for standard properties of GLCM are, for example contrast, energy, homogeneity, and correlation.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

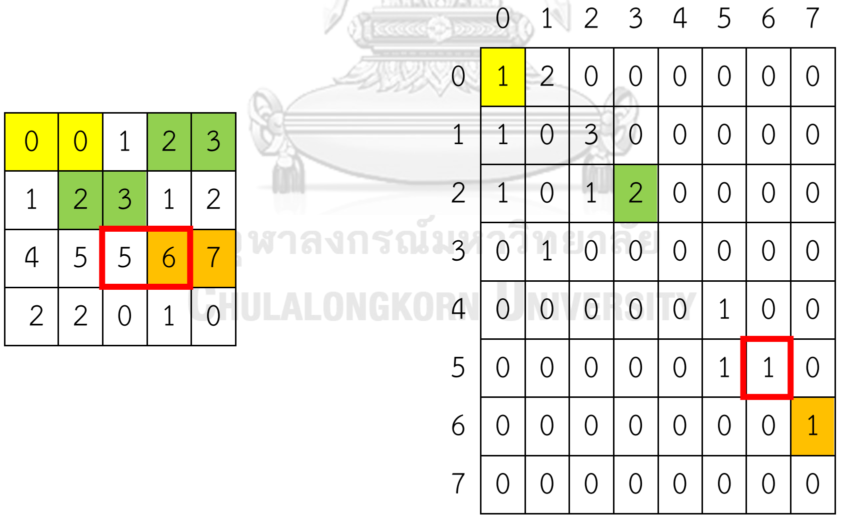


$(i-1, j-1)$	$(i-1, j)$	$(i-1, j+1)$
$(i, j-1)$	(i, j)	$(i, j+1)$
$(i+1, j-1)$	$(i+1, j)$	$(i+1, j+1)$

Figure 2.5. The 4 directions of GLCM

Figure 2.5 shows the directions of the neighbors of the considering pixel, (i, j). GLCM at 0 degree considers pixels (i, j) and $(i, j+1)$. GLCM at 45 degrees considers pixels (i, j) and $(i-1, j+1)$. GLCM at 90 degrees considers pixels (i, j) and $(i-1, j)$. GLCM at 135 degrees considers pixels (i, j) and $(i-1, j-1)$. Figure 2.6(a) shows a sample image of size 4x5 pixels with 8 gray scale levels. GLCM in 0-degree direction of a sample image is shown in Figure 2.6(b).

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY



0	0	1	2	3
1	2	3	1	2
4	5	5	6	7
2	2	0	1	0

	0	1	2	3	4	5	6	7
0	1	2	0	0	0	0	0	0
1	1	0	3	0	0	0	0	0
2	1	0	1	2	0	0	0	0
3	0	1	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0
5	0	0	0	0	0	1	1	0
6	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0

(a) (b)

Figure 2.6. Gray-level co-occurrence matrix

(a) a sample image

(b) GLCM of the image (a) with 8 gray scale levels in 0-degree direction

Contrast is the measurement of intensities between a pixel and its neighbor throughout an image which is known as variance or inertia. Zero contrast suggests a constant image. The equation used to calculate contrast is shown in Eq. (2):

$$\text{Contrast} = \sum_{i=1}^{M}\sum_{j=1}^{N}(i-j)^2\, p(i,j) \qquad (2)$$

where $p(i,j) = \frac{G(i,j)}{M \times (N-1)}$ is a normalized gray level at row i and column j of GLCM,

$G(i,j)$ is the element at row i and column j of GLCM,
N is the number of columns of GLCM, and
M is the number of rows of GLCM.

Energy is the sum of the square of normalized gray levels in GLCM. The energy range is between 0 and 1 where 1 represents a constant image. The formula of energy is described in Eq. (3):

$$\text{Energy} = \sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j)^2 \qquad (3)$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Homogeneity is the measurement of the distribution of closeness of pixels in GLCM to the diagonal of GLCM. The homogeneity range is between 0 and 1 where 1 shows a diagonal GLCM. The homogeneity equation is described in Eq. (4):

$$\text{Homogeneity} = \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)}{1+|i-j|} \qquad (4)$$

Correlation is the measurement of intensity correlation between a pixel and its neighbor throughout an image. The range of correlation is between -1 and 1 where

NaN (Not a Number) suggests a constant image. The formula of correlation is described in Eq. (5):

$$\text{Correlattion} = \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i\sigma_j} \qquad (5)$$

where μ_i is the mean of elements of row i of GLCM,

μ_j is the mean of elements of column j of GLCM,

σ_i is the standard derivation of elements of row i of GLCM, and

σ_j is the standard derivation of elements of column j of GLCM.

Entropy is the measurement of the quality of roughness. The characteristic of malignant tumor is not smooth so the entropy is higher than the benign tumor. The formula of entropy is described in Eq. (6):

$$\text{Entropy} = -\sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j)\log(p(i,j)) \qquad (6)$$

Histogram-based feature

The shape and properties of a histogram are one of the most important features that are used in image classification. Statistical data of intensities of an image can be extracted from a histogram. The probability of the intensities is shown in Eq. (7):

$$P(i) = \frac{H(i)}{N \times M} \qquad (7)$$

where $H(i)$ is the number of pixels with an intensity value i, $i \in \{0,1,2,\ldots,N_g\}$,

i is the intensity value,

N_g is a level of intensity on image,

N is the number of columns of an image, and

M is the number of rows of an image.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

There are four important features of a histogram, namely,

Mean (μ) is the average of intensities as shown in Eq. (8):

$$\mu = \sum_{i=1}^{N_g-1} iP(i) \tag{8}$$

Variance (σ^2) is the square of standard deviation of intensities. It is the change in intensities around the mean, which is calculated from Eq. (9):

$$\sigma^2 = \sum_{i=1}^{N_g-1} (i-\mu)^2 P(i) \tag{9}$$

Skewness is the value that represents the symmetry of a histogram. If a histogram is symmetric the skewness is 0. It can be calculated from Eq. (10):

$$\text{skewness} = \sigma^{-3} \sum_{i=1}^{N_g-1} (i-\mu)^3 P(i) \tag{10}$$

Kurtosis is the measurement from maximum to minimum value of histogram which is related to normal distribution. It can be calculate from Eq. (11):

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$\text{kurtosis} = \sigma^{-4} \sum_{i=1}^{N_g-1} (i-\mu)^4 P(i) \tag{11}$$

2.2.7. Feature selection

From previous section, we can see that there are several features that can be extracted from an image; however, some of them might not be relevant to the classification. Therefore, it is necessary to select the key features to be used for a certain classification and this process is called feature selection. Three basic techniques of feature selections that are used in our classification consist of forward selection [15], backward selection [15], and genetic selection [16].

2.2.8. Classification

The technique that is used in the proposed tumor classification is a supervised learning classifier that creates a model from training data [17]. There are many supervised learning classifiers such as Naïve Bayes [18], decision tree [19], neural network [20], linear regression [21], logistic regression [22], linear discriminant analysis [23], k-nearest neighbors [24], support vector machine [25], and other classifiers. All classifiers try to minimize errors of classification based on training data. However, it usually occurs that the errors increase when applying the model to unknown data. Hence it is necessary to have another set of known data to test the ability of the classification model. These data are called testing data. The classification rate can be improved by using two sets of data called training data and testing data as shown in Figure 2.7.

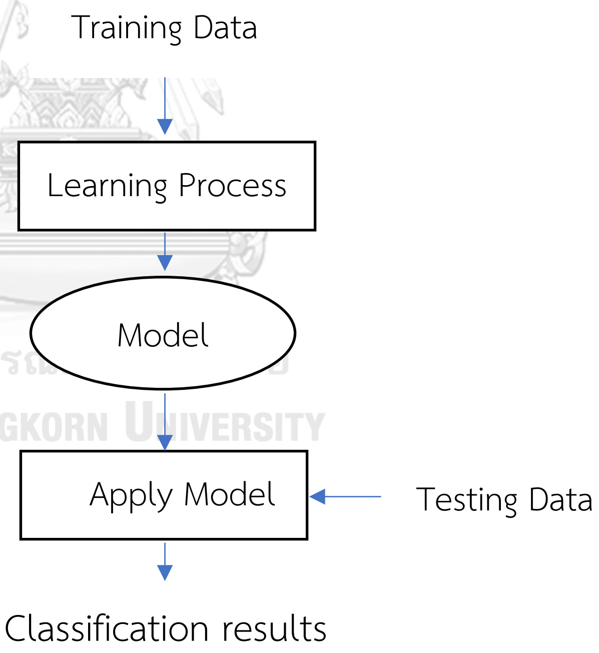


Figure 2.7. The process for creating a classification model

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

2.2.9. Performance measurement

A performance measurement is one of the most important steps for measuring the efficacy of a classification model. There are many tools used in performance measurement [26] as follows:

Confusion Matrix

Table 2.1. Confusion matrix

		Predicted Class	
		Yes	**No**
Actual Class	**Yes**	TP	FN
	No	FP	TN

True Positive (TP) represents the number of positive class and the prediction is correct.

False Positive (FP) represents the number of negative class and the prediction is incorrect.

True Negative (TN) represents the number of negative class and the prediction is correct.

False Negative (FN) represents the number of positive class and the prediction is incorrect.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Accuracy is the precision of classification that is the ratio of the number of correct prediction both Positive and Negative and the total number of data that are classified.

$$\text{Accuracy} = \frac{TP + FN}{TP + TN + FP + FN} \tag{12}$$

Precision is the correction of classification of positive class. It can be calculated by finding the ratio of correct prediction of data in positive class to the number of data in positive class.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

Sensitivity (Recall) is the capacity of accurate prediction which can be calculated by:

$$\text{Senitivity} = \frac{TP}{TP + FN} \tag{14}$$

Specificity is the capacity of test to accurately exclude the wrong prediction which can be calculated by:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{15}$$

$$\text{True Positive Rate } (TPR) = \frac{TP}{TP + FN} \tag{16}$$

$$\text{False Positive Rate } (FPR) = \frac{FP}{TN + FP} \tag{17}$$

$$\text{True Negative Rate } (TNR) = \frac{TN}{TN + FP} \tag{18}$$

$$\text{False Negative Rate } (FNR) = \frac{FN}{TP + FN} \tag{19}$$

where $FNR = 1 - TPR$ and $TNR = 1 - FPR$

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} \tag{20}$$

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ROC curve

ROC curve is a graph that represents the relation between TPR and FPR with adjusting parameters such as threshold value, cost matrix, and size of data [27]. An sample of ROC curve is shown in Figure 2.8.

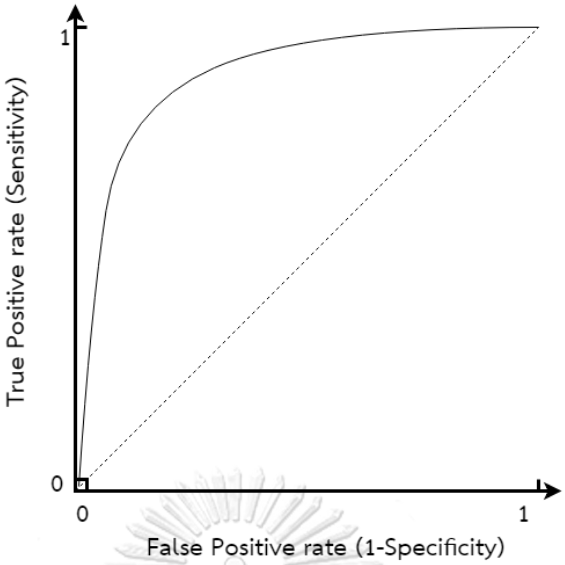


Figure 2.8. An example of ROC curve

The main positions of the ROC curve are:

$(0, 0)$ is the position when a classification model is always Negative. In this case, TPR and FPR are zero since no data were classified to be Positive.

$(1, 1)$ is the position when a classification model is always Positive. Hence, both TPR and FPR are equal to 1.

$(1, 0)$ is the ideal position. In this case, the classification model can predict data correctly.

The position on the diagonal line is in the case when TPR and FPR are equal which means that the correct prediction is nearly equal the incorrect prediction.

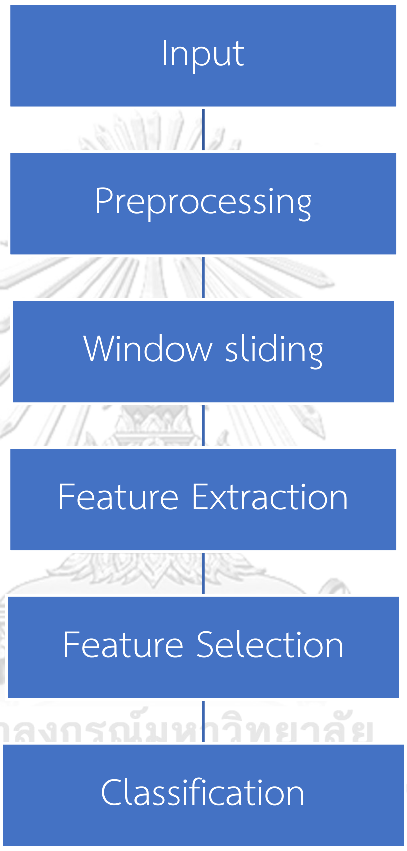
The position below diagonal line is in the case when TPR is less than FPR. On the other hand, the position above diagonal line is in the case when TPR is greater than FPR.

ROC curve shows the ability of classification model. The higher the ROC curve is above the diagonal line, the better the performance of classification model is.

CHAPTER 3

METHODOLOGY

In this chapter, the proposed method for feature extraction and tumor classification is shown in Figure 3.1. It is divided into six parts as follows:



จุฬาลงกรณ์มหาวิทยาลัย

Figure 3.1. The proposed method

3.1. Input data

The input data are video files (.mod) of the patients who have undergone the endobronchial ultrasonography. The video was recorded at the rate of 25 frames per second with frame size of 1080 x 1920 pixels. The original video is RGB color.

3.2. Preprocessing

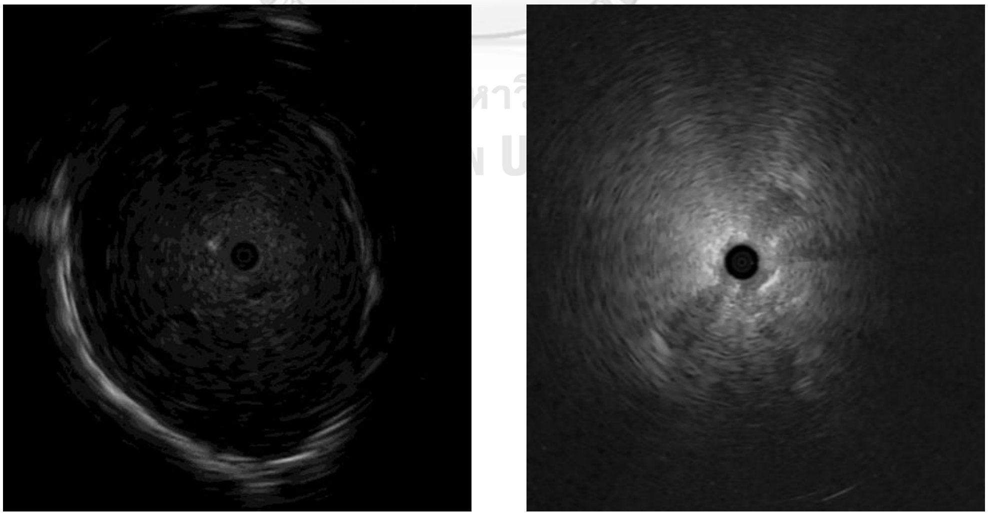
In the preprocessing step, the best frame is selected from each video file manually and the most appropriate region of interest is selected from each frame by performing the following steps:

3.2.1. File format conversion

Since the input video files are saved in .mod file format, thus the first step in preprocessing is to convert a video file format from .mod to .mp4 file format.

3.2.2. Frame selection

In this step, only one frame is selected from a video file by a doctor (a technician in radiology field). The criterion for selection is to look for the perfect tumor (no interference, visible boundary). In each video, there are approximately 25 – 250 frames depending on the length of each video. The length of the video is between 1 second and 10 seconds. Figure 3.2 shows an example of two frames that were selected from the same video file. Figure 3.2(a) is a good frame and Figure 3.2(b) is not a good frame since there are a lot of artifacts in the frame. Thus, only the best frame is selected to represent each video file for better classification result.



(a) a good frame selection (b) a bad frame selection

Figure 3.2. An example of frame selection

3.2.3. Boundary detection

The boundary of the tumor is identified by a doctor (or a technician in radiology field). Figure 3.3 shows the boundary of a tumor with the solid line which is supposed to be the closest one to the real boundary.

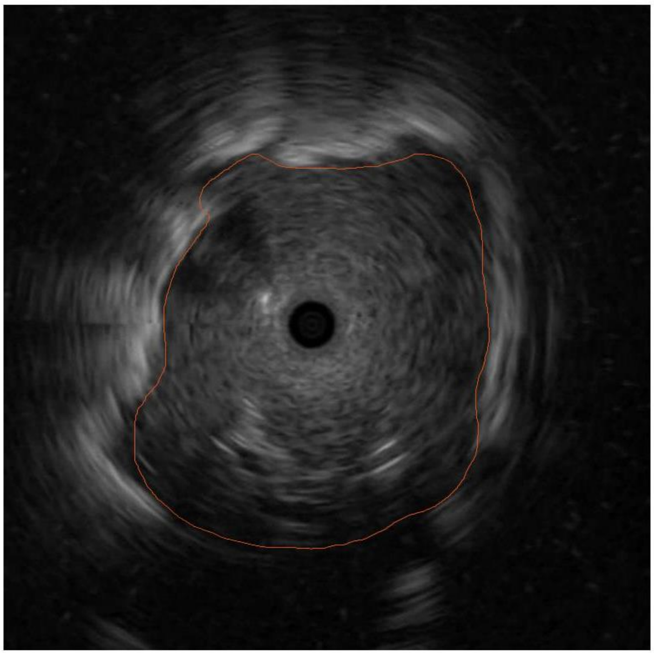


Figure 3.3. Drawing boundary of the tumor

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3.2.4. Region of interest selection

Since the boundary of a tumor (as shown in Figure 3.3) is not symmetric, we need to define the region of interest where the texture of a tumor will be analyzed by using the fact about the EBUS images from previous researches. Kurimoto found that the tumor within the radius less than 3 mm is near the probe of endobronchial ultrasound so the signal has the artifacts and this area is not good for texture analysis. On the other hand, the tumor that is outside the radius of 5 mm from the probe of endobronchial ultrasound has poor quality due to low signal. Thus, the region of interest is defined by the area of the ring between the circle with radius of 3 mm and the circle with radius of 5 mm [10] but some parts of the ring is outside the tumor boundary. Therefore, the region of interest is defined as the intersection between the tumor and the area inside the ring as shown in Figure 3.4.

3.3. Best window selection

In order to analyze the texture of a tumor, a doctor usually selects only a small region (called window or sub-region) within the region of interest that perfectly represents a tumor. This region is selected from the area with uniform texture. This window can be selected automatically according to the following steps:

1. Define a window of size 40 x 40 pixels which is the biggest window that can fit inside the region of interest.
2. Place the window at the upper left part of the ring as shown in Figure 3.4. Then the sum of intensities of the sub-region under the window is calculated and stored for later use.
3. Move the window one pixel to the right of region of interest and calculate the sum of intensities of the sub-region under the window.
4. Repeat step 3 until reaching the right boundary of region of interest. Then move the window down to the next row and to the left boundary of region of interest.
5. Repeat step 3 until reaching the bottom right boundary of region of interest.
6. Rank the sums of intensities of all the sub-regions in ascending order then select the median sub-region and use this sub-region as the best window for texture analysis.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

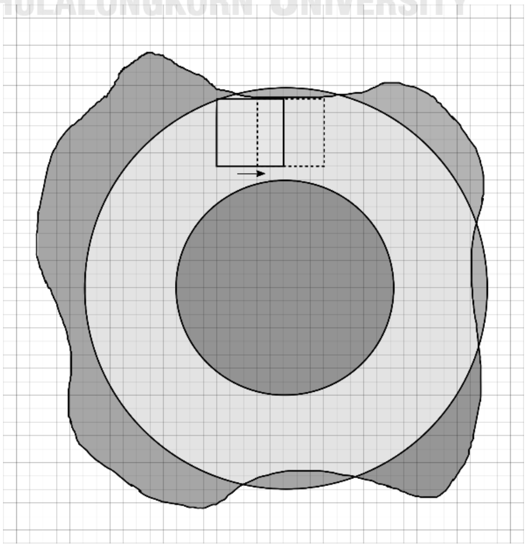


Figure 3.4. Window sliding in the region between the tumor and the ring

The median rank is calculated from

$$Med = \frac{N+1}{2} \tag{21}$$

where Med is the median rank,

N is the number of sub-regions.

The high intensity areas represent the white parts which indicate the air liners or the air dots in a tumor and the low intensity areas represent the black parts which indicate the blood vessels or the liquid in a tumor. But our interest is the areas with median intensity.

3.4. Feature extraction

Feature extraction is a process that tries to extract information from the data. Input data are the intensities of the best window and the output is a real number that represents a feature. The features are created to measure some conjectures depending on the objective of the study. In this work, many features are used including the proposed features, the existing features. Feature extraction is divided into 3 groups:

Group 1 is a group of standard features consists of 22 features, namely, mean, variance standard derivation, skewness, kurtosis, entropy, contrast with 0 degrees, contrast with 45 degree, contrast with 90 degree, contrast with 135 degree, correlation with 0 degree, correlation with 45 degree, correlation with 90 degree, correlation with 135 degree, energy with 0 degree, energy with 45 degree, energy with 90 degree, energy with 135 degree, homogeneity with 0 degree, homogeneity with 45 degree, homogeneity with 90 degree, and homogeneity with 135 degree.

Group 2 is a group of the proposed features called the upper and lower triangular gray level co-occurrence matrices consist of 12 features, namely, upper sum, lower sum and total sum with the 4-degree direction (0, 45, 90, and 135).

The upper and lower triangular gray level co-occurrence matrices are the modification of normal GLCM. These matrices are used to consider homogeneous and heterogeneous internal echoes of the pulmonary lesion. The structures of upper and lower triangular gray level co-occurrence matrices are shown in Figure 3.5(a)-(b).

Let $\mathbf{G}$ be a gray level co-occurrence matrix (GLCM) then the upper triangular GLCM is defined as

$$UG_{d,\theta,\delta}(i,j)=\begin{cases} G_{d,\theta}(i,j), & i\le j-\delta \\ 0, & i>j-\delta \end{cases} \tag{22}$$

and the lower triangular GLCM is defined as

$$UL_{d,\theta,\delta}(i,j)=\begin{cases} 0, & i< j+\delta \\ G_{d,\theta}(i,j), & i\ge j+\delta \end{cases} \tag{23}$$

where $G_{d,\theta}(i,j)$ is the element in GLCM at row i, column j with distance between two pixels d, and θ - degree direction, δ is the dissimilar factor or the difference between intensities of two adjacent pixels.

The reason that elements, $G_{d,\theta}(i,j)$, where $j-\delta<i<j+\delta$ along the main diagonal are not included in either of the triangular matrices are because heterogeneity means the quality of being dissimilar; therefore, the elements that represent homogeneity which are located near the main diagonal of a matrix, are not considered. The dissimilar factor can be specified depending on intensity levels and how heterogeneity is defined.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

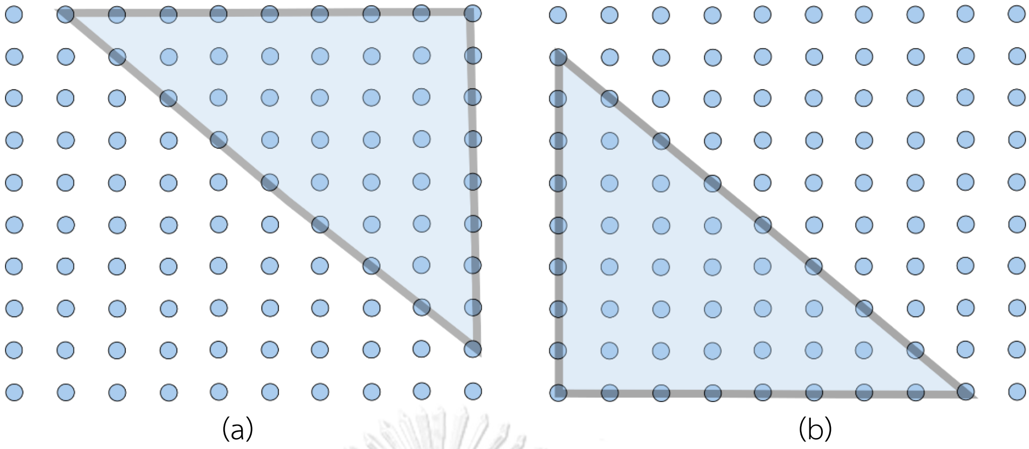

(a) (b)

Figure 3.5. An upper and lower triangular GLCMs

(a) An upper and (b) A lower triangular GLCMs.

The weight-sum of upper and lower GLCM

The upper and lower GLCM is formed by the union of the upper and the lower triangular GLCMs as shown in Figure 3.6.

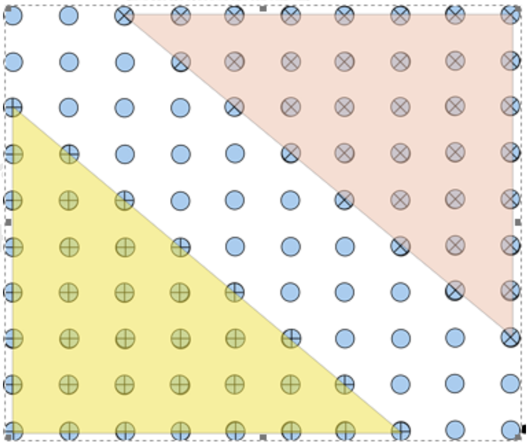


Figure 3.6. An upper and lower triangular GLCMs

As heterogeneity alludes to the difference of intensities between each pixel and its neighbors, hence the more difference between two pixels is, the higher chance of being heterogeneity it will be. In most of the cases, the contrast between intensities changes in a wide range; subsequently, the weight is characterized to relegate the level of differences. The more distinction between intensities of two pixels is, the more weight is assigned. Consequently, the new feature called the weight-sum of upper and lower GLCM, $S_{d,\theta,\delta}$, can be expressed as:

$$S_{d,\theta,\delta} = L_{d,\theta,\delta} + U_{d,\theta,\delta} \tag{24}$$

where $L_{d,\theta,\delta}$ is the weight-sum of lower GLCM which can be calculated by Eq. *(25):*

$$L_{d,\theta,\delta} = \sum_{i=1}^{N} \sum_{j=i+\delta}^{N} |i-j| LG_{d,\theta}(i,j) \tag{25}$$

and $U_{d,\theta,\delta}$ is the weight-sum of upper GLCM which can be calculated from Eq. *(26):*

$$U_{d,\theta,\delta} = \sum_{i=1+\delta}^{N} \sum_{j=1}^{i-\delta} |i-j| UG_{d,\theta}(i,j) \tag{26}$$

where d is the distance between two pixels, θ is the direction that the dissimilarity is determined, δ is the difference between intensities of two adjacent pixels, and N is the dimension of GLCM.

Group 3 is the combination of group 1 and group 2. The total features of group 3 are 34 features.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

3.5. Feature selection

In this work, three techniques are applied to select the most useful features and reduce some useless features. These techniques are forward selection, backward selection, and genetic selection. Since there are many features that are used in classification, thus if all of them are used in this process, it may cause low performance in classifying process. After applying feature selection, the best features are used in the classification process.

Forward selection

Forward selection is a technique that tries to add a new feature one at a time and select only important features. If the added feature improves the performance, then this feature is kept. If the added feature lowers the performance, then this feature is removed. In this experiment, the parameters for the forward selection are set to the values as shown in Table 3.1. The maximum number of attributes of each group to be selected through forward selection has the range between 1 and 22 for group 1, the range between 1 and 12 for group 2, and the range between 1 and 34 for group 3 these the maximal number of features is set to 34. Speculative round is set to equal to 0 and the stopping behavior is set to stop when the performance level is stable.

Table 3.1. The parameters of forward selection

Parameters	Argument setting
Maximal number of features	34
Speculative rounds	0
Stopping behavior	stable

Backward selection

Backward selection is a technique that tries to eliminate one feature at a time and select only important features. If eliminating a feature makes the performance better then this feature is removed. If eliminating a feature makes the performance worse, then this feature is kept. In this experiment, the parameters for backward selection are set to the values as shown in Table 3.2. The maximum number of elimination is equal to 22 for group 1, 12 for group 2, and 34 for group 3, thus the maximal number of features is set to 34. Speculative round which specifies the number of times is set to 0 and the stopping behavior is set to stop when the performance level decreases.

Table 3.2. The parameters of backward selection

Parameters	**Argument setting**
Maximal number of eliminations	34
Speculative rounds	0
Stopping behavior	decrease

Genetic selection

Genetic selection is a technique for finding solutions or the approximate solutions of a problem based on the theory of evolution from biology and natural selection. The principle of genetic algorithm for solving the optimal solution is to replace chromosomes with the existing solutions and then improve each individual solution in various ways that involve evolution, and random gene transformation with genetic operators (evolutionary operator) to get better solutions. In this experiment, the parameters for genetic algorithm are set according to trial and error. The Min number of features represents the minimum number of features that are used in the combinations of features is set to the default value which is 1. The population size that represents the number of individuals per generation is set to 50. The maximum number of generations is set to 500. The weights are normalized to be in the range from 0 to 1. The Maximal fitness is infinity. The selection scheme is set to a tournament. The tournament size is 0.25 with dynamic selection pressure. p is initialized to 0.5 with mutation equal to -1.0, and crossover equal to 0.5. Crossover type is set to uniform as shown in Table 3.3.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 3.3. The parameters of genetic selection

Parameters	Argument setting
Min number of features	1
Population size	50
Maximum number of generations	500
Normalize weights	yes
Maximal fitness	infinity
Selection scheme	tournament
Tournament size	0.25
Dynamic selection pressure	yes
p initial	0.5
p mutation	-1.0
p crossover	0.5
Crossover type	uniform

p initial = The initial probability for an attribute to be switched on is specified by this parameter.

p mutation =The probability for an attribute to be changed is specified by this parameter. If set to -1, the probability will be set to $1/n$ where n is the total number of attributes.

p crossover = The probability for an individual to be selected for crossover is specified by this parameter.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

3.6. Classification

Eight classifiers are used in the classification. These eight classifiers are Naïve Bayes, decision tree, neural network, linear regression, logistic regression, linear discriminant analysis, k-nearest neighbors, and support vector machine. The overall process is shown in Figure 3.7.

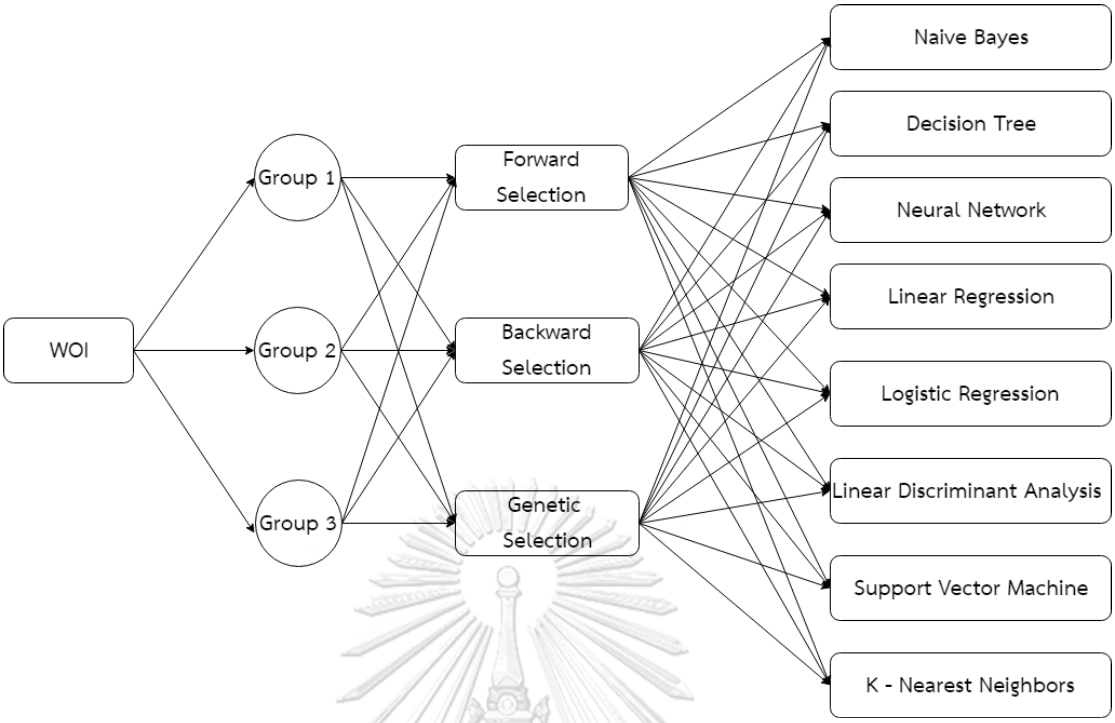


Figure 3.7. Diagram of the proposed process

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Naïve Bayes

Naïve Bayes classifier is a classification model that uses the probability principle based on Bayes' Theorem and the assumption that the occurrence of the events is independent. Naïve Bayes classifier has been used extensively because of its uncomplicated work, but It is effective. In this experiment, Laplace correction is used to improve the performance of Naïve Bayes classifier.

Decision tree

Decision tree is a classification model that is widely used in mathematics. Supervised learning can be used to construct a decision tree and interpret the results. The decision tree consists of internal nodes and leaves. Internal nodes represent the conjunction of features that are used in classification. Leaves represent classes of data. There are many algorithms for decision tree construction, namely, classification and regression trees (CART), induction of decision trees (ID3), C4.5. and chi-square automatic interaction detection (CHAID). In this experiment, CHAID is used and the parameters are set as shown in Table 3.4.

Table 3.4. The parameters of decision tree classifier

Parameters	Argument setting
Criterion	gain ratio
Maximal depth	20
Apply pruning	yes
Confidence	0.25
Apply prepruning	yes
Minimal gain	0.1
Minimal leaf size	2
Minimal size for split	4
Number of prepruning alternatives	3

Neural network

Neural network is a mathematical model or computer model for computational information processing. The concept of this technique is derived from the study of the bioelectric network in the brain, which consists of neurons and synapses. Neural network is constructed from connection between neurons until it is a collaborative network. In this experiment, the parameters are set as shown in Table 3.5.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 3.5. The parameters of neural network classifier

Parameters	Argument setting
Training cycles	500
Learning rate	0.3
Momentum	0.2
Shuffle	yes
Normalize	yes
Error epsilon	1.0E-5

Linear regression

Linear regression is a classification model that tries to fit a linear equation to the observed data. In this experiment, the parameters are set as shown in Table 3.6.

Table 3.6. The parameters of linear regression classifier

Parameters	Argument setting
Feature Selection	M5 prime
Eliminate colinear features	yes
Min tolerance	0.05
Use bias	yes
Ridge	1.0E-8

Logistic regression

Logistic regression is a statistical classifier for binary classification problems (two - class problems). Logistic regression is a simple and powerful linear classifier but the limitation of logistic regression is that the effectiveness of this classifier decreases when the data set is small, or when the data are well separated, or when a problem has more than two classes. In this experiment, the parameters are set as shown in Table 3.7.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 3.7. The parameters of logistic regression classifier

Parameters	Argument setting
Kernel type	dot
Kernel cache	200
C	1.0
Convergence epsilon	0.001
Max iterations	100,000
Scale	yes

Linear discriminant analysis

Linear discriminant analysis is a statistical classifier. The principle of this classifier is to use measurement function to classify unknown data. This classifier estimates linear coefficients that are associated with the variable of data. The linear discriminant analysis is very effective when the data set is in a linearly separating form. For argument setting, approximate covariance inverse is used in this experiment.

K-nearest neighbors

K-nearest neighbors are the clustering algorithms that use the principle of comparing similarity of the observed data with other data. The observed data are set to a class that is the nearest. K-nearest neighbor algorithm is very simple and easy to understand. The k-nearest neighbor algorithm is summarized as follows:

1. Define the size of K. In this experiment, K is set to 1.
2. Calculate the distance between the observed data and sample data by using Euclidean distance.
3. Order the distance and choose the sample that in the closest to the observed data according to the defined K.
4. Consider K classes of data and observe the class that in nearest to the observed data.
5. Determine an appropriate class for the observed data.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Support vector machine

SVM is a classifier that is widely used for digital image processing. The principle of SVM is to train input data into vector in n-dimensional space. In two-dimensional and three-dimensional spaces, the points are in the XY plane and XYZ space, respectively. Then, a hyperplane that separates the input data vector into different classes is created. In the case of two-dimensional and three-dimensional spaces, hyperplanes are straight lines and planes, respectively. The SVM's dominant feature is to store a vector in the input space into the feature space by using kernel. There are many kernels, namely, dot, radial, polynomial, neural, and other kernels. In this experiment, the neural kernel is used and the parameters are set as shown in Table 3.8.

Table 3.8. The parameters of support vector machine classifier

Parameters	**Argument setting**
Kernel type	neural
Kernel a	1.0
Kernel b	0.0
Kernel cache	200
C	0
Convergence epsilon	0.001
L positive	1.0
L negative	1.0
Epsilon	0.0
Epsilon plus	0.0
Epsilon minus	0.0

3.7. The proposed method

In order to classify whether a tumor is benign or malignant, the following process is performed:

1. Convert the original video file from .mod format to .mp4 format. The dimensions of each frame are 1920x1080 pixels with 25 frames per second.
2. Select the best frame from a video file based on the completeness of a tumor and low noise. The best frame is selected by a doctor.
3. Remove metadata from an image. These metadata represent the details of a patient such as a hospital number, first name, age, gender, date and time of recording, the frequency range of ultrasound, zoom distance and other as shown in figure 3.8. Figure 3.9 shows a frame after removing metadata which is now ready to be uses in the next step.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

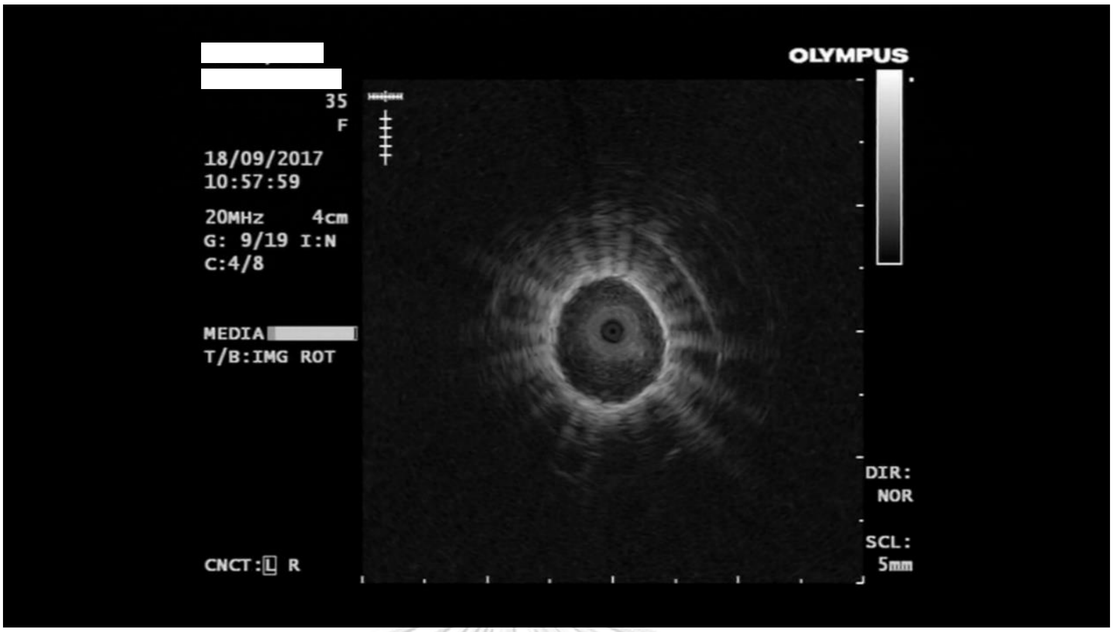


Figure 3.8. An example of frame with Metadata

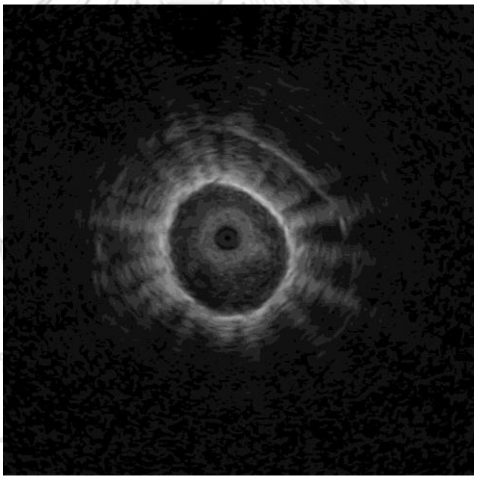


Figure 3.9. The frame after removing metadata

4. Convert the original RGB color image to a gray scale image before applying feature extraction.
5. The boundary of a tumor is identified by a doctor as shown in Figure 3.10.

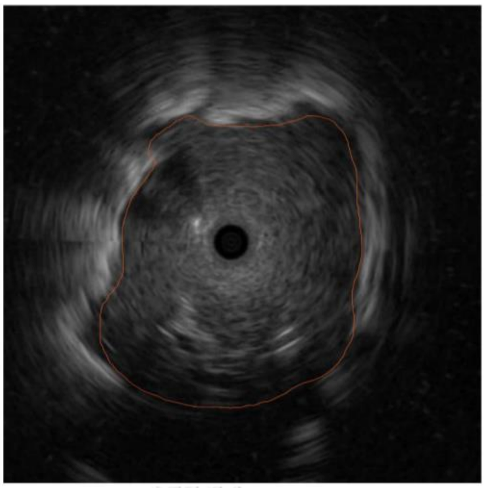


Figure 3.10. Determining boundary of the tumor

6. Find the intersection area between the ring area and the tumor based on Kurimoto’s research [10].

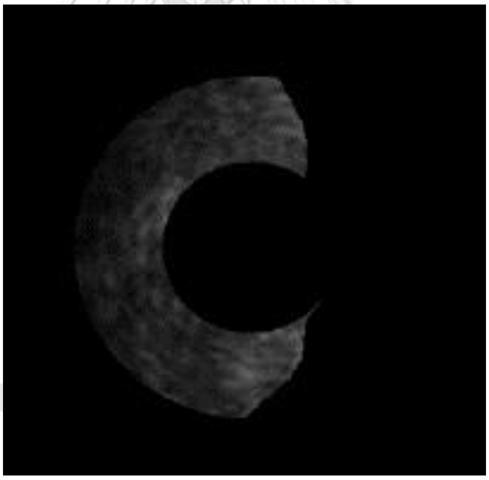


Figure 3.11. The intersection area between the tumor and the ring area

7. Define a window of interest (WOI) of size 40×40 pixels and use it to select the best window of interest whose sum of intensities is the median in the rank. The WOI is slid from top to bottom and from left to right as shown in Figure 3.12.

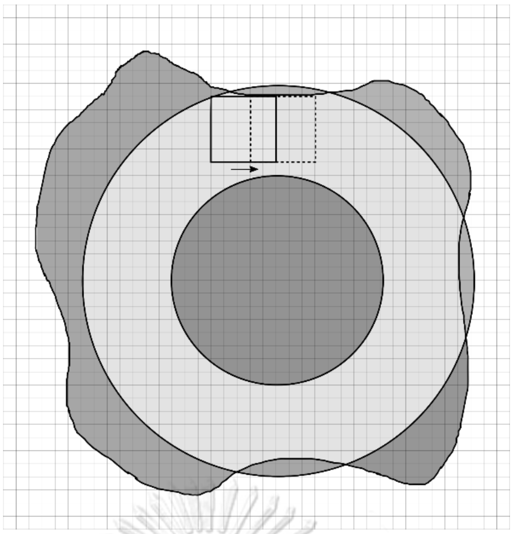


Figure 3.12. Window slicing in the region between tumor and ring

8. Extract the features from WOI and divide the features into three groups, namely, group 1: standard features, group 2: the proposed features, and group 3: mixed features between group 1 and group 2. Figure 3.13 shows the example of WOI that is inside the intersection area of tumor and boundary.

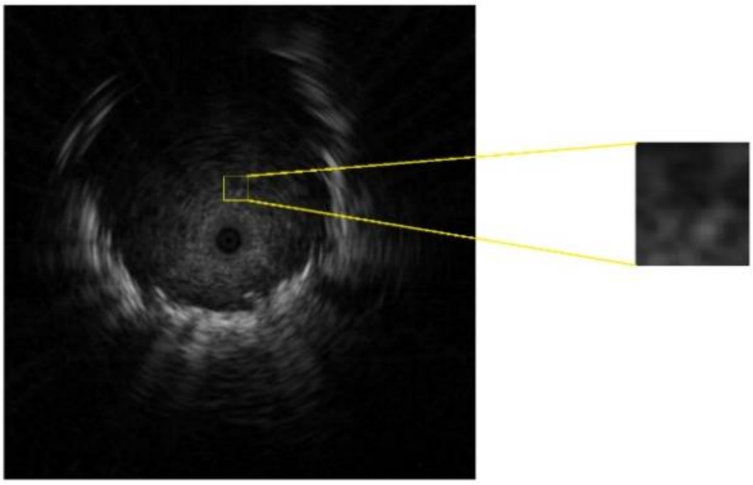


Figure 3.13. The intersection area between the tumor and the ring area

9. Use the existing feature selection methods to choose the efficient features needed for creating the prediction model as shown in figure 3.13.
10. Compare the results obtained from three groups and find which methods give the most accurate results.

CHAPTER 4

EXPERIMENTAL RESULTS

In this chapter, we describe the experiments and the results. The results are divided into 4 parts. First part shows the classification results of using standard features with genetic selection and eight classifiers, the second part shows the classification results of using the proposed features with genetic selection and eight classifiers, the third part shows the classification results of using both standard features and the proposed features with genetic selection and eight classifiers, and the last part shows the comparison of classification results using confusion matrix. Each part of the classification results shows the results of the accuracy, sensitivity, and specificity from eight classifiers and three feature selections.

The input data are video files (.mod) of the patients who had undergone the endobronchial ultrasonography from May 2015 to May 2016 at Phramongkutklao Hospital, Bangkok, Thailand. There are 34 files with benign and 55 files with malignant and the ratio between benign and malignant is shown in Figure 4.1. The dimensions of video are of 1080 x 1920 pixels with 25 frames per second.

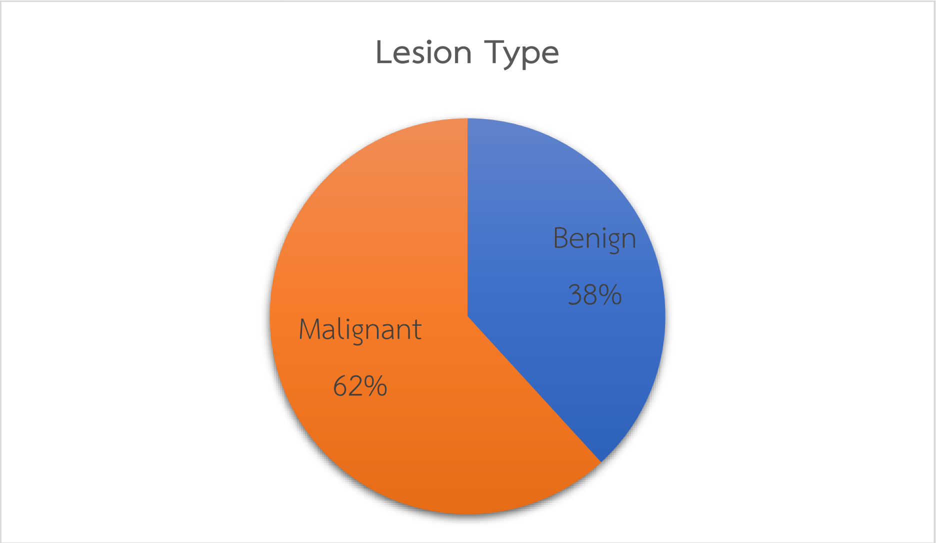


Figure 4.1. The ratio of input data

4.1. The classification results of using standard features with genetic selection and the eight classifiers.

The standard features in group 1 as mentioned in Chapter 3 consist of 22 features, namely, mean, variance, standard derivation, skewness, kurtosis, entropy, contrast with 0 degree, contrast with 45 degrees, contrast with 90 degrees, contrast with 135 degrees, correlation with 0 degree, correlation with 45 degrees, correlation with 90 degrees, correlation with 135 degrees, energy with 0 degree, energy with 45 degrees, energy with 90 degrees, energy with 135 degrees, homogeneity with 0 degree, homogeneity with 45 degrees, homogeneity with 90 degrees, and homogeneity with 135 degrees. All features in group 1 are selected by using the combinations of three feature selections and eight classifiers. After the classification is complete, the eight classification results are shown in Table 4.1. The experimental results reveal that among three feature selections, the genetic selection outperforms the other two feature selections. Hence the eight results in Table 4.1 show the combination of genetic selection and eight classifiers. The accuracies from using eight classifiers in descending order are: support vector machine (78.652 %), k-nearest neighbors (78.652%), decision tree (78.652%), neural network (77.528%), linear regression (74.157%), logistic regression (74.157%), linear discriminant analysis (73.034%), and Naïve Bayes (70.787%). The sensitivities from using eight classifiers in descending order are: logistic regression (98.182%), neural network (96.364%), decision tree (94.545%), linear regression (94.545%), k-nearest neighbors (94.545%), linear discriminant analysis (90.909%), Naïve Bayes (85.455%), and support vector machine (85.455%). The specificities from using eight classifiers in descending order are: support vector machine (67.647%), decision tree (52.941%), k-nearest neighbors (52.941%), Naïve Bayes (47.059%), neural network (47.059%), linear discriminant analysis (44.118%), linear regression (41.176%), and logistic regression (35.294%).

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 4.1. The classification results of using standard features with the eight classifiers.

Method	Accuracy	Sensitivity	Specificity
Support vector machine	**78.652**	85.455	**67.647**
K-nearest neighbors	**78.652**	94.545	52.941
Decision tree	**78.652**	94.545	52.941
Neural network	77.528	96.364	47.059
Linear regression	74.157	94.545	41.176
Logistic regression	74.157	**98.182**	35.294
Linear discriminant analysis	73.034	90.909	44.118
Naïve Bayes	70.787	85.455	47.059

4.2. The classification results of using the proposed features with genetic selection and the eight classifiers.

In this experiment, the proposed features that are used for the classification are the features from the proposed Upper and Lower Triangular Gray Level Co-occurrence Matrices. There are 12 features, namely, the upper sum, the lower sum and the total sum with the 4-degree direction (0, 45, 90, and 135). All features were selected by using the combinations of three feature selections and eight classifiers. After the classification is complete, the eight classification results are shown in Table 4.2. The experimental results reveal that among three feature selections, the genetic selection outperforms the other two feature selections. Hence the eight results in Table 4.2 show the combination of genetic selection and eight classifiers. The accuracies from using eight classifiers in descending order are: support vector machine (80.899 %), neural network (78.652%), k-nearest neighbors (76.404%), decision tree (76.404%), linear regression (73.034%), Naïve Bayes (73.034 %), linear discriminant analysis (71.910%), and logistic regression (70.787%). The sensitivities from using eight classifiers in descending order are: logistic regression (100.00%), decision tree (98.182%), linear regression (96.364%), k-nearest neighbors (94.545%), linear discriminant analysis

(89.091%), support vector machine (87.273%), Naïve Bayes (85.455%), and neural network (85.455%). The specificities from using eight classifiers in descending order are: support vector machine (70.588%), neural network (67.647%), Naïve Bayes (52.941%), k-nearest neighbors (47.059%), linear discriminant analysis (44.118%), decision tree (41.176%), linear regression (35.294%), and logistic regression (23.529%).

Table 4.2. The classification results of using the proposed features with the eight classifiers.

Method	Accuracy	Sensitivity	Specificity
Support vector machine	**80.899**	87.273	**70.588**
Neural network	78.652	85.455	67.647
K-nearest neighbors	76.404	94.545	47.059
Decision tree	76.404	98.182	41.176
Linear regression	73.034	96.364	35.294
Naïve Bayes	73.034	85.455	52.941
Linear discriminant analysis	71.910	89.091	44.118
Logistic regression	70.787	**100.00**	23.529

4.3. The classification results of using both standard features and the proposed features with genetic selection and the eight classifiers.

In this experiment, the combination of the proposed features and the standard features were used. The total features are 34 features. All of the features were selected by using the combinations of three feature selections with eight classifiers. Hence the eight results in Table 4.3 show the combination of genetic selection and eight classifiers. The experimental results reveal that genetic selection is the best feature selection among all three feature selections. Table 4.3 shows the results from using genetic selection with eight classifiers. The accuracies from using eight classifiers in descending order are: support vector machine (86.517%), neural network (78.652%), linear discriminant analysis (75.281%), k-nearest neighbors (74.157%), decision tree

(74.157%), linear regression (73.034%), logistic regression (71.910%), and Naïve Bayes (70.787%). The sensitivities from using eight classifiers in descending order are: logistic regression (98.182%), k-nearest neighbors (96.364%, decision tree (92.727%), linear discriminant analysis (90.909%), linear regression (90.909%), support vector machine (87.273%), Naïve Bayes (85.455%), and neural network (83.636%). The specificities from using eight classifiers in descending order are: support vector machine (85.294%), neural nets (70.588%), linear discriminant analysis (50.000%), linear regression (44.118%), k-nearest neighbors (38.235%), Naïve Bayes (47.059%), decision tree (44.118%), and logistic regression (29.412%).

Table 4.3. The classification results of using both standard features and the proposed features with the eight classifiers.

Method	Accuracy	Sensitivity	Specificity
Support vector machine	**86.517**	87.273	**85.294**
Neural nets	78.652	83.636	70.588
Linear discriminant analysis	75.281	90.909	50.000
K-nearest neighbors	74.157	96.364	38.235
Decision tree	74.157	92.727	44.118
Linear regression	73.034	90.909	44.118
Logistic regression	71.910	**98.182**	29.412
Naïve Bayes	70.787	85.455	47.059

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4.4. The comparison of classification results using confusion matrix

The confusion matrix in Table 4.4 shows the accuracy of classification results of using standard features with genetic selection and support vector machine when comparing with the actual data. The classification results reveal that eight malignant tumors were misclassified as benign tumors, whereas, 11 benign tumors were misclassified as malignant tumors. Table 4.5 shows the confusion matrix of the accuracies from using the weight-sum upper and lower GLCM features with genetic selection and support vector machine. The results reveal that seven malignant tumors were misclassified as benign tumors, whereas, 10 benign tumors were misclassified as malignant tumors.

Finally, the confusion matrix in Table 4.6 shows the classification accuracies from using the combination of standard features and the weight-sum upper and lower GLCM features with genetic selection and support vector. The results reveal that the use of the weight-sum upper and lower GLCM and the standard features are much better than the other two groups because only five benign tumors were misclassified as malignant tumors and seven malignant tumors were misclassified as benign tumors.

Table 4.4. The confusion matrix of classification results from using standard features with genetic selection and support vector machine

Classified as	Correct class	
	Malignant	Benign
Malignant	47	11
Benign	8	23

Table 4.5. The confusion matrix of classification results from using the proposed features with genetic selection and support vector machine

Classified as	Correct class	
	Malignant	Benign
Malignant	48	10
Benign	7	24

Table 4.6. The confusion matrix of classification results from using the proposed feature and standard features with genetic selection with support vector machine

Classified as	Correct class	
	Malignant	Benign
Malignant	48	5
Benign	7	29

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

CHAPTER 5

DISCUSSION AND CONCLUSION

The weight-sum of upper and lower GLCM features are the proposed features which can be used to measure homogeneity and heterogeneity of internal echoes of an image. The principles of the weight-sum of upper and lower GLCM features are the modification of GLCM which records the frequency of a pair of intensities of neighbor pixels with a specific distance according to the considering degrees of direction by adding the weight to that frequency. The weight of each frequency depends on the difference between the intensities of the two pixels. For example, the weight of the frequency of a pair of intensity with values 0 and 1 have the weight less than the frequency of a pair of intensity with values 0 and 15.

The group of features are divided into three groups: group 1 consists of standard feature, group 2 consists of the proposed features (weight-sum of upper and lower GLCM feature), and group 3 consists of the combination of group 1 and group 2. In this study, three feature selections, namely, forward selection, backward selection, and genetic selection are used as feature selections, and eight classifiers, namely, Naïve Bayes, decision tree, neural network, linear regression, logistic regression, linear discriminant analysis, K-nearest neighbors, and support vector machine are used as classifiers. The combination of feature selections and classifiers are applied to three groups of features to find the best features and the best combination of the techniques that yield the most accurate classification results.

Table 5.1 shows the results from applying genetic selection with support vector machine to three groups of features. The results reveal that the features in group 3 yield the highest accuracy, the highest sensitivity, and the highest specificity. Thus, the proposed features can be used to improve the performance of classification.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 5.1. The classification results of three group feature extraction with genetic selection with support vector machine classifier

Feature Extraction Group	**Accuracy**	**Sensitivity**	**Specificity**
Group 1	78.652	85.455	67.647
Group 2	80.899	**87.273**	70.588
Group 3	**86.517**	**87.273**	**85.294**

Moreover, the performance of the proposed features and the classification can be shown by the ROC curve in Figure 5.1. It can be seen that the largest area under the curve is the results from using features in group 3. The areas under the curves of features in group 1, group 2 and group 3 are shown in Figure 5.1, respectively.

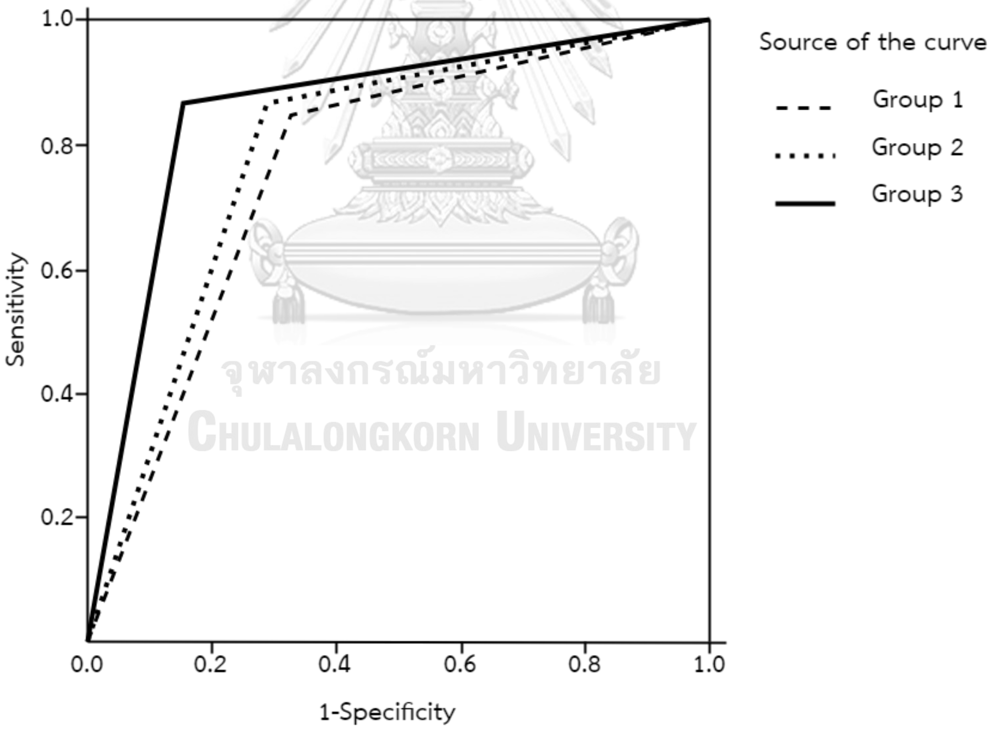


Figure 5.1. ROC curve

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 5.2. Area under the curve

Feature Extraction Group	**Area under the curve**
Group 1	0.766
Group 2	0.789
Group 3	**0.863**

In summary, we propose the new features and the classification method that can classify pulmonary lesion with an acceptable accuracy rate. As a result, the proposed features, called weight-sum of upper and lower GLCM features, together with genetic selection and support vector machine could help the doctors to classify pulmonary lesion in an EBUS image.

There are some limitations in this study that could be covered in the future work. These limitations are as follows: First, in the frame selection, the automatic frame selection can be applied. Second, in the boundary detection, the automatic boundary detection can be developed to perform the task without human interaction. And finally, in the region of interest selection, the squared area of 40×40 pixels can be extended to the whole ring area in order to include more features in the classification. These extra features are the features that represent other characteristics of echoes of malignant; such as linear air, dot, and continuous boundary.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

REFERENCES

1. Wee, R.Y. *Leading Causes Of Death In Thailand*. 2017 April 25, 2017 [cited 2018 8 May]; Available from: https://www.worldatlas.com/articles/leading-causes-of-death-in-thailand.html.
2. International Association for the Study o Lung Cancer. *LUNG CANCER FACTS AND STATISTICS*. 2017 [cited 2018 23 Apr]; Available from: http://wclc2017.iaslc.org/wp-content/uploads/2017/09/2017-WCLC-Fact-Sheet-Lung-Cancer-Final.pdf.
3. Kurimoto, N., et al., *Analysis of the internal structure of peripheral pulmonary lesions using endobronchial ultrasonography.* Chest, 2002. **122**(6): p. 1887-1894.
4. Kuo, C.-H., et al., *Diagnosis of peripheral lung cancer with three echoic features via endobronchial ultrasound.* Chest, 2007. **132**(3): p. 922-929.
5. BAI, X.-b., K.-q. WANG, and H. Wang, *Research on the classification of wood texture based on Gray Level Co-occurrence Matrix [J].* Journal of Harbin Institute of Technology, 2005. **12**: p. 021.
6. Wu, C.-M., Y.-C. Chen, and K.-S. Hsieh, *Texture features for classification of ultrasonic liver images.* IEEE Transactions on medical imaging, 1992. **11**(2): p. 141-152.
7. Mahmoud, M.K.A., et al. *Classification of malignant melanoma and benign nevi from skin lesions based on support vector machine*. in *Computational Intelligence, Modelling and Simulation (CIMSim), 2013 Fifth International Conference on*. 2013. IEEE.
8. Chao, T.-Y., et al., *Differentiating peripheral pulmonary lesions based on images of endobronchial ultrasonography.* Chest, 2006. **130**(4): p. 1191-1197.
9. Lie, C.-H., et al., *New image characteristics in endobronchial ultrasonography for differentiating peripheral pulmonary lesions.* Ultrasound in medicine and biology, 2009. **35**(3): p. 376-381.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

10. Morikawa, K., et al., *Histogram-based quantitative evaluation of endobronchial ultrasonography images of peripheral pulmonary lesion.* Respiration, 2015. **89**(2): p. 148-154.

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

12. Abdel-Nasser, M., et al., *Breast tumor classification in ultrasound images using texture analysis and super-resolution methods.* Engineering Applications of Artificial Intelligence, 2017. **59**: p. 84-92.

13. Walker, R.F., P.T. Jackway, and D. Longstaff, *Genetic algorithm optimization of adaptive multi-scale GLCM features.* International Journal of Pattern Recognition and Artificial Intelligence, 2003. **17**(01): p. 17-39.

14. Zainudin, F.L., et al. *Comparison between GLCM and modified Zernike moments for material surfaces identification from photo images*. in *Computational Science and Technology (ICCST), 2014 International Conference on*. 2014. IEEE.

15. Pudil, P., J. Novovičová, and J. Kittler, *Floating search methods in feature selection.* Pattern recognition letters, 1994. **15**(11): p. 1119-1125.

16. Holland, J. and D. Goldberg, *Genetic algorithms in search, optimization and machine learning.* Massachusetts: Addison-Wesley, 1989.

17. Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*. 2011: Elsevier.

18. Murphy, K.P., *Naive bayes classifiers.* University of British Columbia, 2006. **18**.

19. Safavian, S.R. and D. Landgrebe, *A survey of decision tree classifier methodology.* IEEE transactions on systems, man, and cybernetics, 1991. **21**(3): p. 660-674.

20. Lippmann, R., *An introduction to computing with neural nets.* IEEE Assp magazine, 1987. **4**(2): p. 4-22.

21. Neter, J., et al., *Applied linear statistical models*. Vol. 4. 1996: Irwin Chicago.

22. Ng, A.Y. and M.I. Jordan. *On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes*. in *Advances in neural information processing systems*. 2002.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

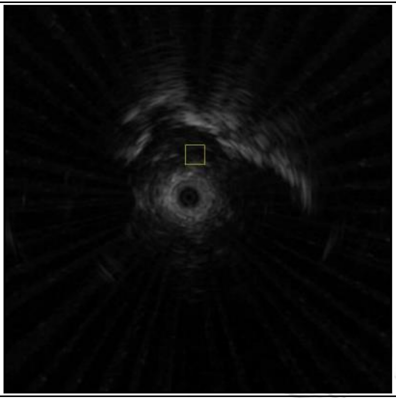
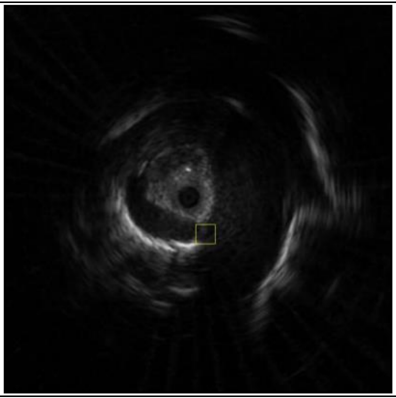
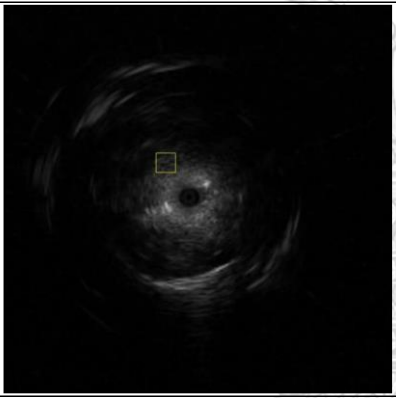
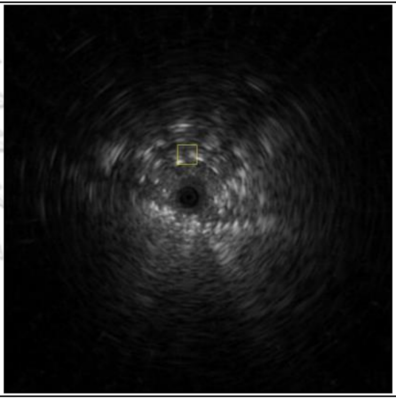
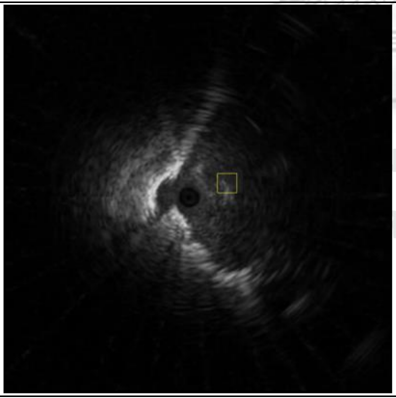
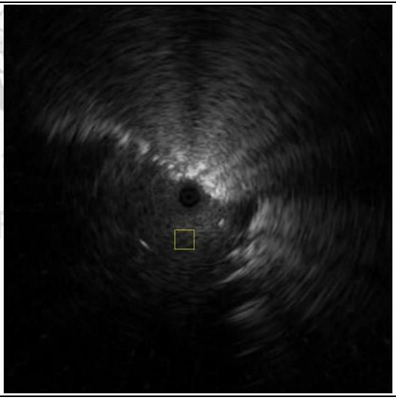
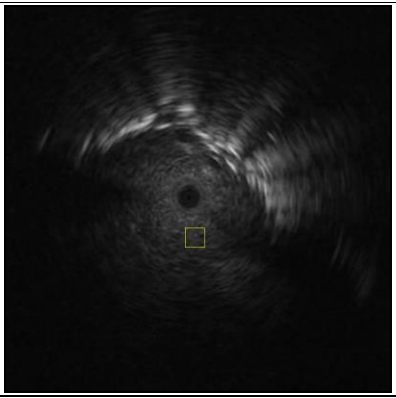
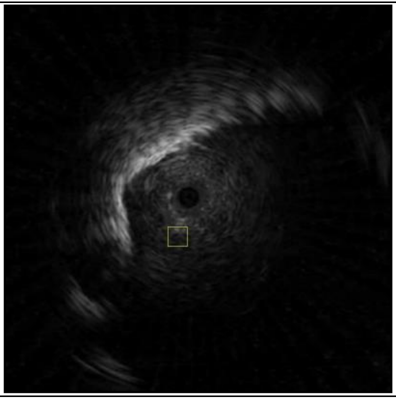
23. Xanthopoulos, P., P.M. Pardalos, and T.B. Trafalis, *Linear discriminant analysis*, in *Robust data mining*. 2013, Springer. p. 27-33.
24. Kramer, O., *K-nearest neighbors*, in *Dimensionality Reduction with Unsupervised Nearest Neighbors*. 2013, Springer. p. 13-23.
25. Scholkopf, B. and A.J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. 2001: MIT press.
26. Sokolova, M. and G. Lapalme, *A systematic analysis of performance measures for classification tasks.* Information Processing & Management, 2009. **45**(4): p. 427-437.
27. Bradley, A.P., *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern recognition, 1997. **30**(7): p. 1145-1159.

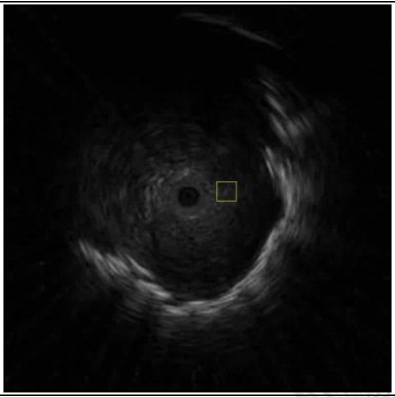
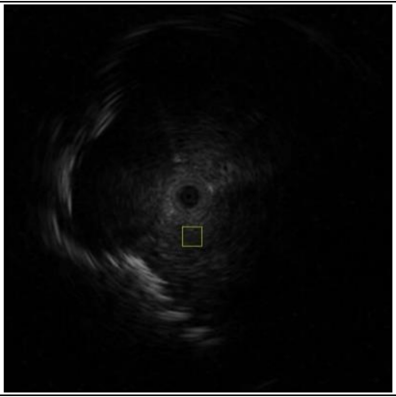
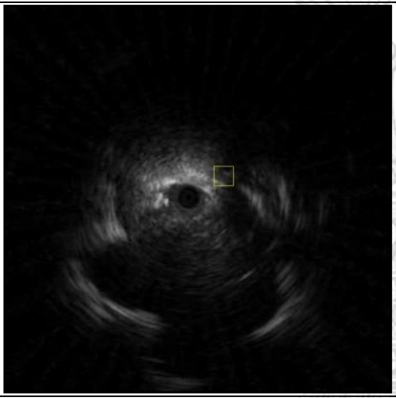
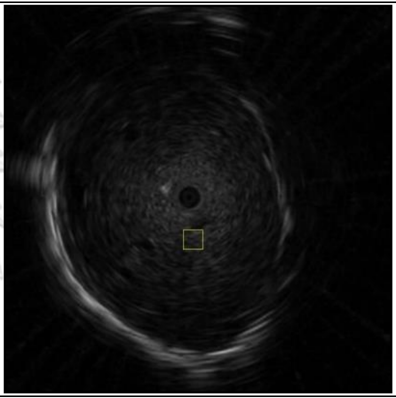
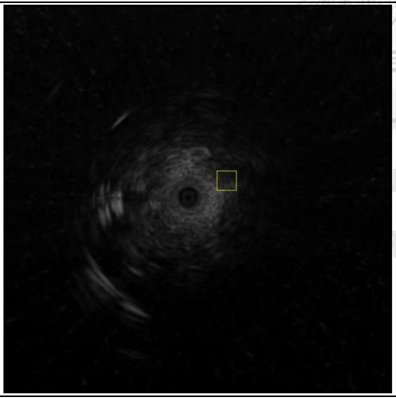
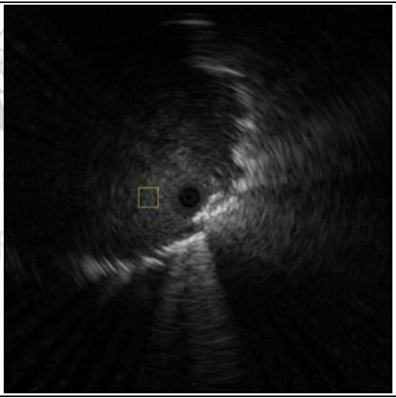
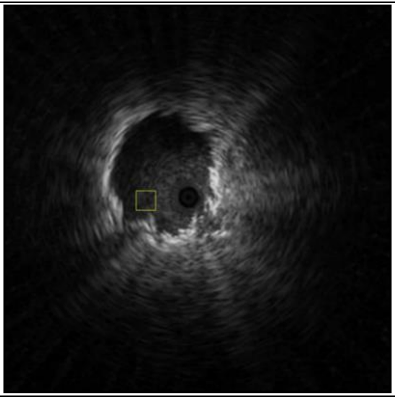
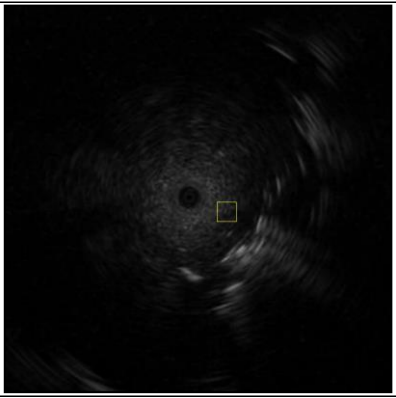
จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

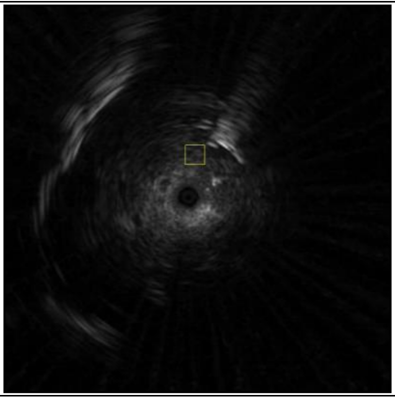
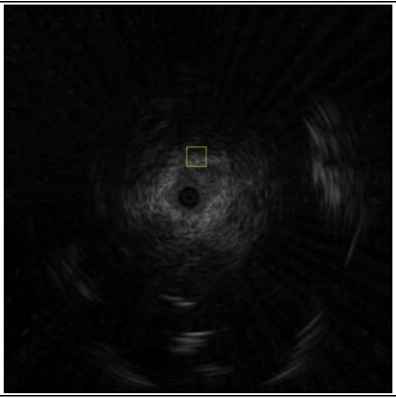
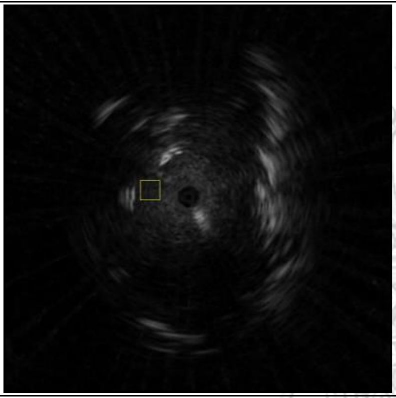
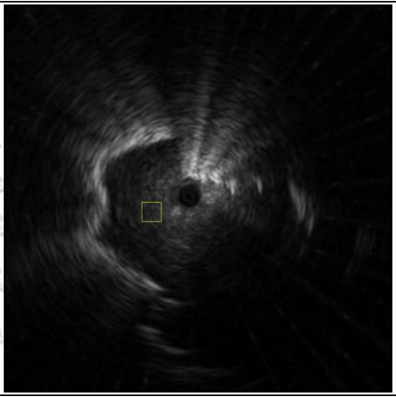
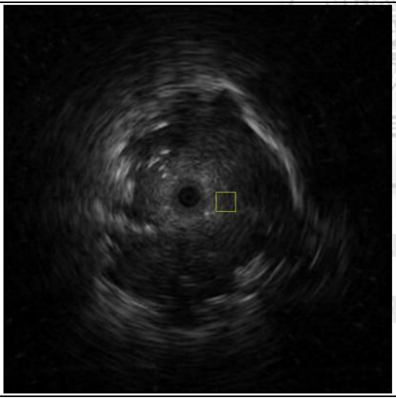
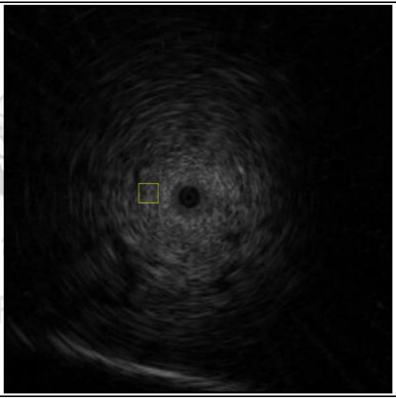
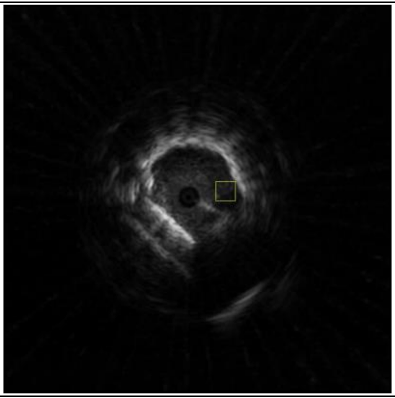
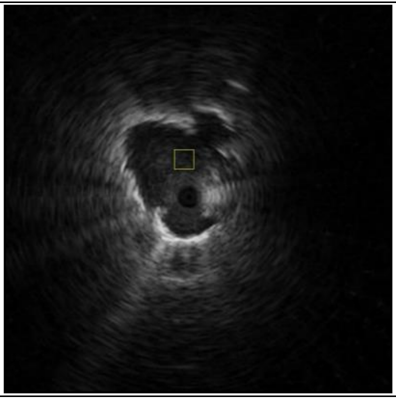
APPENDIX

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

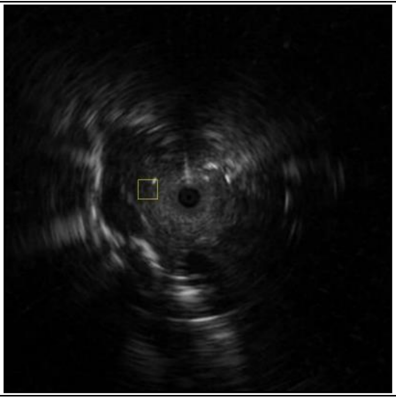
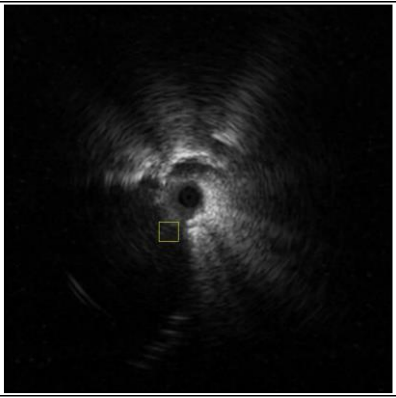
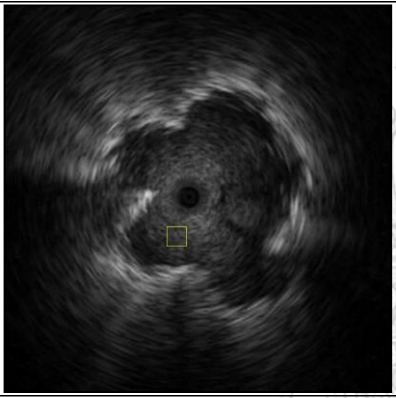
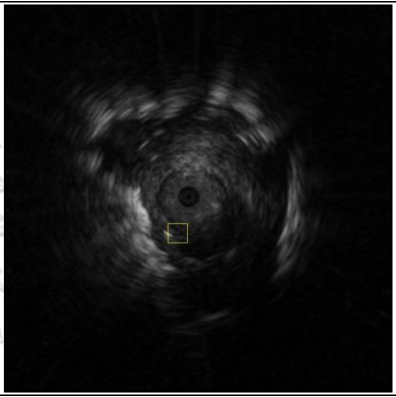
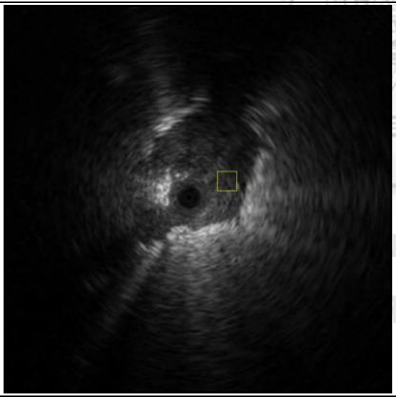
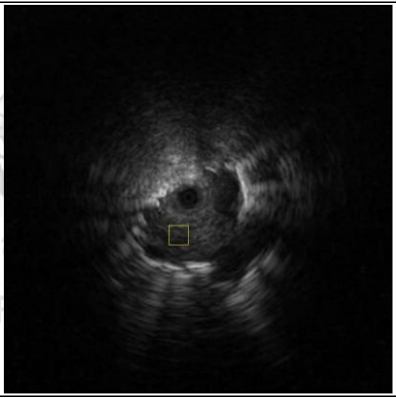
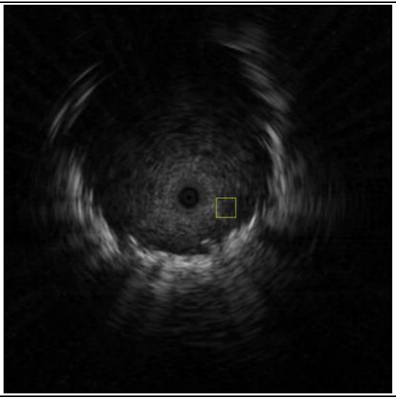
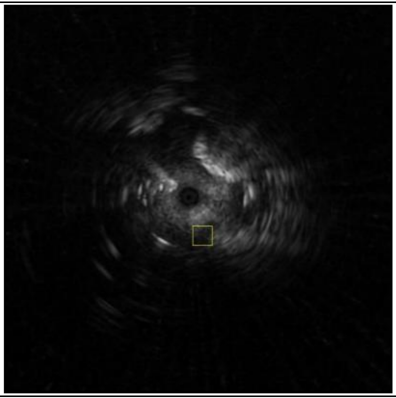
APPENDIX A Best Frame with window of interest

File	Best Frame	File	Best Frame
1		5	
2		6	
3		7	
4		8	

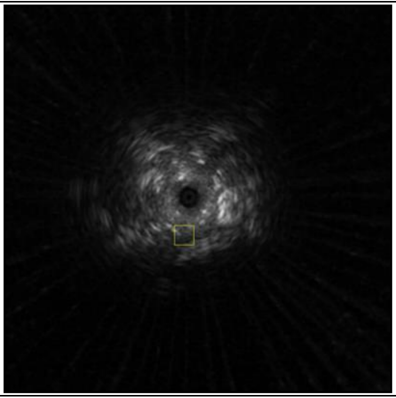
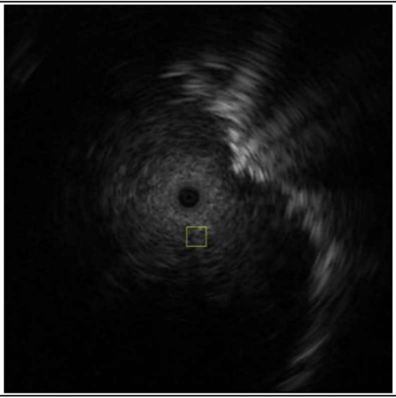
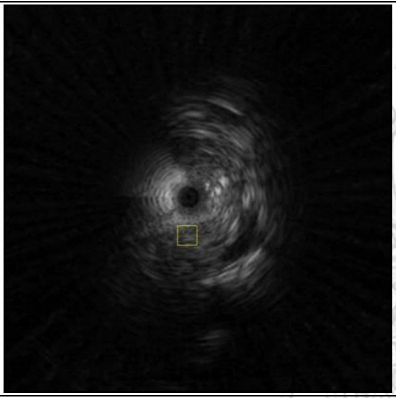
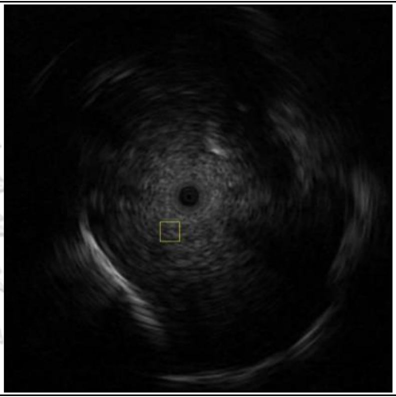
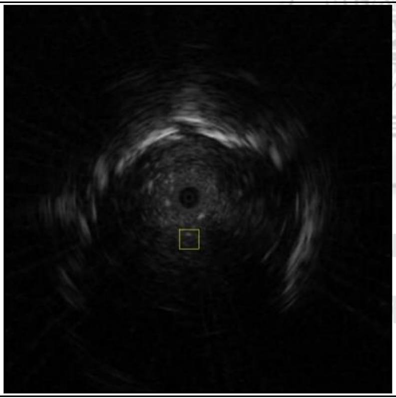
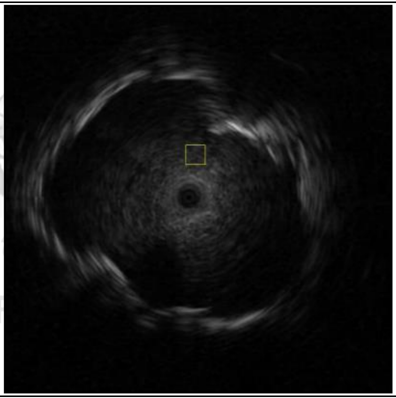
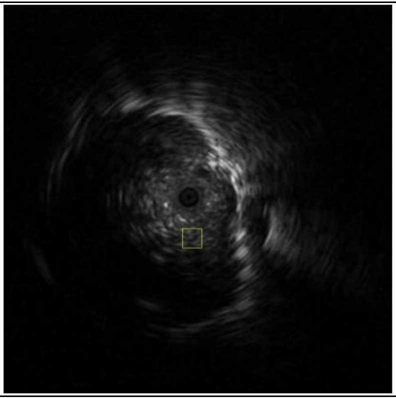
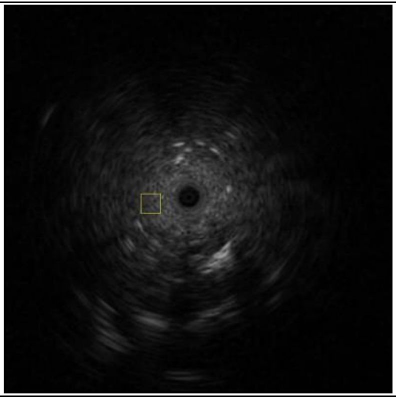
File	Best Frame	File	Best Frame
9		13	
10		14	
11		15	
12		16	

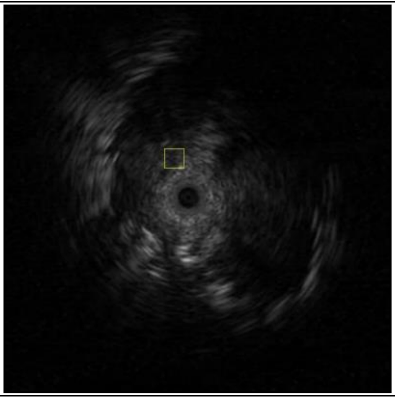
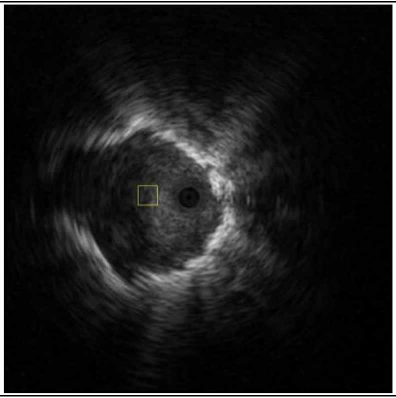
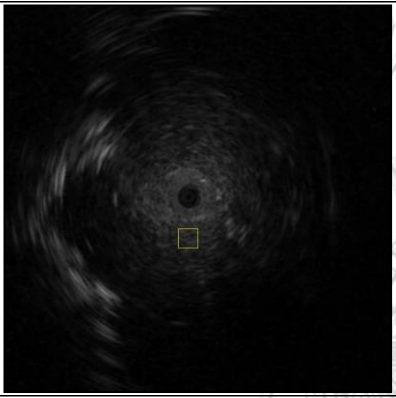
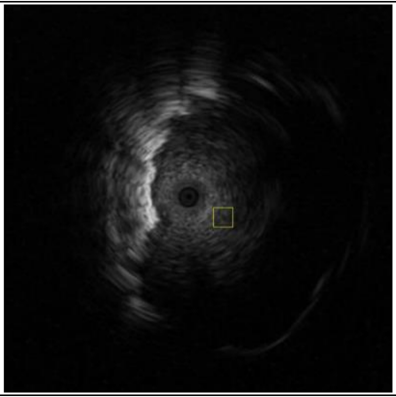
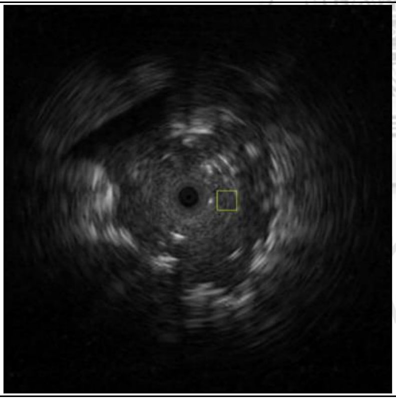
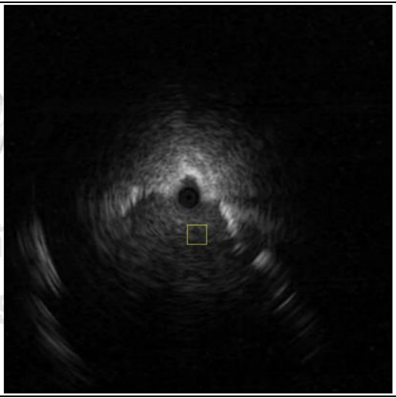
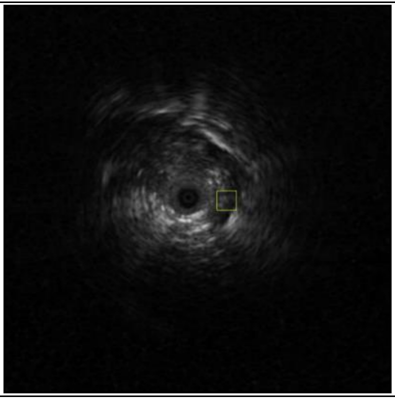
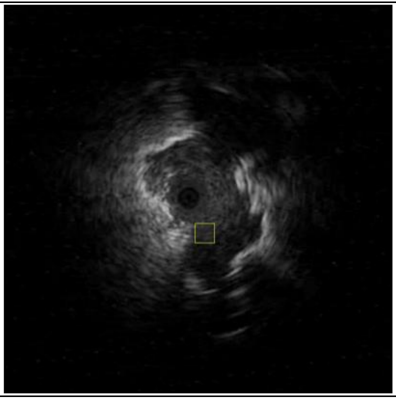
File	Best Frame	File	Best Frame
17		21	
18		22	
19		23	
20		24	

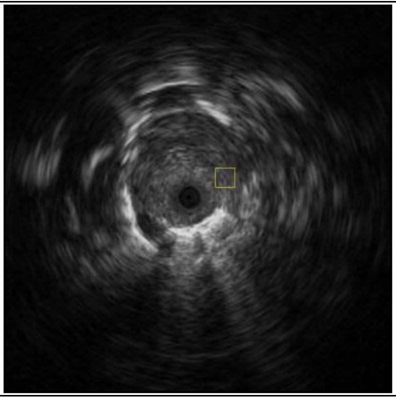
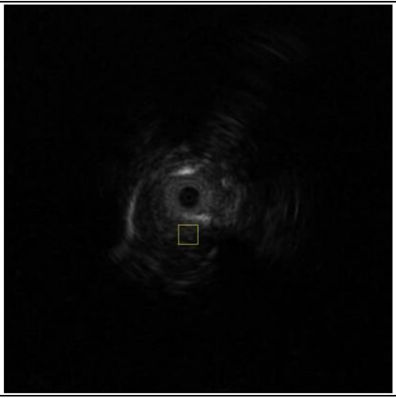
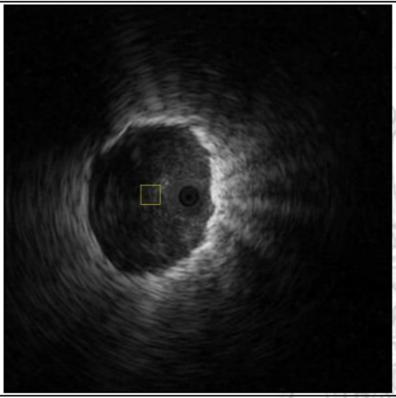
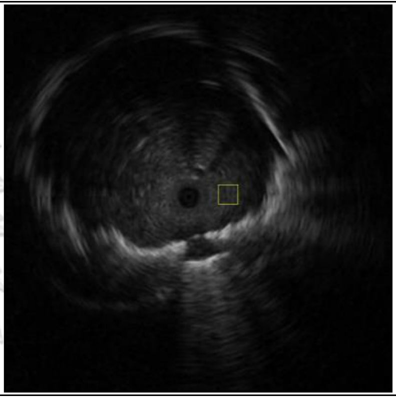
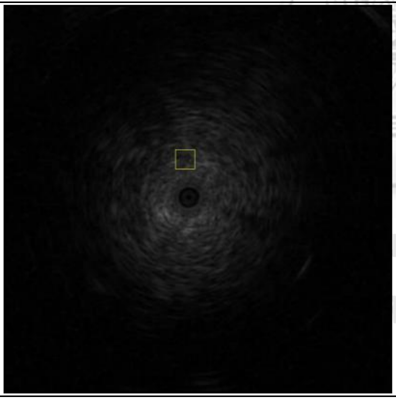
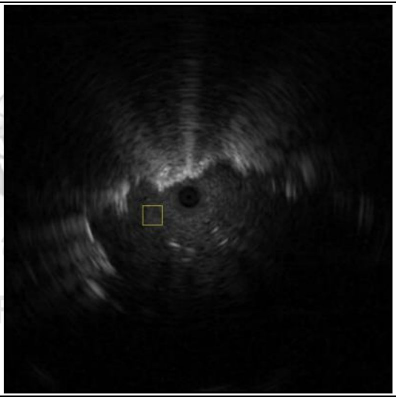
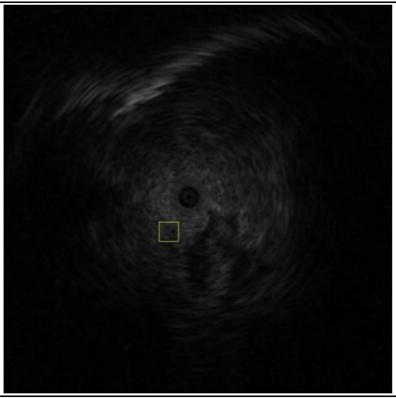
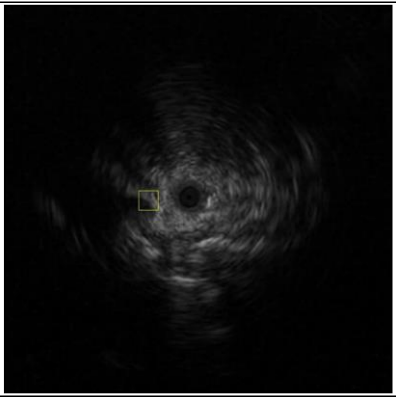
มหาวิทยา
RN UNIVER

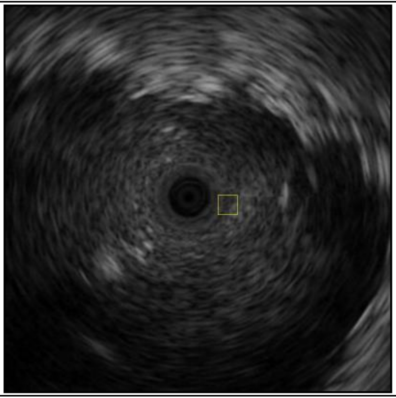
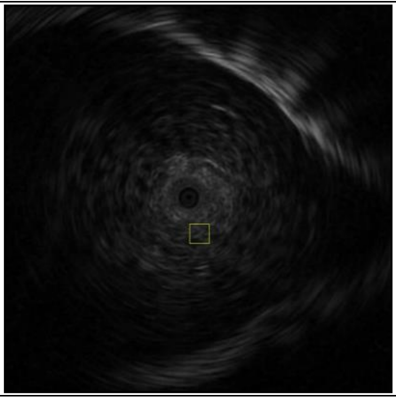
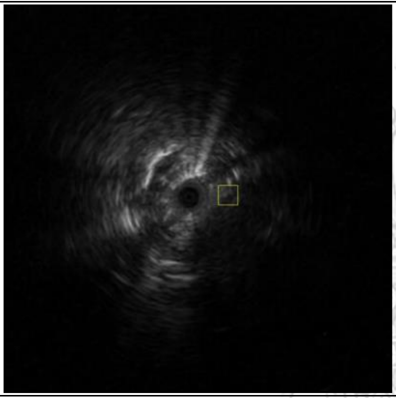
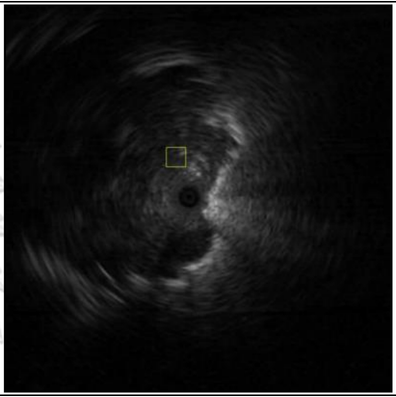
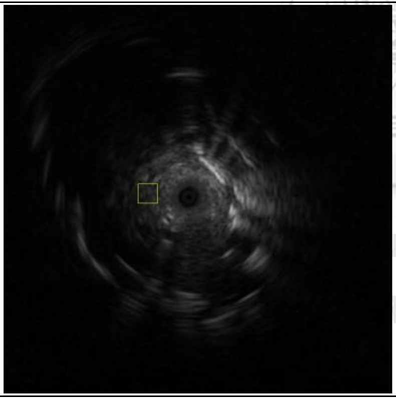
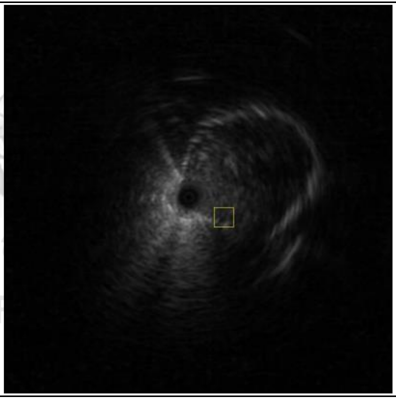
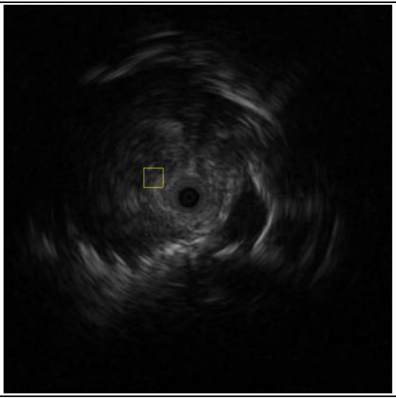
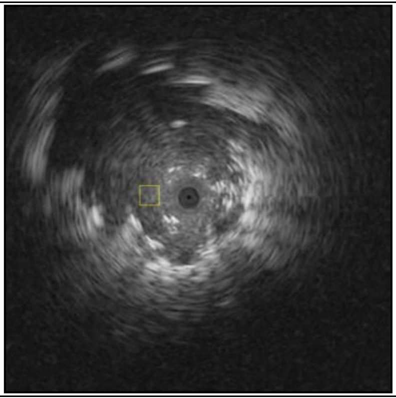
File	Best Frame	File	Best Frame
25		29	
26		30	
27		31	
28		32	

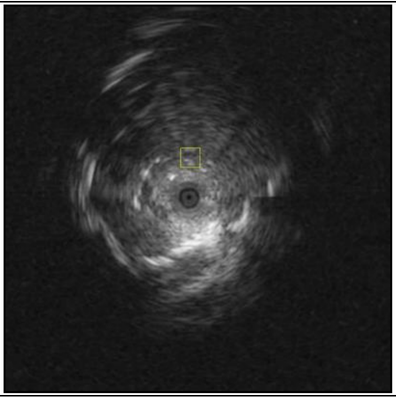
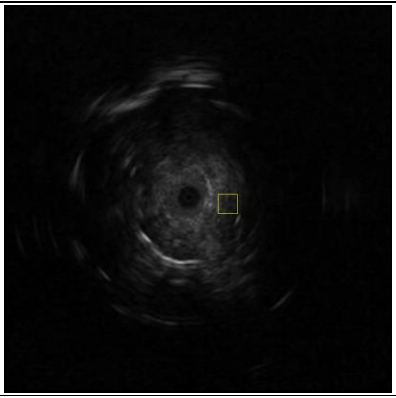
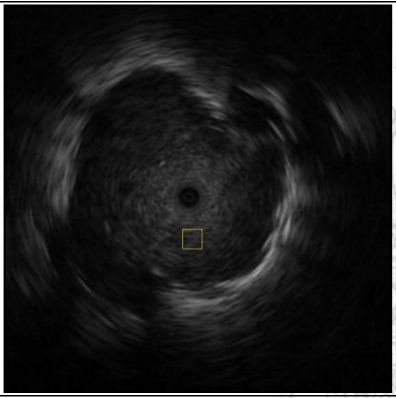
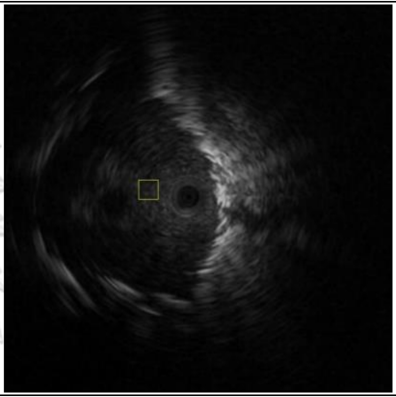
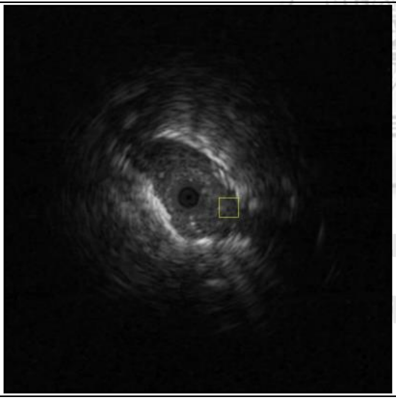
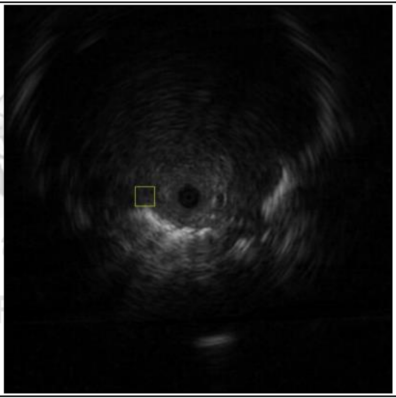
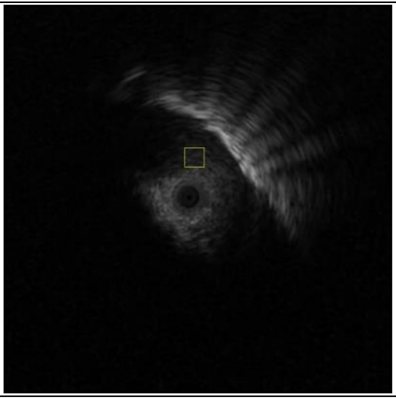
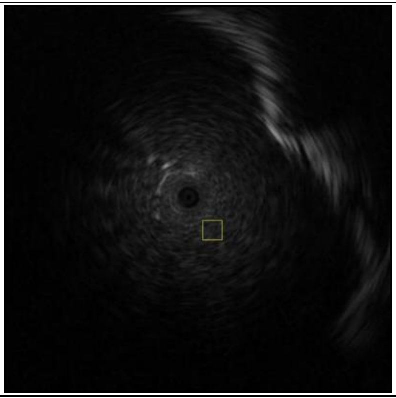
หาวิทยา
RN UNIVER

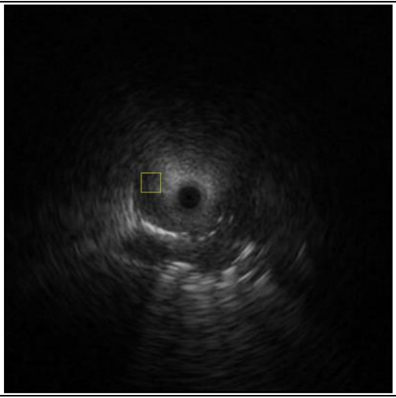
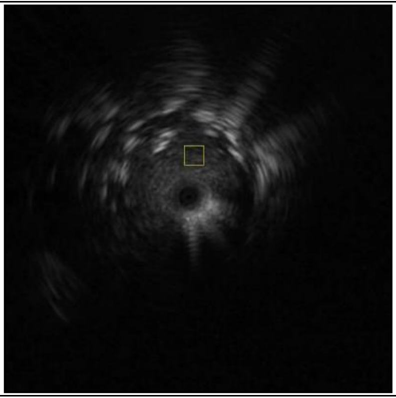
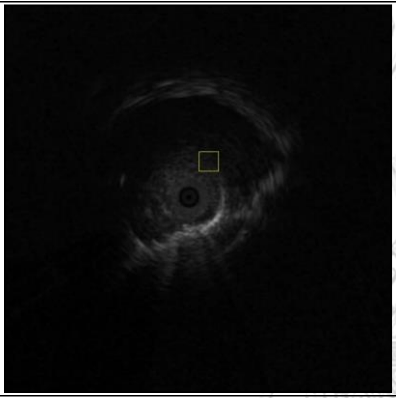
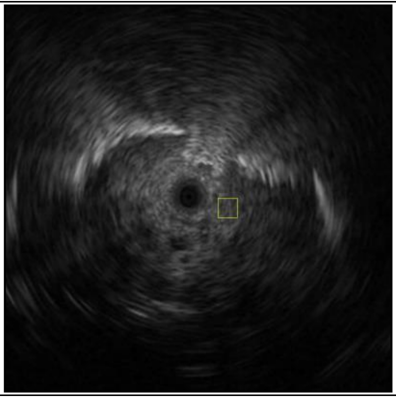
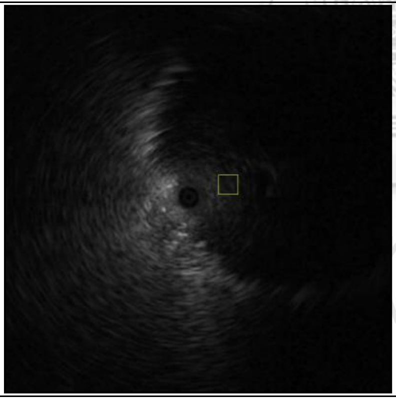
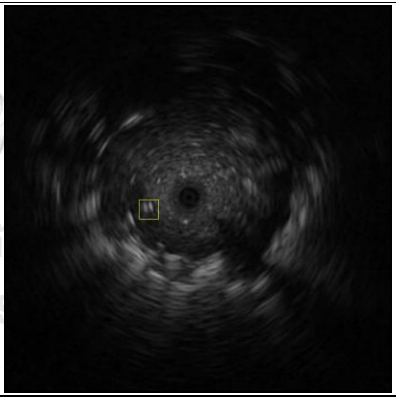
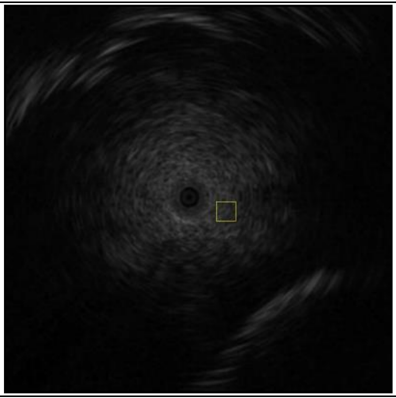
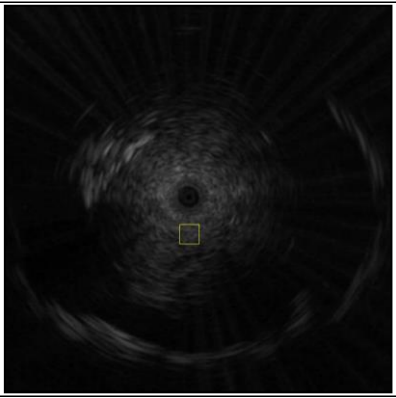
File	Best Frame	File	Best Frame
33		37	
34		38	
35		39	
36		40	

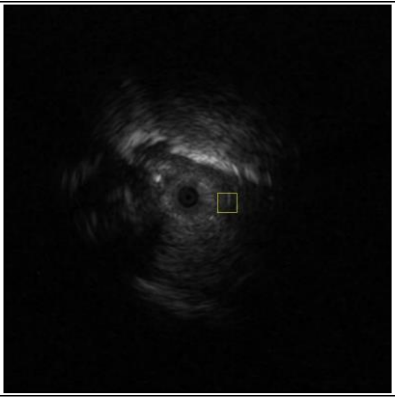
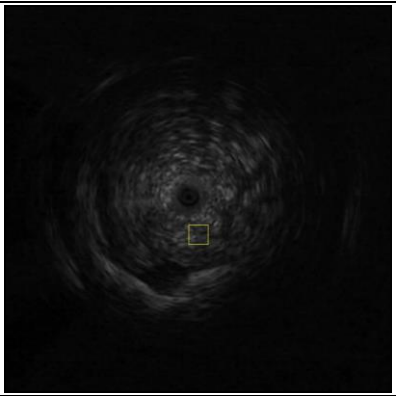
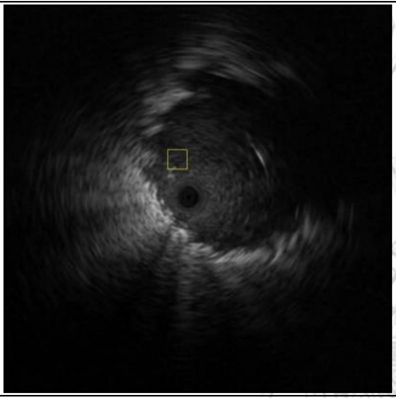
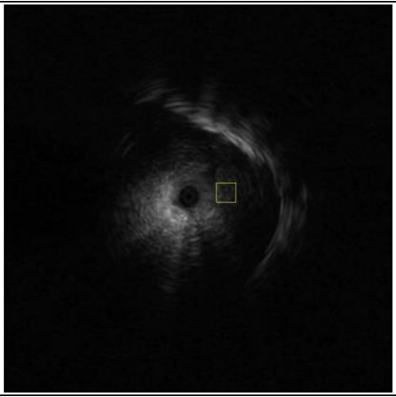
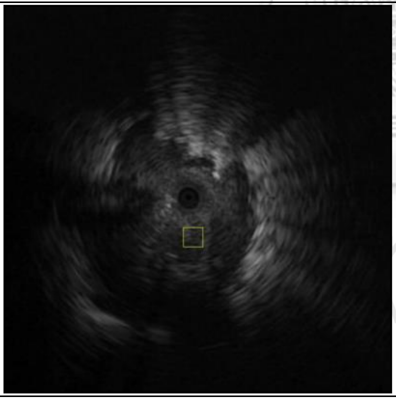
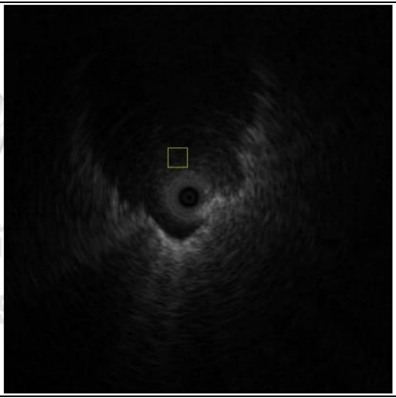
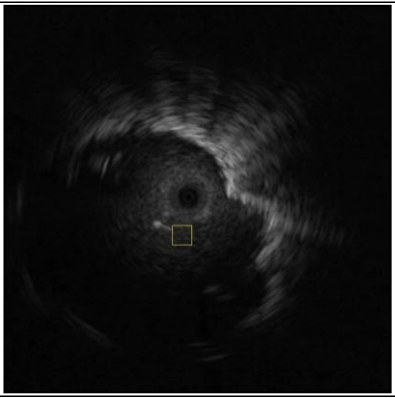
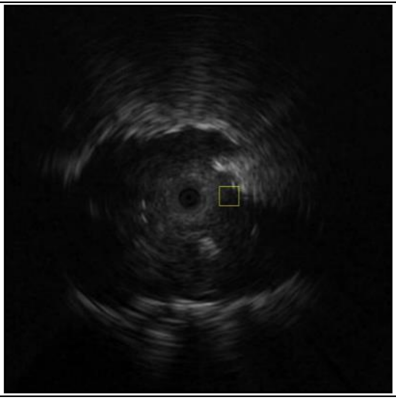
File	Best Frame	File	Best Frame
41		45	
42		46	
43		47	
44		48	

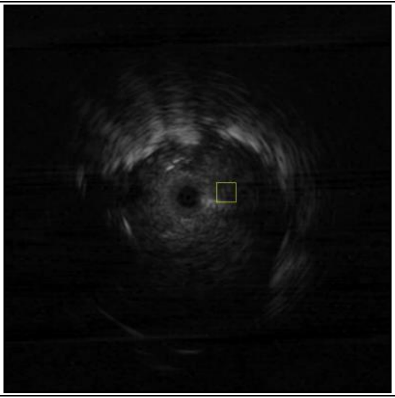
File	Best Frame	File	Best Frame
49		53	
50		54	
51		55	
52		56	

File	Best Frame	File	Best Frame
57		61	
58		62	
59		63	
60		64	

File	Best Frame	File	Best Frame
65		69	
66		70	
67		71	
68		72	

File	Best Frame	File	Best Frame
73		77	
74		78	
75		79	
76		80	

File	Best Frame	File	Best Frame
81		85	
82		86	
83		87	
84		88	

File	Best Frame	File	Best Frame
89			

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

APPENDIX B Feature extraction value

Table Feature using weight sum GLCM

No.	L0	U0	S0	L45	U45	S45	L90	U90	S90	L135	U135	S135	target
1	169	191	360	556	326	882	533	311	844	522	339	861	m
3	375	403	778	595	447	1042	602	455	1057	630	540	1170	m
5	625	432	1057	583	582	1165	385	535	920	380	686	1066	m
8	496	608	1104	671	634	1305	759	558	1317	774	524	1298	m
9	351	409	760	474	654	1128	435	593	1028	491	553	1044	m
10	541	416	957	599	487	1086	382	402	784	431	603	1034	m
12	296	367	663	412	602	1014	419	527	946	527	521	1048	m
17	296	348	644	678	391	1069	650	361	1011	629	410	1039	m
19	264	430	694	497	605	1102	535	568	1103	591	521	1112	m
20	427	556	983	475	587	1062	407	454	861	593	488	1081	m
21	505	318	823	523	461	984	340	468	808	400	630	1030	m
27	533	345	878	570	402	972	371	395	766	414	614	1028	m
29	445	376	821	668	446	1114	678	420	1098	661	473	1134	m
30	411	612	1023	425	587	1012	444	434	878	649	518	1167	m
31	478	616	1094	572	531	1103	558	402	960	721	470	1191	m
35	398	482	880	654	514	1168	668	537	1205	662	569	1231	m
41	539	438	977	603	503	1106	488	409	897	488	522	1010	m
42	214	529	743	285	684	969	228	446	674	471	400	871	m
46	422	587	1009	668	530	1198	702	471	1173	705	528	1233	m
48	461	516	977	548	699	1247	504	716	1220	535	713	1248	m
49	269	380	649	433	619	1052	394	534	928	451	484	935	m
50	243	539	782	511	620	1131	590	519	1109	697	423	1120	m
52	577	389	966	677	422	1099	580	392	972	569	545	1114	m
53	325	513	838	476	756	1232	488	681	1169	501	608	1109	m
54	396	372	768	608	541	1149	563	481	1044	522	489	1011	m
55	358	550	908	496	629	1125	451	521	972	563	450	1013	m
56	649	399	1048	585	487	1072	370	521	891	437	718	1155	m
57	350	334	684	460	601	1061	447	588	1035	499	618	1117	m
62	414	337	751	582	494	1076	529	483	1012	511	531	1042	m
66	442	379	821	712	525	1237	698	529	1227	628	577	1205	m
68	463	379	842	472	626	1098	424	627	1051	449	674	1123	m
70	462	608	1070	469	583	1052	496	522	1018	641	554	1195	m
73	320	394	714	431	615	1046	352	496	848	448	585	1033	m
74	207	241	448	378	541	919	334	440	774	389	456	845	m
76	453	358	811	524	443	967	350	306	656	427	504	931	m
79	350	446	796	459	511	970	374	338	712	520	424	944	m
80	384	707	1091	458	664	1122	526	425	951	763	441	1204	m
81	493	602	1095	523	563	1086	487	406	893	649	497	1146	m
85	460	448	908	633	514	1147	596	383	979	585	448	1033	m
86	329	376	705	457	555	1012	411	528	939	481	553	1034	m
87	335	488	823	478	452	930	443	225	668	619	341	960	m
89	349	673	1022	489	672	1161	527	452	979	691	426	1117	m
91	333	305	638	494	551	1045	513	587	1100	539	594	1133	m
96	304	395	699	525	467	992	514	323	837	568	312	880	m
97	370	345	715	455	449	904	450	433	883	503	566	1069	m

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

99	140	200	340	451	322	773	446	251	697	543	312	855	m
102	523	348	871	523	379	902	359	404	763	463	598	1061	m
103	361	523	884	400	620	1020	320	386	706	522	423	945	m
104	629	501	1130	603	561	1164	434	488	922	541	628	1169	m
105	533	486	1019	536	586	1122	315	486	801	478	606	1084	m
106	352	430	782	447	629	1076	375	616	991	458	598	1056	m
110	489	395	884	584	450	1034	410	324	734	454	522	976	m
114	297	461	758	409	500	909	405	361	766	574	486	1060	m
122	160	174	334	414	257	671	366	192	558	422	255	677	m
130	536	181	717	713	275	988	470	270	740	375	493	868	m
6	376	487	863	474	562	1036	379	412	791	548	485	1033	b
7	283	360	643	501	441	942	507	436	943	566	496	1062	b
11	341	167	508	545	264	809	366	234	600	341	401	742	b
15	658	284	942	642	379	1021	285	397	682	306	747	1053	b
18	214	200	414	388	399	787	355	416	771	412	490	902	b
24	536	292	828	630	396	1026	431	404	835	451	599	1050	b
25	279	326	605	300	615	915	278	572	850	388	612	1000	b
28	448	423	871	414	544	958	220	457	677	418	589	1007	b
32	328	228	556	541	357	898	443	321	764	439	437	876	b
33	351	406	757	513	628	1141	443	603	1046	435	604	1039	b
36	419	416	835	622	535	1157	576	503	1079	547	480	1027	b
43	608	473	1081	619	610	1229	391	522	913	439	648	1087	b
44	314	445	759	458	653	1111	440	616	1056	483	608	1091	b
45	634	305	939	566	428	994	334	456	790	316	696	1012	b
61	530	555	1085	540	554	1094	468	442	910	620	595	1215	b
63	332	784	1116	374	710	1084	464	473	937	752	396	1148	b
69	568	419	987	576	556	1132	454	576	1030	456	691	1147	b
71	328	431	759	480	560	1040	424	470	894	461	448	909	b
72	298	341	639	523	450	973	533	438	971	598	451	1049	b
82	632	308	940	556	451	1007	322	485	807	370	732	1102	b
84	348	494	842	364	523	887	307	348	655	549	428	977	b
88	415	352	767	521	445	966	306	338	644	405	504	909	b
90	686	478	1164	671	553	1224	480	570	1050	504	728	1232	b
92	298	476	774	546	585	1131	576	517	1093	629	477	1106	b
93	477	254	731	543	343	886	308	297	605	352	534	886	b
94	655	227	882	615	327	942	293	411	704	320	731	1051	b
95	362	370	732	469	390	859	379	290	669	517	433	950	b
98	428	337	765	690	392	1082	609	362	971	578	461	1039	b
101	351	183	534	483	328	811	275	206	481	296	362	658	b
109	371	233	604	437	503	940	376	559	935	421	648	1069	b
111	314	408	722	380	686	1066	428	648	1076	486	631	1117	b
115	397	481	878	497	476	973	457	376	833	575	454	1029	b
120	547	276	823	652	419	1071	435	394	829	387	564	951	b
129	293	499	792	356	498	854	347	253	600	534	379	913	b

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table Feature selection using standard feature

No	Mean	var	SD	skew	kurt	entro	Con0	Con45	Con90	Con135	Cor0	Cor45	Cor90	Cor135	Ener0	Ener45	Ener90	Ener135	Ho0	Ho45	Ho90	Ho135	target
1	19.926	105.073	10.254	0.937	3.392	5.130	2.476	9.846	7.288	8.394	0.988	0.954	0.967	0.963	0.009	0.005	0.005	0.005	0.603	0.430	0.448	0.432	m
3	32.491	91.017	9.543	-0.123	2.292	5.236	6.812	16.214	14.661	24.407	0.962	0.909	0.920	0.864	0.004	0.003	0.003	0.002	0.464	0.352	0.373	0.318	m
5	44.155	253.773	15.935	1.154	4.039	5.744	16.073	28.614	9.065	18.870	0.969	0.944	0.983	0.963	0.003	0.003	0.004	0.003	0.390	0.339	0.442	0.370	m
8	79.236	1105.285	33.256	1.496	4.995	6.573	28.249	87.078	65.559	94.487	0.987	0.960	0.971	0.957	0.002	0.001	0.001	0.001	0.345	0.235	0.254	0.241	m
9	36.441	112.609	10.615	0.728	3.479	5.309	5.474	16.283	13.600	18.986	0.976	0.928	0.940	0.914	0.005	0.003	0.003	0.003	0.511	0.375	0.385	0.347	m
10	35.686	277.097	16.651	1.250	4.752	5.835	10.351	30.275	18.637	25.850	0.982	0.948	0.967	0.955	0.003	0.002	0.002	0.002	0.425	0.302	0.368	0.358	m
12	43.316	720.300	26.847	1.131	4.264	6.503	13.304	44.463	20.899	21.160	0.992	0.973	0.986	0.985	0.002	0.002	0.002	0.002	0.414	0.308	0.390	0.375	m
17	35.801	73.289	8.564	0.155	2.589	5.110	6.480	10.945	6.038	12.468	0.956	0.925	0.959	0.916	0.004	0.004	0.005	0.003	0.447	0.392	0.469	0.396	m
19	33.841	94.306	9.714	0.696	3.703	5.229	4.983	20.448	17.398	22.582	0.974	0.893	0.909	0.882	0.005	0.003	0.003	0.003	0.508	0.326	0.352	0.324	m
20	47.767	135.109	11.627	0.447	3.677	5.499	17.639	25.681	8.116	23.879	0.934	0.905	0.970	0.911	0.003	0.002	0.003	0.002	0.340	0.297	0.434	0.313	m
21	29.486	123.206	11.103	0.698	3.241	5.378	8.795	14.412	7.542	16.274	0.965	0.940	0.970	0.936	0.004	0.003	0.004	0.003	0.452	0.386	0.451	0.374	m
27	40.919	112.073	10.590	0.402	2.792	5.378	12.832	16.409	7.115	21.240	0.942	0.927	0.969	0.903	0.003	0.003	0.004	0.002	0.368	0.340	0.440	0.317	m
29	42.912	380.387	19.510	0.221	1.904	6.059	9.258	29.484	19.041	25.117	0.988	0.963	0.976	0.968	0.003	0.002	0.002	0.002	0.480	0.316	0.342	0.326	m
30	38.121	86.054	9.279	0.287	2.711	5.206	10.036	20.361	7.838	13.461	0.942	0.883	0.955	0.922	0.003	0.003	0.004	0.003	0.394	0.326	0.452	0.370	m
31	52.284	170.214	13.051	-0.061	2.435	5.682	16.518	23.500	9.012	25.139	0.951	0.931	0.974	0.927	0.002	0.002	0.003	0.002	0.353	0.320	0.412	0.313	m
35	55.285	135.099	11.627	0.106	2.437	5.505	9.205	27.655	15.298	18.824	0.966	0.898	0.944	0.927	0.003	0.002	0.003	0.002	0.428	0.314	0.378	0.345	m
41	41.396	223.758	14.963	0.854	3.007	5.700	13.327	19.202	8.354	21.787	0.972	0.960	0.982	0.953	0.003	0.002	0.003	0.002	0.376	0.349	0.463	0.327	m
42	30.643	106.350	10.316	1.396	4.965	5.098	3.797	14.287	9.206	10.194	0.983	0.935	0.958	0.950	0.007	0.004	0.005	0.005	0.542	0.384	0.440	0.418	m
46	53.038	502.777	22.430	1.008	4.567	6.285	16.331	56.675	35.611	41.977	0.984	0.944	0.965	0.959	0.002	0.002	0.002	0.002	0.406	0.294	0.317	0.277	m
48	60.041	238.895	15.461	0.286	2.778	5.928	11.594	43.044	35.090	46.936	0.976	0.909	0.927	0.902	0.002	0.002	0.002	0.001	0.426	0.269	0.285	0.255	m
49	29.648	119.444	10.932	1.381	5.996	5.239	5.316	17.567	12.344	16.293	0.978	0.927	0.949	0.934	0.005	0.004	0.004	0.003	0.521	0.390	0.403	0.391	m
50	44.162	207.783	14.419	0.247	2.934	5.802	13.361	36.743	27.815	42.088	0.968	0.912	0.933	0.898	0.002	0.002	0.002	0.002	0.371	0.281	0.287	0.251	m
52	44.518	249.485	15.800	1.325	5.779	5.796	12.778	29.323	24.008	41.264	0.975	0.942	0.952	0.917	0.003	0.002	0.002	0.002	0.400	0.298	0.329	0.289	m
53	46.271	181.665	13.483	0.334	2.324	5.685	8.316	33.634	18.540	17.683	0.978	0.909	0.950	0.951	0.003	0.002	0.002	0.002	0.427	0.278	0.333	0.342	m
54	36.914	62.283	7.894	0.096	2.452	4.980	5.428	18.226	11.131	12.847	0.956	0.848	0.908	0.893	0.005	0.003	0.004	0.004	0.485	0.331	0.387	0.381	m
55	54.366	113.861	10.674	0.466	3.171	5.381	13.130	27.955	11.966	18.801	0.943	0.878	0.947	0.918	0.004	0.003	0.004	0.003	0.405	0.330	0.410	0.374	m

No	Mean	var	SD	skew	kurt	entro	Con0	Con45	Con90	Con135	Cor0	Cor45	Cor90	Cor135	Ener0	Ener45	Ener90	Ener135	Ho0	Ho45	Ho90	Ho135	target
56	42.686	315.551	17.769	2.547	12.503	5.667	12.690	26.417	21.803	38.131	0.980	0.959	0.969	0.943	0.003	0.002	0.002	0.002	0.420	0.325	0.349	0.306	m
57	38.788	74.557	8.637	-0.181	2.961	5.121	5.012	17.469	13.484	17.427	0.966	0.882	0.909	0.882	0.005	0.003	0.003	0.003	0.496	0.349	0.364	0.339	m
62	39.029	233.397	15.282	0.675	2.874	5.817	10.799	19.775	10.752	21.677	0.977	0.958	0.977	0.953	0.003	0.002	0.003	0.002	0.435	0.345	0.435	0.338	m
66	41.836	133.427	11.555	0.460	2.597	5.471	10.451	14.281	8.240	20.647	0.962	0.945	0.970	0.926	0.003	0.003	0.003	0.002	0.426	0.370	0.431	0.336	m
68	44.077	171.150	13.087	0.491	2.587	5.605	6.728	18.623	12.159	17.125	0.980	0.945	0.966	0.952	0.004	0.002	0.003	0.002	0.466	0.360	0.402	0.352	m
70	67.204	131.016	11.450	0.376	2.911	5.495	19.287	30.139	8.697	23.267	0.927	0.885	0.967	0.910	0.002	0.002	0.004	0.002	0.345	0.290	0.433	0.316	m
73	31.997	71.892	8.482	0.040	2.739	5.085	4.878	12.506	7.132	10.082	0.967	0.914	0.950	0.929	0.005	0.004	0.005	0.004	0.488	0.371	0.465	0.392	m
74	22.753	67.116	8.195	0.021	2.440	4.978	2.513	9.451	6.758	8.098	0.981	0.929	0.950	0.941	0.008	0.004	0.005	0.004	0.582	0.400	0.451	0.423	m
76	39.516	37.912	6.159	-0.081	2.591	4.639	8.449	12.033	3.950	11.491	0.888	0.840	0.948	0.847	0.005	0.004	0.007	0.004	0.424	0.370	0.522	0.383	m
79	40.641	63.666	7.982	-0.289	2.581	4.993	5.024	10.012	3.756	6.398	0.961	0.922	0.970	0.949	0.006	0.004	0.006	0.005	0.484	0.405	0.519	0.460	m
80	47.078	689.575	26.268	0.339	2.205	6.475	32.046	62.705	15.332	29.811	0.978	0.957	0.989	0.979	0.002	0.002	0.002	0.002	0.350	0.291	0.396	0.333	m
81	52.402	187.435	13.695	0.665	2.951	5.687	12.353	15.827	8.794	24.330	0.968	0.958	0.977	0.936	0.003	0.002	0.003	0.002	0.368	0.349	0.426	0.326	m
85	44.995	282.456	16.812	0.707	2.379	5.798	10.317	10.446	8.655	25.168	0.982	0.981	0.985	0.958	0.003	0.003	0.003	0.002	0.416	0.389	0.439	0.321	m
86	31.606	100.350	10.021	0.018	2.443	5.284	5.431	14.939	12.549	18.961	0.973	0.926	0.937	0.905	0.005	0.003	0.003	0.003	0.506	0.380	0.411	0.359	m
87	37.344	198.884	14.107	1.560	5.946	5.517	5.919	13.360	14.151	24.910	0.986	0.967	0.965	0.939	0.005	0.004	0.004	0.003	0.509	0.401	0.404	0.360	m
89	96.751	271.852	16.493	0.341	2.654	6.001	17.474	27.191	11.585	28.360	0.970	0.951	0.979	0.949	0.002	0.002	0.002	0.002	0.347	0.298	0.385	0.313	m
91	39.403	92.721	9.632	-0.035	2.837	5.265	4.135	17.389	15.372	20.295	0.978	0.906	0.917	0.891	0.005	0.003	0.003	0.003	0.512	0.362	0.374	0.323	m
96	32.394	101.309	10.068	1.294	5.662	5.164	5.394	10.083	5.485	10.471	0.974	0.951	0.977	0.953	0.006	0.004	0.007	0.005	0.502	0.418	0.519	0.434	m
97	33.204	45.196	6.725	0.057	2.876	4.767	5.027	8.559	7.449	14.894	0.944	0.904	0.917	0.832	0.007	0.005	0.005	0.004	0.496	0.429	0.430	0.359	m
99	18.963	26.244	5.125	0.105	2.725	4.351	2.251	5.699	3.318	4.303	0.958	0.893	0.937	0.916	0.014	0.008	0.010	0.009	0.621	0.476	0.544	0.511	m
102	39.998	62.814	7.928	0.266	2.762	4.982	7.284	9.247	5.592	14.799	0.942	0.926	0.956	0.882	0.005	0.004	0.005	0.003	0.444	0.414	0.480	0.356	m
103	38.623	87.792	9.373	0.206	2.619	5.203	4.313	18.421	13.844	16.519	0.975	0.896	0.922	0.907	0.005	0.003	0.003	0.003	0.520	0.335	0.356	0.349	m
104	61.399	103.791	10.191	0.213	2.469	5.351	19.767	22.861	6.891	28.628	0.905	0.889	0.967	0.861	0.002	0.002	0.004	0.002	0.334	0.327	0.447	0.295	m
105	49.077	655.865	25.618	1.606	5.576	6.238	68.361	111.790	20.124	64.524	0.949	0.917	0.985	0.952	0.002	0.002	0.002	0.002	0.307	0.253	0.363	0.285	m
106	43.118	92.145	9.602	-0.216	2.954	5.249	5.802	15.387	10.779	16.016	0.968	0.915	0.942	0.911	0.004	0.003	0.003	0.003	0.474	0.355	0.406	0.342	m
110	39.198	132.166	11.500	1.921	9.160	5.226	7.767	20.108	18.942	30.997	0.971	0.925	0.928	0.882	0.006	0.004	0.004	0.003	0.499	0.390	0.410	0.356	m
114	33.202	46.354	6.810	0.075	2.462	4.772	3.658	10.185	6.892	9.694	0.960	0.890	0.926	0.895	0.007	0.004	0.005	0.004	0.540	0.383	0.426	0.397	m
122	16.791	29.145	5.400	0.786	3.534	4.346	1.765	4.135	3.329	5.245	0.970	0.929	0.943	0.910	0.017	0.011	0.013	0.010	0.635	0.511	0.532	0.481	m

No	Mean	var	SD	skew	kurt	entro	Con0	Con45	Con90	Con135	Cor0	Cor45	Cor90	Cor135	Ener0	Ener45	Ener90	Ener135	Ho0	Ho45	Ho90	Ho135	target
98	47.327	245.615	15.677	0.602	2.753	5.854	14.139	19.128	8.506	23.962	0.973	0.963	0.983	0.953	0.002	0.002	0.003	0.002	0.389	0.346	0.409	0.317	b
101	21.279	31.099	5.578	0.801	4.145	4.431	3.016	5.525	2.151	3.950	0.952	0.911	0.966	0.935	0.012	0.009	0.013	0.010	0.565	0.474	0.601	0.527	b
109	32.111	115.032	10.729	0.718	3.605	5.337	15.140	18.634	4.965	19.963	0.935	0.919	0.979	0.914	0.004	0.003	0.005	0.003	0.440	0.387	0.504	0.384	b
111	39.768	121.928	11.046	0.399	2.722	5.415	5.496	19.185	14.753	19.623	0.977	0.922	0.940	0.919	0.004	0.003	0.003	0.002	0.507	0.345	0.369	0.337	b
115	43.259	143.464	11.981	0.783	4.063	5.501	9.446	26.034	19.081	28.331	0.968	0.909	0.934	0.902	0.004	0.003	0.003	0.002	0.446	0.324	0.364	0.320	b
120	31.565	103.172	10.161	0.565	2.943	5.292	7.154	13.029	6.157	11.904	0.967	0.938	0.970	0.943	0.004	0.003	0.004	0.003	0.452	0.374	0.479	0.392	b
129	27.194	225.810	15.032	2.349	11.157	5.376	18.002	26.703	7.321	20.295	0.960	0.945	0.987	0.953	0.005	0.004	0.006	0.004	0.464	0.416	0.525	0.404	b

APPENDIX C Confusion matrix of all classifier

TABLE RESULTS OF FEATURE GROUP 1

	F S.	Correct class			B E.	Correct class			GS.	Correct class			
	Fna	Malig	Benign		Bna	Malig	Benign		Gna	Malig	Benign		
Classified as	Malignant	45	25	60.674		52	27	66.292		47	18	70.787	acc
	Benign	10	9	81.818		3	7	94.545		8	16	85.455	sen
				26.471				20.588				47.059	spec
	Ftree				Btree				Gtree				
Classified as	Malignant	54	29	66.292		52	31	61.798		52	16	78.652	acc
	Benign	1	5	98.182		3	3	94.545		3	18	94.545	sen
				14.706				8.824				52.941	spec
	Fneu				Bneu				Gneu				
Classified as	Malignant	55	31	65.169		55	34	61.798		53	18	77.528	acc
	Benign	0	3	100.000		0	0	100.000		2	16	96.364	sen
				8.824				0.000				47.059	spec
	FLR				BLR				GLR				
Classified as	Malignant	48	26	62.921		53	30	64.045		52	20	74.157	acc
	Benign	7	8	87.273		2	4	96.364		3	14	94.545	sen
				23.529				11.765				41.176	spec
	Flog				Blog				Glog				
Classified as	Malignant	55	33	62.921		55	34	61.798		54	22	74.157	acc
	Benign	0	1	100.000		0	0	100.000		1	12	98.182	sen
				2.941				0.000				35.294	spec
	Fsvm				Bsvm				Gsvm				
Classified as	Malignant	35	14	61.798		38	19	59.551		47	11	78.652	acc
	Benign	20	20	63.636		17	15	69.091		8	23	85.455	sen
				58.824				44.118				67.647	spec
	FLDA				BLDA				GLDA				
Classified as	Malignant	49	24	66.292		55	34	61.798		50	19	73.034	acc
	Benign	6	10	89.091		0	0	100.000		5	15	90.909	sen
				29.412				0.000				44.118	spec
	Fknn				Bknn				Gknn				
Classified as	Malignant	42	20	62.921		36	18	58.427		52	16	78.652	acc
	Benign	13	14	76.364		19	16	65.455		3	18	94.545	sen
				41.176				47.059				52.941	spec

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

TABLE RESULTS OF FEATURE GROUP 2

	F S.	Correct class			B E.	Correct class			GS.	Correct class			
	Fna	Malig	Benign		Bna	Malig	Benign		Gna	Malig	Benign		
Classified as	Malignant	52	26	67.416		48	21	68.539		47	16	73.034	acc
	Benign	3	8	94.545		7	13	87.273		8	18	85.455	sen
				23.529				38.235				52.941	spec
	Ftree				Btree				Gtree				
Classified as	Malignant	53	28	66.292		54	28	67.416		54	20	76.404	acc
	Benign	2	6	96.364		1	6	98.182		1	14	98.182	sen
				17.647				17.647				41.176	spec
	Fneu				Bneu				Gneu				
Classified as	Malignant	50	24	67.416		46	22	65.169		47	11	78.652	acc
	Benign	5	10	90.909		9	12	83.636		8	23	85.455	sen
				29.412				35.294				67.647	spec
	FLR				BLR				GLR				
Classified as	Malignant	47	20	68.539		51	32	59.551		53	22	73.034	acc
	Benign	8	14	85.455		4	2	92.727		2	12	96.364	sen
				41.176				5.882				35.294	spec
	Flog				Blog				Glog				
Classified as	Malignant	55	29	67.416		55	29	67.416		55	26	70.787	acc
	Benign	0	5	100		0	5	100		0	8	100.000	sen
				14.706				14.706				23.529	spec
	Fsvm				Bsvm				Gsvm				
Classified as	Malignant	40	20	60.674		43	16	68.539		48	10	80.899	acc
	Benign	15	14	72.727		12	18	78.182		7	24	87.273	sen
				41.176				52.941				70.588	spec
	FLDA				BLDA				GLDA				
Classified as	Malignant	49	23	67.416		50	27	64.045		49	19	71.910	acc
	Benign	6	11	89.091		5	7	90.909		6	15	89.091	sen
				32.353				20.588				44.118	spec
	Fknn				Bknn				Gknn				
Classified as	Malignant	44	23	61.798		30	16	53.933		52	18	76.404	acc
	Benign	11	11	80		25	18	54.545		3	16	94.545	sen
				32.353				52.941				47.059	spec

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

TABLE RESULTS OF FEATURE GROUP 3

	F S.	Correct class			B E.	Correct class			GS.	Correct class			
	Fna	Malig	Benign		Bna	Malig	Benign		Gna	Malig	Benign		
Classified as	Malignant	47	21	67.416		47	21	67.416		47	18	70.787	acc
	Benign	8	13	85.455		8	13	85.455		8	16	85.455	sen
				38.235				38.235				47.059	spec
	Ftree				Btree				Gtree				
Classified as	Malignant	54	29	66.292		54	28	67.416		51	19	74.157	acc
	Benign	1	5	98.182		1	6	98.182		4	15	92.727	sen
				14.706				17.647				44.118	spec
	Fneu				Bneu				Gneu				
Classified as	Malignant	50	25	66.292		51	24	68.539		46	10	78.652	acc
	Benign	5	9	90.909		4	10	92.727		9	24	83.636	sen
				26.471				29.412				70.588	spec
	FLR				BLR				GLR				
Classified as	Malignant	49	23	67.416		47	19	69.663		50	19	73.034	acc
	Benign	6	11	89.091		8	15	85.455		5	15	90.909	sen
				32.353				44.118				44.118	spec
	Flog				Blog				Glog				
Classified as	Malignant	55	29	67.416		55	34	61.798		54	24	71.910	acc
	Benign	0	5	100		0	0	100		1	10	98.182	sen
				14.706				0				29.412	spec
	Fsvm				Bsvm				Gsvm				
Classified as	Malignant	38	18	68.354		38	23	55.056		48	5	86.517	acc
	Benign	7	16	84.444		17	11	69.091		7	29	87.273	sen
				47.059				32.353				85.294	spec
	FLDA				BLDA				GLDA				
Classified as	Malignant	48	21	68.539		50	27	64.045		50	17	75.281	acc
	Benign	7	13	87.273		5	7	90.909		5	17	90.909	sen
				38.235				20.588				50.000	spec
	Fknn				Bknn				Gknn				
Classified as	Malignant	39	15	65.169		36	20	56.18		53	21	74.157	acc
	Benign	16	19	70.909		19	14	65.455		2	13	96.364	sen
				55.882				41.176				38.235	spec

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

VITA

Name	Banphatree Khomkham
Date of Birth	27 March 1992
Place of Birth	Bangkok, Thailand
Education	B.Sc. (Mathematics), Khon Kaen University, 2014 (First Class Honor)
Scholarship	Development and Promotion of Science and Technology Talents Project (DPST)
Publication	

1. Khomkham, B., & Thongjunthuk, T. (2014). Necessary Conditions of Existence of Solutions. The 9th Conference on Science and Technology for Youths, Bangkok, Thailand. (Presentation)

2. Khomkham, B., A. Wattanathum, and R. Lipikorn. (2017). Peripheral pulmonary lesion classification from endobronchial ultrasonography images using weight-sum of upper and lower GLCM feature. in the 7th International Conference on System Engineering and Technology (ICSET), 2017.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY