รหัสโครงการ SUT7-705-60-24-28

รายงานการวิจัย

การวิเคราะห์ภาพแมมโมแกรมเพื่อวินิจฉัยมะเร็งเต้านมด้วยเทคนิคการปรับปรุงภาพและซัพพอร์ตเวกเตอร์แมชชีน

(Mammogram Image Analysis to Diagnose Breast Cancer with Image Enhancement and Support Vector Machine Techniques)

ได้รับทุนอุดหนุนการวิจัยจาก
มหาวิทยาลัยเทคโนโลยีสุรนารี

ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว

รหัสโครงการ SUT7-705-60-24-28

รายงานการวิจัย

การวิเคราะห์ภาพแมมโมแกรมเพื่อวินิจฉัยมะเร็งเต้านมด้วยเทคนิคการปรับปรุงภาพและซัพพอร์ตเวกเตอร์แมชชีน

(Mammogram Image Analysis to Diagnose Breast Cancer with Image Enhancement and Support Vector Machine Techniques)

ผู้วิจัย

หัวหน้าโครงการ

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ

ผู้วิจัยร่วม

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ

อาจารย์ ดร.เกตุกาญจน์ ไชยขันธุ์

สาขาวิชาวิศวกรรมคอมพิวเตอร์

สำนักวิชาวิศวกรรมศาสตร์

ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยเทคโนโลยีสุรนารี ปีงบประมาณ พ.ศ. 2560 และ 2561

ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว

กุมภาพันธ์ 2563

กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบคุณมหาวิทยาลัยเทคโนโลยีสุรนารี และสำนักงานคณะกรรมการวิจัยแห่งชาติ ที่สนับสนุนโครงการวิจัยนี้ด้วยการจัดสรรงบประมาณให้ในปีงบประมาณ พ.ศ. 2560 และพ.ศ.2561 ขอขอบคุณสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์และสถาปัตยกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ที่อนุญาตให้ ดร.เกตุกาญจน์ ไชยขันธุ์ เข้าร่วมดำเนินงานในโครงการวิจัยนี้ รวมถึงขอขอบคุณผู้ทรงคุณวุฒิทั้งภายนอกและภายในมหาวิทยาลัย ที่ได้เสียสละเวลาทำหน้าที่ตรวจข้อเสนอโครงการวิจัยและตรวจร่างรายงานการวิจัยฉบับสมบูรณ์ ข้อเสนอแนะจากผู้ทรงคุณวุฒิทุกท่านเป็นประโยชน์อย่างมากต่อคณะผู้วิจัยในการปรับปรุงการออกแบบและการดำเนินงานของโครงการวิจัย งานวิจัยนี้สำเร็จได้อย่างดีด้วยการมีส่วนร่วมจากนักศึกษาทั้งในระดับปริญญามหาบัณฑิตและปริญญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ได้ทำหน้าที่เป็นผู้ช่วยวิจัยในโครงการวิจัยนี้

บทคัดย่อภาษาไทย

ภาพแมมโมแกรมเป็นภาพถ่ายทางรังสีใช้ช่วยในการวินิจฉัยมะเร็งเต้านมและกำหนดขอบเขตในการตัดชิ้นเนื้อพิสูจน์การวิเคราะห์ภาพแมมโมแกรมมีจุดประสงค์เพื่อจำแนกประเภทเนื้องอกในภาพแมมโมแกรมว่าเป็นชนิดอันตรายหรือไม่อันตราย ประโยชน์ของการจำแนกอัตโนมัติจะช่วยรังสีแพทย์ให้วินิจฉัยโรคมะเร็งเต้านมได้ถูกต้อง ในปัจจุบันมีนักวิจัยจำนวนมากพัฒนาประสิทธิภาพของการจำแนกภาพแมมโมแกรมโดยใช้เทคนิควิธีต่าง ๆ ของการประมวลผลภาพร่วมกับเทคนิควิธีการเรียนรู้ของเครื่อง เพื่อเพิ่มความแม่นในการจำแนก การปรับปรุงภาพก่อนการนำไปจำแนกเป็นขั้นตอนที่สำคัญ เนื่องจากภาพแมมโมแกรมอาจมีความไม่ชัดเจนหรือมีสัญญาณรบกวนในภาพ ทำให้การจำแนกได้ผลที่ไม่ดีนัก ดังนั้นงานวิจัยนี้จึงเสนอวิธีการปรับปรุงภาพคือการกำจัดสัญญาณรบกวนภายในภาพออกไปแล้วจึงทำการปรับปรุงภาพโดยทำให้ความเข้มสีบริเวณก้อนเนื้อในภาพชัดเจนขึ้น จากนั้นจึงใช้เทคนิคการประมวลผลภาพด้วยวิธีการหาขอบเขตที่น่าสนใจ โดยใช้ขั้นตอนวิธีในการตัดเฉพาะบริเวณก้อนเนื้อในภาพแมมโมแกรมเพื่อนำมาประมวลผล หลังจากได้บริเวณขอบเขตที่สนใจแล้ว ขั้นตอนก่อนการจำแนกอีกขั้นตอนหนึ่งคือการหาลักษณะสำคัญภายในบริเวณขอบเขตที่สนใจ โดยงานวิจัยนี้จะพิจารณาลักษณะสำคัญ 3 ลักษณะ คือ ลักษณะสำคัญของลวดลาย ลักษณะสำคัญของฮิสโตแกรม และ ลักษณะสำคัญของรูปร่าง โดยเฉพาะลักษณะสำคัญของรูปร่างได้มีการเพิ่มชุดข้อมูลต่อท้ายชุดข้อมูลเดิมโดยพิจารณาจากความถี่ของกราฟฮิสโตแกรมของรอยหยักบริเวณเส้นขอบของก้อนเนื้อ และในขั้นตอนสุดท้าย ลักษณะสำคัญทั้ง 3 แบบจะถูกนำไปใช้ในการจำแนก ด้วยเทคนิควิธีในการจำแนกข้อมูลแบบมีผู้สอนที่ชื่อว่าซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ร่วมกับเคอร์เนลฟังก์ชันหลายแบบ งานวิจัยนี้จะเปรียบเทียบประสิทธิภาพการจำแนกระหว่างเทคนิควิธีซัพพอร์ตเวกเตอร์แมชชีนกับเทคนิคการจำแนกอื่น ๆ เช่น โครงข่ายประสาทเทียม และนาอีฟเบย์

บทคัดย่อภาษาอังกฤษ

Mammography is a special type of low-powered x-ray method that has been used to improve diagnostic and decrease the number of unneeded biopsies. Detection breast cancer in early stage can help treatment successful. Many researches show that malignant breast tumors tend to demonstrate irregular and undulated shapes, whereas benign breast tumors are regularly round and smooth shapes. Consequently, many researches about tumor shape may help in maintaining diagnosis. Thus, the contour feature of tumor contour is very significant feature to distinguish between malignant and benign tumor. In this paper, we propose an approach to automatically appraise the density and contrast of breast images using gamma correction to increase the intensity of dense pixels with light intensity and vice versa to decrease the sparse intensity pixels showing dark intensity. In the segmentation process, we use region growing technique to get region of interest. We also extract three important features including texture, shape, and intensity histogram. Especially add data of shape feature into the original data by considering histogram of serrated contour in each tumor. In the classification process, we use SVM to classify tumor into two classes: malignant and benign. Moreover, we also compare between SVM classification with Artificial Neural Network and naïve Bayes. Neural Network and Naïve Bays. The results of classification show that SVM gives good classification accuracy more than Artificial Neural Network and naïve Bays.

สารบัญ

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

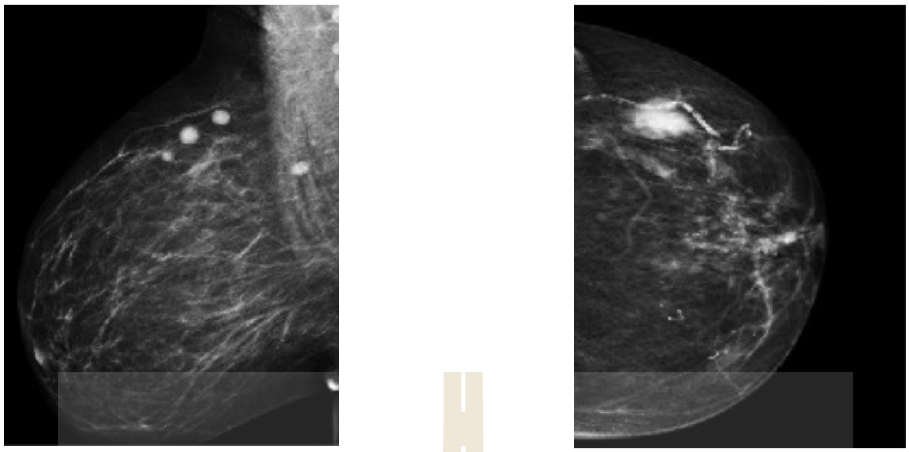
บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

มะเร็งเป็นกลุ่มของโรคที่เกิดจากความผิดปกติของเซลล์ในร่างกายมนุษย์ ซึ่งเซลล์ที่ผิดปกตินั้นจะมีการเปลี่ยนแปลงและมีขนาดใหญ่ลุกลามขึ้นเรื่อย ๆ หากปล่อยไว้โดยไม่มีการตรวจวิเคราะห์และรักษาอาจทำให้ผู้ป่วยเสียชีวิตได้ ส่วนใหญ่นั้นเซลล์มะเร็งจะเป็นลักษณะก้อนเนื้อที่มีความหนาแน่นสูงตรวจสอบได้จากการตรวจชิ้นเนื้อ (biopsy) การวิเคราะห์จากภาพแมมโมแกรม (mammography) การวิเคราะห์จากภาพคลื่นเสียงความถี่สูงหรืออุลตราซาวด์ (ultrasound)

มะเร็งเต้านม (breast cancer) เป็นเซลล์ที่มีความผิดปกติซึ่งเกิดขึ้นภายในทรวงอกของผู้หญิง และเป็นสาเหตุหลักของโรคมะเร็งที่ทำให้ผู้หญิงเสียชีวิตมากที่สุด และพบบ่อยมากกว่ามะเร็งปากมดลูกมะเร็งเต้านมนั้นสามารถพบได้ในผู้ชายเช่นกันแต่พบในอัตราส่วนที่น้อยมากมะเร็งเต้านมมักจะไม่แสดงอาการในเบื้องต้น ผู้หญิงส่วนใหญ่จึงรู้สึกถึงอาการของมะเร็งเต้านมในขณะที่มะเร็งได้ลุกลามไปมากแล้ว ดังนั้นหากมีการวิเคราะห์และวินิจฉัยก้อนเนื้อว่าเป็นเนื้อร้าย (malignant) หรือเป็นเพียงเนื้องอก (benign) ในเบื้องต้นแล้วก็จะทำให้ผู้ป่วยได้มีโอกาสในการรักษาและหายจากอาการป่วยและรอดชีวิตค่อนข้างสูง ในการตรวจวิเคราะห์ก้อนเนื้อด้วยวิธีการผ่าตัดนำเอาก้อนเนื้อนั้นออกมาพิสูจน์ จะค่อนข้างยุ่งยากและมีผลข้างเคียงที่อาจก่อให้เกิดอันตรายต่อคนไข้ ดังนั้นวิธีที่นิยมใช้กันมากในปัจจุบันคือการนำภาพแมมโมแกรม และ ภาพอุลตราซาวน์ของเต้านมมาวิเคราะห์หาความผิดปกติมะเร็งเต้านมมักจะก่อกำเนิดขึ้นจากเนื้อเยื่อในทรวงอกภายในต่อมน้ำนม

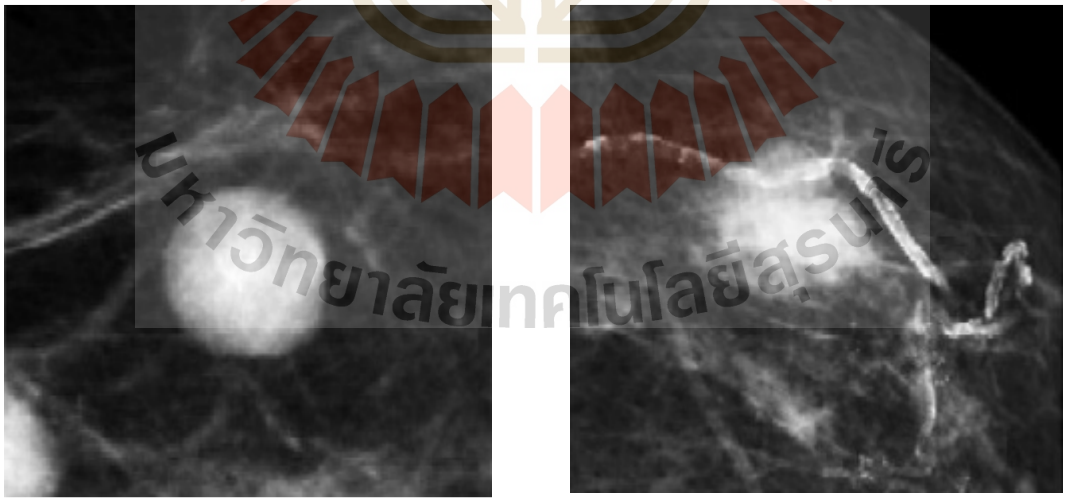
เนื้อเยื่อที่มีความผิดปกตินั้นจากการวิเคราะห์ของรังสีแพทย์จะเห็นเป็นก้อนเนื้อที่มีความหนาแน่นสูงและมีก้อนหินปูนเกาะอยู่ด้วย จากรูปที่ 1.1 แสดงภาพแมมโมแกรมของเต้านม ซึ่งเนื้อเยื่อหรือก้อนเนื้อที่มีความผิดปกตินี้สามารถเป็นได้ทั้งก้อนเนื้องอกที่ไม่ลุกลาม หรืออาจจะเป็นก้อนเนื้อร้ายที่ลุกลามไปยังอวัยวะต่าง ๆ ได้ ทั้งนี้การแบ่งแยกระหว่างเนื้องอกและเนื้อร้ายขึ้นอยู่กับรูปร่างของก้อนเนื้อที่ตรวจพบ



(ก) benign tumor (ข) malignant tumor

รูปที่ 1.1 ภาพแมมโมแกรมจากทรวงอกที่ตรวจพบ (ก) ก้อนเนื้องอกที่ไม่ใช่เนื้อร้าย (ข) ก้อนเนื้อร้าย

จากความรู้ของรังสีแพทย์และนักรังสีวิทยาสรุปได้ว่า ก้อนเนื้องอกส่วนใหญ่มักจะมีรูปร่างขอบที่เป็นทรงกลมหรือทรงรี ในขณะที่ก้อนเนื้อที่ก่อตัวเป็นมะเร็งเต้านมนั้นจะมีขอบบางส่วนที่เป็นทรงกลม และขอบบางส่วนที่มีลักษณะผิดรูปหรือมีรอยหยักค่อนข้างมาก และโดยทั่วไปแล้วก้อนเนื้อที่มีความผิดปกตินั้นเมื่อดูจากภาพแมมโมแกรมมักจะมีความสว่างของสี และความหนาแน่นในบริเวณก้อนเนื้อร้ายมากกว่าบริเวณข้างเคียงหรือบริเวณที่เป็นไขมัน รูปที่ 1.2 แสดงลักษณะรูปร่างของก้อนเนื้องอกที่ไม่ร้ายแรงและก้อนเนื้อที่พบว่าเป็นมะเร็งเต้านม



(ก) benign tumor shape (ข) malignant tumor shape

รูปที่ 1.2 แสดงลักษณะรูปร่างของก้อนเนื้อ (ก) รูปร่างก้อนเนื้องอกที่ไม่ใช่เนื้อร้าย (ข) รูปร่างของก้อนเนื้อร้าย

อย่างไรก็ตามภาพแมมโมแกรมที่ได้มานั้นอาจยังมีสิ่งรบกวนภายในภาพ (noise) ซึ่งอาจทำให้ภาพไม่ชัดเจน ส่งผลให้การวิเคราะห์ของแพทย์อาจมีความผิดพลาดเกิดขึ้นได้ เพราะต้องทำการวิเคราะห์โดยใช้สายตา ในปัจจุบันมีระบบคอมพิวเตอร์ช่วยในการตรวจหาหรือวิเคราะห์โรคเรียกว่า Computer Aided Detection (CAD) เป็นเทคนิคที่ช่วยให้รังสีแพทย์วิเคราะห์เพื่อวินิจฉัยบริเวณที่เป็นมะเร็งได้แม่นยำยิ่งขึ้น โดยเทคนิคนี้ช่วยในการปรับปรุงให้ภาพถ่ายทางการแพทย์มีความชัดเจนและช่วยให้แพทย์วิเคราะห์ได้ถูกต้องมากยิ่งขึ้น

ในงานวิจัยนี้ได้เสนอแนวทางการเพิ่มขีดความสามารถของระบบคอมพิวเตอร์ช่วยในการตรวจหาหรือวิเคราะห์โรคให้มีขีดความสามารถสูงขึ้น โดยใช้การวินิจฉัยภาพแมมโมแกรมเป็นกรณีต้นแบบ ในงานวิจัยนี้เสนอแนวทางการปรับปรุงภาพเพื่อให้ได้ภาพที่ชัดเจนขึ้นในแนวทางเดียวกับ CAD แต่เพิ่มขีดความสามารถในการวินิจฉัยด้วยการใช้เทคนิคการเรียนรู้ของเครื่องเพื่อช่วยให้การวินิจฉัยเพื่อแยกแยะก้อนเนื้อปกติ ออกจากก้อนเนื้อร้ายทำได้แบบอัตโนมัติด้วยการประยุกต์ใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน ซึ่งจัดเป็นอัลกอริทึมการเรียนรู้ที่มีประสิทธิภาพการจำแนกได้แม่นยำมากที่สุดในเทคโนโลยียุคปัจจุบัน การเพิ่มขีดความสามารถของระบบคอมพิวเตอร์นี้จะช่วยให้รังสีแพทย์วิเคราะห์ภาพได้ถูกต้องแม่นยำมากยิ่งขึ้น

1.2 วัตถุประสงค์ของโครงการวิจัย

เพื่อพัฒนาความสามารถของระบบคอมพิวเตอร์ช่วยในการตรวจหาและวิเคราะห์โรคมะเร็งเต้านมจากภาพแมมโมแกรมให้มีขีดความสามารถสูงขึ้นด้วยการใช้เทคนิคการปรับปรุงภาพ ได้แก่ ปรับความสว่างของภาพให้บริเวณก้อนเนื้อปรากฏชัดเจน หาขอบเขตสำคัญของภาพ คัดเลือกเฉพาะฟีเจอร์ที่สำคัญของภาพ จากนั้นจำแนกภาพก้อนเนื้อปกติหรือก้อนเนื้องอกที่ไม่ร้ายแรงออกจากภาพก้อนเนื้อร้าย ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

1.3 ขอบเขตของการวิจัย

- ข้อมูลที่ใช้ในการทดสอบอัลกอริทึม เป็นชุดข้อมูลมาตรฐานจาก University of South Florida Digital Mammography Home Page ชื่อ Digital Database for Screening Mammography (DDSM) (http://marathon.csee.usf.edu/Mammography/Database.html) ข้อมูลนี้ได้มาจากกลุ่มตัวอย่างคนไข้ 2,500 คน โดยในชุดข้อมูล DDSM มีการเก็บภาพแมมโมแกรมของเต้านมด้านซ้ายและด้านขวาในมุมมอง 2 แบบ คือ MLO (mediolateral view) และ CC (caudocranial view)

- การตรวจสอบความถูกต้องของการวิเคราะห์ภาพเพื่อจำแนกภาพก้อนเนื้อไม่ร้ายแรงออกจากภาพก้อนเนื้อร้าย จะใช้การทดสอบด้วยภาพที่แยกไว้สำหรับทำหน้าที่เป็นชุดทดสอบ โดยภาพในชุดทดสอบนี้รังสีแพทย์ได้วินิจฉัยไว้แล้วว่าเป็นก้อนเนื้อประเภทใด ผลการทดสอบของงานวิจัยนี้จึงไม่จำเป็นต้องใช้แพทย์ผู้เชี่ยวชาญยืนยันซ้ำ

1.4 ประโยชน์ที่ได้รับ

งานวิจัยนี้ได้รับประโยชน์จากการดำเนินงานโครงการ ในหลายด้านได้แก่

1) งานวิจัยนี้เป็นการออกแบบแนวคิด และพัฒนาขั้นตอนวิธีในการปรับปรุงและจำแนกภาพแมมโมแกรม เพื่อให้ได้องค์ความรู้ใหม่ในด้านการวิเคราะห์และวินิจฉัยภาพด้วยกระบวนการอัตโนมัติ ประโยชน์ที่ได้รับโดยตรงคือเทคนิคที่พัฒนาขึ้นนี้จะช่วยให้รังสีแพทย์และนักรังสีวิทยา ทำงานได้ง่ายและมีความถูกต้องมากขึ้น ซึ่งจะเป็นประโยชน์ต่อผู้ป่วยโดยตรง เทคนิคและอัลกอริทึมใหม่ที่พัฒนาขึ้นสามารถเผยแพร่ผลงานที่เป็นความก้าวหน้าใหม่ต่อที่ประชุมวิชาการระดับนานาชาติได้จำนวน 3 บทความ และตีพิมพ์ผลงานในวารสารวิชาการระดับนานาชาติได้ 1 บทความ

2) การดำเนินงานวิจัยในส่วนที่เป็นการวิจัยเชิงประจักษ์ ที่จะต้องมีการเก็บรวบรวมข้อมูลและทำการทดลองกับข้อมูลเหล่านั้น จะต้องใช้ผู้ช่วยวิจัยที่เป็นนักศึกษาระดับปริญญาโทและเอก ในสาขาวิชาวิศวกรรมคอมพิวเตอร์ โครงการวิจัยนี้จึงมีประโยชน์ในการพัฒนานักวิจัยรุ่นใหม่ให้สามารถทำงานวิจัยในระดับสูงได้

3) การพัฒนาอัลกอริทึมให้สามารถใช้งานได้ในลักษณะของโปรแกรมต้นแบบหรือ prototype ทำให้ได้โปรแกรมที่สามารถจดลิขสิทธิ์ได้จำนวน 1 โปรแกรม ได้แก่ โปรแกรมจำแนกภาพรังสีเพื่อการวินิจฉัยมะเร็งเต้านม (mammography image classification for breast cancer diagnosis program) ลิขสิทธิ์เลขที่ ว1. 6552 ออกให้ ณ วันที่ 12 มิถุนายน พ.ศ. 2560

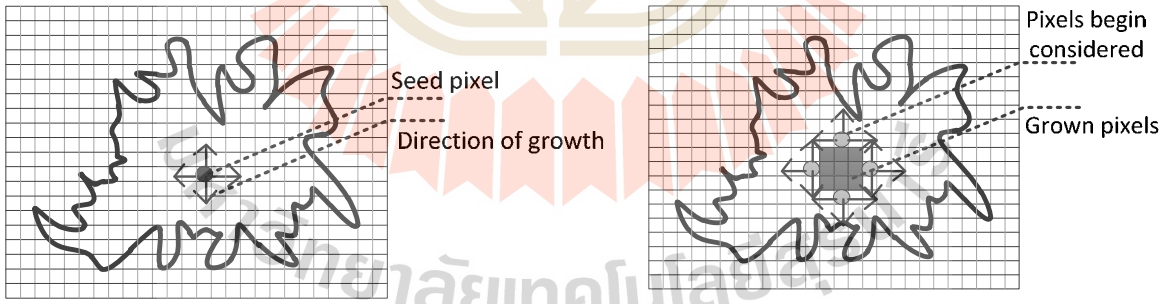
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 เทคนิคการปรับปรุงภาพ

วิธีการปรับปรุงภาพเพื่อให้ภาพแมมโมแกรมมีขอบเขตและลักษณะเด่นของภาพที่ชัดเจน เหมาะสมสำหรับการจำแนกภาพด้วยวิธีอัตโนมัติ ใช้วิธีการพื้นฐานสองวิธีคือการกำหนดขอบเขตของภาพด้วยวิธีการขยายพื้นที่ และการหาลักษณะสำคัญของภาพ เทคนิคพื้นฐานทั้งสองแบบมีรายละเอียดดังนี้

การกำหนดขอบเขตของภาพด้วยวิธีการขยายพื้นที่

การขยายพื้นที่ของส่วนภาพ (region growing) เป็นการแบ่งส่วนภาพบนพื้นฐานของพื้นที่ (region based) ที่มีการกำหนดจุดกึ่งกลาง (seed point) ในภาพ แล้วทำการขยายพื้นที่ (growing) ออกไปยังจุดพิกเซลใกล้เคียง โดยพิจารณาจากค่าระดับสีเทา จนกระทั่งขอบเขตนั้นสิ้นสุดเมื่อมีความเข้มสีของพิกเซลปัจจุบันและพิกเซลใกล้เคียงที่ต่างกันมาก ภายหลังเสร็จสิ้นขั้นตอนการขยายพื้นที่ จะได้จำนวนขอบเขตในภาพหลาย ๆ ขอบเขต ซึ่งแบ่งแยกจากกันอย่างชัดเจน ช่วยให้ขั้นตอนต่อไปสามารถรวมส่วนภาพที่อยู่ติดกันและมีค่าระดับสีเทาใกล้เคียงกันเข้าไว้ด้วยกัน (merging) รูปที่ 2.1 แสดงตัวอย่างของการขยายพื้นที่ของส่วนภาพโดยการพิจารณาพิกเซลใกล้เคียง



รูปที่ 2.1 การขยายส่วนพื้นที่ของภาพ

จุดประสงค์ในการขยายพื้นที่ของส่วนภาพคือการตัดเฉพาะขอบเขตที่น่าสนใจ (Region of Interest: ROI) หรือบริเวณก้อนเนื้อที่มีความหนาแน่นสูง เนื่องจากภาพแมมโมแกรมที่ได้นั้นจะมีขนาดใหญ่และประกอบด้วยส่วนที่เป็นไขมันและก้อนเนื้อ แต่การนำภาพแมมโมแกรมมาจำแนกนั้นจะสนใจเฉพาะบริเวณก้อนเนื้อที่มีความหนาแน่นสูง ดังนั้นจึงต้องตัดบริเวณที่ไม่สำคัญหรือบริเวณที่เป็นไขมันออกไป เพื่อลดขนาดภาพและลดเวลาในการประมวลผลภาพและการจำแนกภาพ

การคำนวณเพื่อการขยายพื้นที่ของส่วนภาพนั้น แสดงดังสมการที่ 2-1 ถึง 2-3 โดยที่ R แทนพื้นที่ที่สนใจ S แทนพิกเซลทั้งหมดที่พิจารณาและ P แทน logical predicate หมายถึงการพิจารณาพิกเซลใกล้เคียง ยกตัวอย่างเช่น หากค่าพิกเซลปัจจุบันมีค่าความเข้มสีคล้ายหรืออยู่ในช่วงขีดแบ่ง (threshold) ที่กำหนด นั่นคือ P(Ri) = TRUE แต่ทั้งนี้ค่า logical predicate ก็อาจขึ้นอยู่กับการพิจารณาด้วยวิธีการหรือค่าอื่น ๆ ได้ด้วย

$$R = \bigcup_{i=1}^{S} R_i , R_i \cap R_j = 0 , \quad i \neq j \tag{2-1}$$

$$P(R_i) = \text{TRUE} , \quad i = 1, 2, \ldots, S \tag{2-2}$$

$$P(R_i \cup R_j) = \text{FALSE} , \quad i \neq j, \quad R_i \ \text{adjacent to} \ R_j \tag{2-3}$$

สมการที่ 2-1 หมายถึงการแบ่งขอบเขตต้องสมบูรณ์ (completeness) โดยที่ทุก ๆ พิกเซลจะต้องอยู่ภายในขอบเขตใด ๆ และแต่ละขอบเขตจะต้องไม่ซ้อนทับกัน (disjointness) สมการที่ 2-2 หมายถึงคุณสมบัติหรือความเข้มสีในระดับสีเทาของพิกเซลใด ๆ ที่อยู่ในขอบเขตเดียวกันจะต้องมีคุณสมบัติที่คล้ายกัน (satisfiability) เช่น ความเข้มสีใกล้เคียงกัน และสมการที่ 2-3 หมายถึงขอบเขตใด ๆ ที่ทำการแบ่งแยกแล้วจะแบ่งแยกจากกันอย่างชัดเจน หรือ แต่ละขอบเขตภายหลังการขยายพื้นที่ของส่วนภาพจะต้องแบ่งแยกกันอย่างสิ้นเชิง (segmentability) ซึ่งในบางกรณี เช่น การแบ่งขอบเขตภาพก้อนเนื้อในเต้านม บางขอบเขตที่อยู่ติดกันหรือมีความห่างเพียงเล็กน้อยก็สามารถนำขอบเขตเหล่านี้มาผสานกันได้ เพื่อความสะดวกในการนำไปดึงลักษณะสำคัญต่อไป

การขยายส่วนพื้นที่ในภาพจะใช้การพิจารณาค่าของพิกเซลปัจจุบันร่วมกับค่าของพิกเซลใกล้เคียง (neighborhood) โดยพิจารณาได้สองแบบ คือ การพิจารณาพิกเซลใกล้เคียงจำนวน 4 พิกเซล (4-neighberhoods) ดังแสดงในรูป 2.2(ก) และการพิจารณาพิกเซลใกล้เคียงจำนวน 8 พิกเซล (8-neighberhoods) ดังแสดงในรูป 2.2(ข)

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

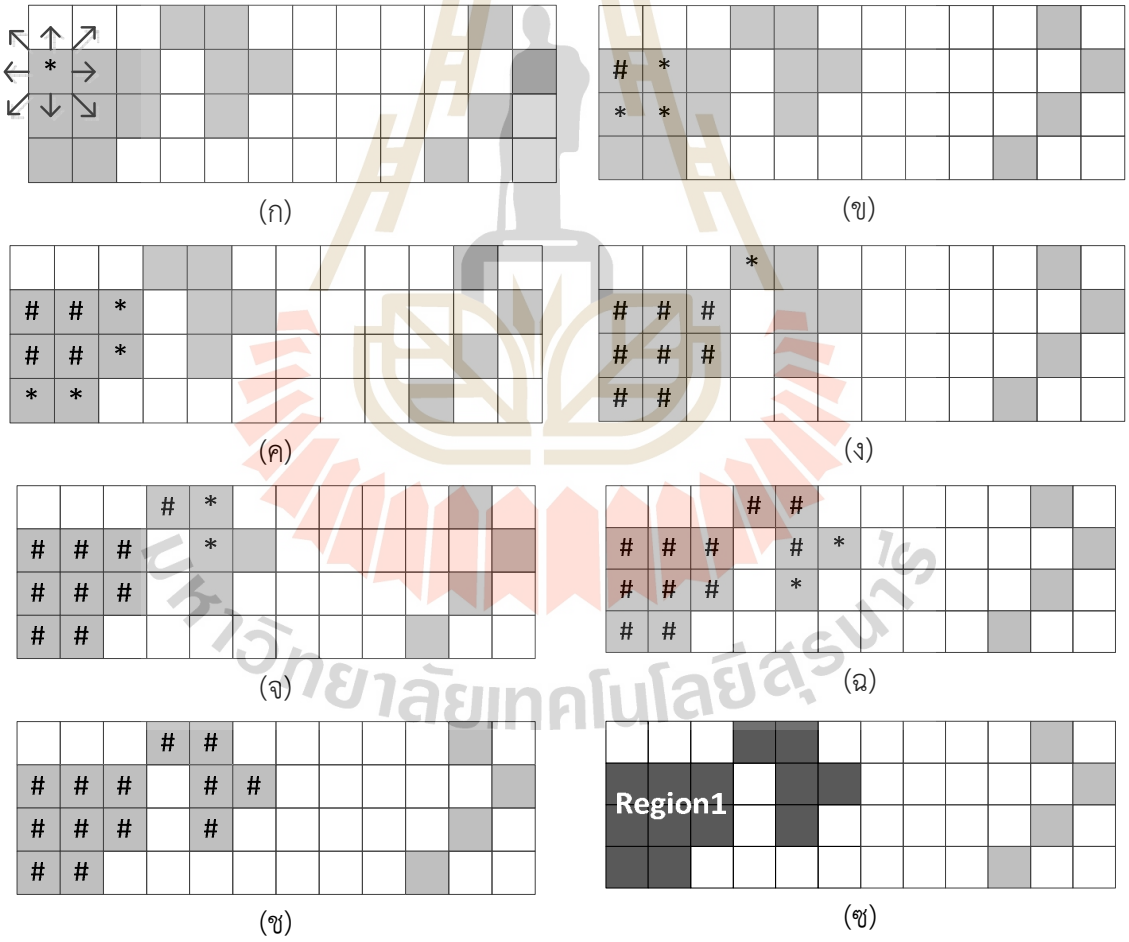
(ก) 4 พิกเซล

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

(ข) 8 พิกเซล

รูปที่ 2.2 ตำแหน่งการพิจารณาพิกเซลใกล้เคียง

ตัวอย่างการขยายส่วนพื้นที่ในรูปที่ 2.3 ใช้การพิจารณาพิกเซลใกล้เคียงจำนวน 8 พิกเซล เริ่มต้นที่รูป 2.3(ก) กำหนดจุดเริ่มต้นในการขยายส่วนพื้นที่โดยแทนด้วยเครื่องหมายดอกจัน (asterisk) จากนั้นพิจารณาพิกเซลข้างเคียงจำนวน 8 พิกเซลโดยเปรียบเทียบค่าพิกเซลปัจจุบันกับพิกเซลใกล้เคียงหากมีค่าความเข้มสีใกล้เคียงกัน หรือมีค่าความเข้มสีอยู่ในช่วงขีดแบ่งหรือ threshold ที่กำหนดก็ให้ทำเครื่องหมายชาร์ป (sharp) ในพิกเซลใกล้เคียงเหล่านั้น ดังแสดงในรูปที่ 2.3(ข) ถึง 2.3(ช) แต่หากพิกเซลใกล้เคียงใดที่มีความเข้มสีต่างกัน หรือมีค่าความเข้มสีไม่อยู่ในช่วงขีดแบ่ง ก็ไม่ต้องทำเครื่องหมายและไม่ต้องนำมาพิจารณา จากนั้นจึงทำกระบวนการเดิมซ้ำจนกระทั่งไม่มีจุดพิกเซลใกล้เคียงที่มีความเข้มสีคล้ายกันจึงหยุดการขยายพื้นที่ สุดท้ายจะได้บริเวณขอบเขตที่สนใจโดยใช้วิธีการขยายพื้นที่ดังแสดงในรูป 2.3(ช) แต่หากยังมีขอบเขตอื่น ๆ ในภาพที่ยังไม่ได้ทำการขยายพื้นที่ก็ทำตามกระบวนการเดิมคือกำหนดจุดเริ่มต้นของขอบเขตอื่น ๆ และพิจารณาพิกเซลใกล้เคียงต่อไปจนครบทั้งภาพ



รูปที่ 2.3 ตัวอย่างการขยายส่วนพื้นที่โดยพิจารณาพิกเซลใกล้เคียง 8 พิกเซล

การหาลักษณะสำคัญของภาพ

ลักษณะสำคัญของภาพที่ใช้ในงานวิจัยนี้ ประกอบด้วย ลักษณะลำคัญของลวดลาย (texture feature) ลักษณะสำคัญของฮิสโตแกรม (histogram based feature) และลักษณะสำคัญของรูปร่าง (shape feature) รายละเอียดของแต่ละเทคนิคอธิบายได้ดังนี้

- ลักษณะลำคัญของลวดลาย (texture feature)

ลวดลายภายในภาพเป็นหนึ่งในลักษณะสำคัญที่ใช้ในการระบุวัตถุหรือขอบเขตที่น่าสนใจในภาพ ลักษณะสำคัญหรือคุณลักษณะของลวดลายนั้นจะเป็นข้อมูลที่เกี่ยวกับการกระจายของรูปแบบโทนสี (tone) ภายในภาพ ดังนั้นลักษณะการเปลี่ยนแปลงของโทนสีภายในภาพจึงเป็นข้อมูลสำคัญซึ่งนำมาใช้ในการจำแนกภาพได้ ลักษณะสำคัญของลวดลายนั้นสามารถหาได้จากเมตริกซ์ของระดับสีเทาที่เกิดขึ้นร่วมกัน (Grey-level Co-occurrence Matrix : GLCM) ฟังก์ชันใน GLCM นั้นจะทำการคำนวณและเปรียบเทียบการเกิดขึ้นของระดับสีเทาในภาพหรือรูปแบบ (pattern) ของระดับสีเทาระหว่างพิกเซลในภาพ โดยใช้ความน่าจะเป็นในการแสดงผลของความชัดเจนของลวดลาย (contrast) การเกิดขึ้นร่วมกันของลวดลาย (correlation) และการเป็นเนื้อเดียวกันของลวดลาย (homogeneity) ซึ่งลักษณะสำคัญของลวดลายเหล่านี้จะถูกนำไปใช้ในกระบวนการจำแนกภาพ

- ลักษณะสำคัญของฮิสโตแกรม (histogram based feature)

รูปร่างลักษณะและคุณสมบัติของฮิสโตแกรมเป็นลักษะสำคัญอีกประเภทหนึ่งที่นิยมนำมาใช้เป็นลักษณะสำคัญในการจำแนกภาพ ซึ่งกราฟฮิสโตแกรมจะให้ข้อมูลสถิติของความเข้มสีที่เกิดขึ้นในภาพ และสามารถหาความน่าจะเป็นของความเข้มสีระดับสีเทาที่เกิดขึ้นในภาพ กราฟฮิสโตแกรมนั้นมีคุณลักษณะสำคัญทางสถิติ (statistic feature) ทั้งหมด 4 ค่า ได้แก่ (1) ค่าเฉลี่ย (mean) คือ ค่าเฉลี่ยความเข้มสี (2) ค่าความแปรปรวน (variance) คือ การเปลี่ยนแปลงความเข้มสีรอบค่าเฉลี่ย (3) ความเบ้ (skewness) คือ ค่าที่แสดงถึงความสมมาตรของฮิสโตแกรม ถ้าฮิสโตแกรมสมมาตรแล้วความเบ้จะมีค่าเป็น 0 ถ้าฮิสโตแกรมเบ้ขวา ค่าความเบ้จะเป็นบวก และถ้าฮิสโตแกรมเบ้ซ้าย ค่าความเบ้จะเป็นค่าลบ และ (4) ความโด่ง (kurtosis) คือ ค่าที่วัดจุดสูงสุดและต่ำสุดภายในกราฟในฮิสโตแกรมซึ่งมีความสัมพันธ์กับการกระจายข้อมูลแบบปกติ (normal distribution)

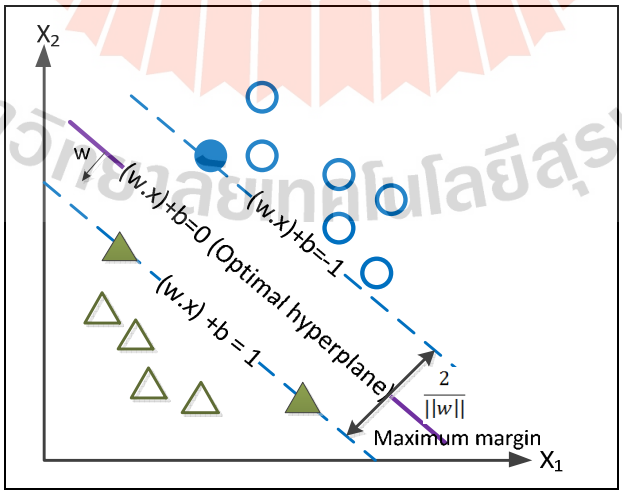
- ลักษณะสำคัญของรูปร่าง (shape feature)

การหาลักษณะสำคัญของรูปร่างนั้นเป็นองค์ประกอบที่สำคัญอย่างหนึ่งในการจำแนกภาพ สมมติว่าต้องการที่จะจำแนกภาพวัตถุสองชนิดซึ่งมีรูปร่างต่างกัน ลักษณะสำคัญของรูปร่างของวัตถุจะเป็นตัวระบุความแตกต่างของวัตถุแต่ละชนิด ลักษณะสำคัญของรูปร่างที่สำคัญได้แก่ พื้นที่ (area) เส้นผ่าศูนย์กลาง (diameter) ส่วนนูนของรูปร่าง (convex area) โครงร่าง (skeleton) เส้นรอบรูป (perimeter) ระยะทางจากจุดศูนย์กลางไปยังเส้นขอบ (centroid to distance) ในการ

จำแนกมะเร็งเต้านมจากภาพแมมโมแกรมนั้น โดยทั่วไปแล้วลักษณะรูปร่างของก้อนเนื้อที่ไม่เป็นอันตรายและก้อนเนื้อร้ายจะมีรูปร่างที่ต่างกัน คือ ก้อนเนื้อที่ไม่เป็นอันตรายจะมีลักษณะเป็นรูปร่างค่อนข้างกลมและขอบของก้อนเนื้อจะมีรอยหยักน้อย ซึ่งในทางกลับกันก้อนเนื้อร้ายที่มีแนวโน้มจะเป็นมะเร็งนั้นลักษณะรูปร่างจะไม่เป็นรูปทรง มีความบิดเบี้ยวและโค้งงอมาก และขอบของก้อนเนื้อจะมีรอยหยักมาก ดังนั้นในงานวิจัยนี้จึงเลือกใช้ลักษณะสำคัญของรูปร่าง โดยทำการวัดระยะทางจากจุดศูนย์กลางของก้อนเนื้อไปยังเส้นขอบ ซึ่งวิธีการนี้จะเป็นการดูความเปลี่ยนแปลงของความโค้งจากจุดศูนย์กลางไปยังเส้นขอบ ซึ่งถ้ามีการเปลี่ยนแปลงความโค้งจากจุดศูนย์กลางไปยังเส้นขอบน้อยก็สันนิษฐานได้ว่าเป็นก้อนเนื้อไม่อันตราย และหากมีการเปลี่ยนแปลงความโค้งจากจุดศูนย์กลางไปยังเส้นขอบค่อนข้างมากก็สันนิษฐานได้ว่าเป็นก้อนเนื้อที่มีโอกาสเป็นมะเร็ง

2.2 เทคนิคซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) เป็นวิธีการสำหรับจำแนกข้อมูลที่นิยมใช้อย่างแพร่หลายในปัจจุบัน ซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้แบบมีผู้แนะนำ (supervised learning) ซึ่งสามารถนำไปประยุกต์ใช้ได้กับปัญหาการจำแนกข้อมูล (data classification) และ การวิเคราะห์การถดถอย (regression analysis) การจำแนกโดยซัพพอร์ตเวกเตอร์แมชชีนนั้น มีหลักการพื้นฐานคือจะทำการสร้างไฮเปอร์เพลนที่มีขอบทั้งสองด้านกว้างมากที่สุด (maximum-margin hyperplane) เพื่อทำการจำแนกข้อมูลที่นำเข้ามา และในขณะเดียวกันเมื่อได้ไฮเปอร์เพลนที่มีขอบทั้งสองด้านกว้างมากที่สุดแล้ว เส้นของขอบแต่ละด้านนั้นจะต้องตัดผ่านหรือครอบคลุมข้อมูลนำเข้าให้น้อยที่สุด ดังนั้นการหาสมการไฮเปอร์เพลนและขนาดของมาร์จินที่เหมาะสม จะทำให้สามารถจำแนกข้อมูลได้อย่างมีประสิทธิภาพ ดังแสดงในรูปที่ 2.4



รูปที่ 2.4 การจำแนกข้อมูล 2 คลาส ด้วยซัพพอร์ตเวกเตอร์แมชชีน

2.3 งานวิจัยที่เกี่ยวข้อง

การวิเคราะห์ภาพแมมโมแกรมเพื่อจำแนกมะเร็งเต้านมนั้นเป็นงานวิจัยที่น่าสนใจ เนื่องจากการวิเคราะห์ด้วยภาพนั้นเป็นขั้นตอนเบื้องต้นที่มีประสิทธิภาพและไม่เป็นอันตราย ช่วยให้การวิเคราะห์สะดวกขึ้นโดยคนไข้ที่มีก้อนเนื้อผิดปกติในบริเวณทรวงอกและต้องการทราบว่าตนป่วยเป็นมะเร็งเต้านมหรือไม่นั้นสามารถเข้ารับการตรวจวิเคราะห์ได้ โดยไม่ต้องเข้ารับการผ่าตัดชิ้นเนื้อ ดังนั้นจึงมีงานวิจัยที่หลากหลายในการพยายามที่จะวิเคราะห์และจำแนกมะเร็งเต้านมจากภาพให้มีความแม่นยำและมีประสิทธิภาพสูง

จากความรู้ของนักรังสีวิทยาที่สามารถวิเคราะห์จากลักษณะรูปร่างของก้อนเนื้อในทรวงอก เพื่อระบุว่าเป็นก้อนเนื้อดีและก้อนเนื้อร้ายนั้น เป็นการสังเกตจากรูปร่าง ลักษณะ และความหนาแน่นที่แตกต่างกันของภาพก้อนเนื้อ แต่ในหลายปีที่ผ่านมาการวิเคราะห์ของนักรังสีวิทยาอาจมีความผิดพลาดได้ การวิเคราะห์นั้นจึงต้องใช้ประสบการณ์ การฝึกฝน และความรู้เฉพาะทาง แต่ถึงกระนั้นก็ยังมีการสำรวจพบว่าประมาณ 10% ของก้อนเนื้อร้ายทั้งหมดในภาพแมมโมแกรม ถูกอ่านและวิเคราะห์ผิดโดยนักรังสีวิทยา ซึ่งทำให้มีค่า false positive ค่อนข้างสูง และก้อนเนื้อส่วนใหญ่ที่มีการวิเคราะห์ผิดพลาดนั้นจะอยู่ในทรวงอกที่มีความหนาแน่นสูง (Jackson et al., 1993) ยกตัวอย่างเช่นมีการวิเคราะห์ว่าก้อนเนื้อที่สงสัยนั้นเป็นก้อนเนื้อซึ่งมีความเสี่ยงในการเป็นมะเร็งเต้านม แต่เมื่อผ่าตัดเพื่อนำก้อนเนื้อไปพิสูจน์แล้ว กลับได้ผลว่าไม่ใช่ก้อนเนื้อร้าย ซึ่งความผิดพลาดจากการวิเคราะห์ภาพแมมโมแกรมของนักรังสีวิทยานั้นเกิดจากสาเหตุสำคัญ 3 ประการ คือ (1) ลักษณะเด่นของก้อนเนื้อจากภาพแมมโมแกรมที่จะแปรผลนั้นมีความไม่ชัดเจน (2) สัญญาณรบกวนภายในภาพแมมโมแกรมซึ่งอาจเกิดจากคุณภาพของเครื่องเอ็กซเรย์และ(3) ลักษณะสำคัญบางประการของภาพยังคลุมเครือและไม่สามารถวิเคราะห์ได้ด้วยตาเปล่า (Sivaramakrishna et al., 2002)

ในไม่กี่ปีผ่านมานี้มีการนำเสนอวิธีการหลากหลายวิธีในการจำแนกภาพแมมโมแกรมโดยนำคอมพิวเตอร์มาใช้ในการวิเคราะห์ภาพแมมโมแกรม (mammographic computer-aided diagnosis: CAD) ทีมวิจัยหลายทีมได้ทำการเสนอระบบ CAD ในการช่วยวิเคราะห์ภาพซึ่งได้ผลในการวิเคราะห์ที่ดีและมีประสิทธิภาพ (Cheng et al., 2006; Rangayyan et al., 2007; Elter et al., 2009) ทีมวิจัยของ Oliver (2010) ได้เสนอการจำแนกความหนาแน่นของก้อนเนื้อในทรวงอกด้วยวิธีการทางสถิติ (statistical for breast density segmentation) โดยพิจารณาความหนาแน่นของจุดพิกเซลภายในภาพ วิธีการนี้ทำการพิจารณาลักษณะเฉพาะของเนื้อเยื่อในภาพแมมโมแกรม และทำการจำแนกภาพออกเป็นบริเวณที่เป็นไขมันและบริเวณที่เป็นเนื้อเยื่อที่มีความหนาแน่นสูง โดยใช้เทคนิค PCA และ LDA ในการจำแนกภาพ และนำผลมาเปรียบเทียบกัน

การพิจารณาขอบเขตที่สนใจ (Region of interest : ROI) เป็นอีกวิธีหนึ่งที่นำมาใช้ในการแบ่งขอบเขตของภาพแมมโมแกรมโดยตัดเอามาเฉพาะขอบเขตของก้อนเนื้อที่มีความหนาแน่นสูง

โดยทำการพิจารณาความเข้มสีของพิกเซลข้างเคียง โดยวิธีการนี้จะทำการลดเวลาในการประมวลผล และเพิ่มความแม่นตรงในกระบวนการจำแนกเนื่องจาก การพิจารณาเพราะบริเวณ ROI นั้น เป็นการนำบางส่วนของภาพเฉพาะบริเวณก้อนเนื้อมาพิจารณาเท่านั้นโดยไม่จำเป็นต้องใช้ภาพแมมโมแกรมขนาดใหญ่ วิธีการ region growing เป็นเทคนิคที่ใช้กันอย่างกว้างขวางในการแบ่งส่วนของภาพแมมโมแกรม โดยทำการแบ่งส่วนของก้อนเนื้อออกมาจากพื้นหลัง Huoและคณะ (1995) ได้ทำการพัฒนาเทคนิค semi-automatic region growing ขึ้นมา โดยขั้นตอนแรกจะต้องมีการเลือกจุด seed point โดยนักรังสีวิทยาก่อน แล้วจึงทำ region growing ตามขั้นตอนปกติ ซึ่งวิธีการนี้ยังเป็นการทำ region growing กึ่งอัตโนมัติเพราะต้องมีการคัดเลือก seed point โดยนักรังสีวิทยา

วิธีการ feature extraction เป็นเทคนิคที่นำเอาลักษณะที่สำคัญของ ROI จากภาพแมมโมแกรมออกมาก่อนที่จะส่งไปให้ classifier จำแนก Jiang และคณะ (1998) ได้นำ feature ที่เกี่ยวกับลักษณะทางสัณฐานวิทยา (morphological feature) มาใช้ โดย feature ประเภทนี้จะบ่งบอกถึงลักษณะของรูปร่าง ความโค้ง ความบิดเบี้ยว เส้นรอบวง เส้นผ่าศูนย์กลาง และพื้นที่ของ ROI นอกจากนี้ feature อีกประเภทหนึ่งที่นิยมใช้มากสำหรับภาพแมมโมแกรมคือ texture feature โดยได้ถูกนำเสนอในงานวิจัยของ Kegelmeyer และคณะ (1994) ซึ่งพบว่า texture feature ก็เป็น feature อีกประเภทหนึ่งที่เมื่อนำไปสู่ขั้นตอนการจำแนก (classification) แล้ว ทำให้ได้ประสิทธิภาพในการจำแนกที่ดีขึ้น

ในขั้นตอนการจำแนก Campaniniและคณะ (2004) ใช้ซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกภาพแมมโมแกรม โดยใช้ค่า grey level ของ ROI จากตัวอย่างที่เป็นเนื้อเยื่อที่มีความหนาแน่น และเนื้อเยื่อปกติ นอกจากการใช้ซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกแล้ว neural network ก็เป็น classifier อีกตัวหนึ่งที่มีประสิทธิภาพ ยกตัวอย่างเช่น Stathaki และคณะ (1994) ได้ใช้ neural network ในการจำแนกภาพแมมโมแกรมโดยใช้ feature ที่ได้จาก autoregressive model แบบสองมิติ Christoyianni และคณะ (2000) ได้ใช้ radial-based function neural network ทำการจำแนก feature ที่ได้จาก histogram ของภาพ ROI แต่ละภาพ

จากการทบทวนวรรณกรรมที่เกี่ยวข้อง พบว่างานวิจัยที่เกี่ยวข้องกับการจำแนกมะเร็งเต้านมด้วยการวิเคราะห์ภาพแมมโมแกรม กลุ่มนักวิจัยส่วนใหญ่จะเสนอแนวคิดไปในทางเดียวกันคือมีการทำการปรับปรุงภาพก่อนด้วยวิธีการประมวลผลภาพที่หลากหลาย เช่น การกำจัดสัญญาณรบกวนภายในภาพและการปรับความเข้มสี หลังจากนั้นจึงนำภาพที่ปรับปรุงแล้วมาทำการแบ่งแยกขอบเขตเพื่อจะพิจารณาเฉพาะบริเวณก้อนเนื้อซึ่งโดยปกติจะมีความเข้มสีมากกว่าพื้นหลัง เช่นการหา ROI โดยวิธี region growing ต่อจากนั้นจึงนำ ROI ที่ได้ ไปเข้ากระบวนการ feature extraction ซึ่งเป็นการนำเอาลักษณะเฉพาะของ ROI ออกจากพื้นหลังอื่น ๆ เช่น morphological feature, texture feature เป็นต้น ส่วนขั้นตอนสุดท้ายคือกระบวนการ classification ซึ่งจากการทบทวน

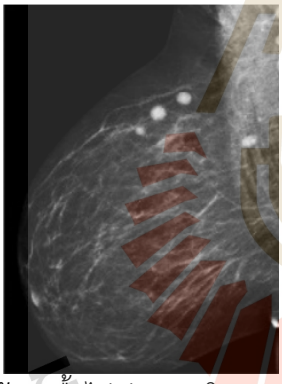
วรรณกรรมที่เกี่ยวข้องการจำแนกมะเร็งเต้านมพบว่า นักวิจัยส่วนใหญ่จะใช้ซัพพอร์ตเวกเตอร์แมชชีนและนิวรอลเน็ตเวิร์กในการจำแนก

ในงานวิจัยนี้มีวัตถุประสงค์เช่นเดียวกับงานวิจัยอื่น ๆ ที่กล่าวถึงข้างต้น แต่แนวทางการปรับปรุงภาพและการจำแนกภาพ จะมีความแตกต่างจากงานอื่นตรงที่การคัดเลือกลักษณะสำคัญในภาพจะมีการพิจารณาส่วนประกอบเส้นโค้งของภาพที่เป็นลักษณะเฉพาะของภาพมะเร็งเต้านม ลักษณะเฉพาะนี้จะเป็นปัจจัยสำคัญให้การจำแนกภาพด้วยซัพพอร์ตเวกเตอร์แมชชีนมีความแม่นยำสูงขึ้น

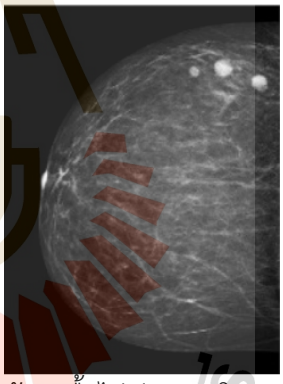
บทที่ 3
การออกแบบและพัฒนาวิธีการวิเคราะห์ภาพแมมโมแกรมเพื่อวินิจฉัยมะเร็งเต้านม

3.1 ลักษณะของภาพแมมโมแกรม

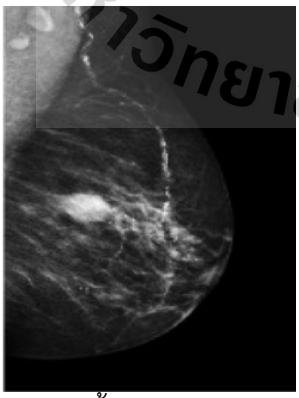
ภาพแมมโมแกรมเป็นภาพถ่ายทางรังสีที่แพทย์ใช้ช่วยในการวินิจฉัยอาการผิดปกติ เช่น มะเร็งเต้านมและพยาธิสภาพอื่น ๆ ในงานวิจัยนี้เน้นที่ภาพแมมโมแกรมที่เกี่ยวข้องกับมะเร็งเต้านม โดยใช้ภาพจาก University of South Florida Digital Mammography ชื่อชุดข้อมูล Digital Database for Screening Mammography (http://marathon.csee.usf.edu/Mammography/Database.html) ข้อมูลนี้ได้มาจากกลุ่มตัวอย่างคนไข้ 2,500 คน โดยมีการเก็บภาพแมมโมแกรมของเต้านมด้านซ้ายและด้านขวาในมุมมอง 2 แบบ คือ มุมมองแบบ MLO (mediolateral view) ซึ่งเป็นการถ่ายภาพรังสีในแนวขวางลำตัว เช่น จากด้านขวาไปด้านซ้าย และ มุมมองแบบ CC (caudocranial view) ที่เป็นการถ่ายภาพรังสีในแนวดิ่งจากส่วนบนลำตัวพุ่งลงด้านล่าง โดยมีทั้งภาพแมมโมแกรมของผู้ป่วยที่มีก้อนเนื้อที่ไม่เป็นอันตรายและก้อนเนื้อร้าย ดังแสดงตัวอย่างในรูปที่ 3.1



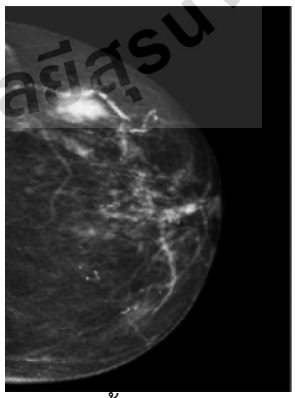
(ก) ก้อนเนื้อไม่อันตรายในมุมมอง MLO



(ข) ก้อนเนื้อไม่อันตรายในมุมมอง CC



(ค) ก้อนเนื้อมะเร็งในมุมมอง MLO



(ง) ก้อนเนื้อมะเร็งในมุมมอง CC

รูปที่ 3.1 ภาพแมมโมแกรมในมุมมอง MLO และ CC

3.2 กรอบแนวคิดของงานวิจัย

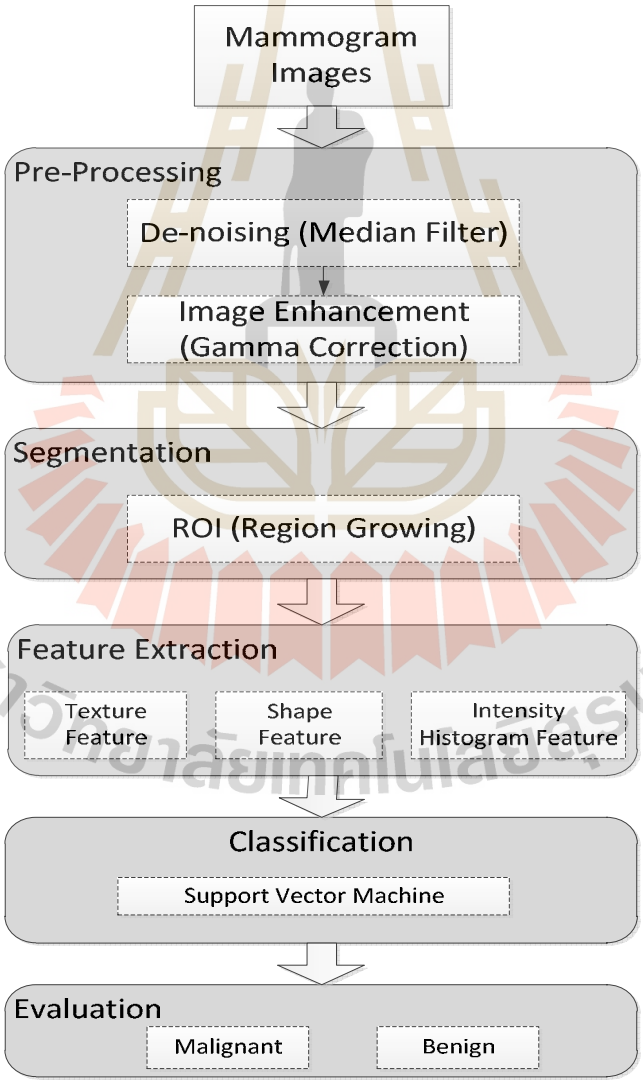
กรอบแนวคิดหลักของงานวิจัยนี้ คือ การพัฒนาเทคนิคการวิเคราะห์ภาพแมมโมแกรมเพื่อจำแนกมะเร็งเต้านมได้โดยอัตโนมัติ จุดมุ่งหมายหลักต้องการช่วยสนับสนุนการวิเคราะห์และวินิจฉัยโรคของรังสีแพทย์ โครงสร้างพื้นฐานของระบบที่จะพัฒนาขึ้นนี้ประกอบด้วย 3 ส่วนประกอบหลัก คือ

(1) ส่วนการปรับปรุงภาพ และการดึงฟีเจอร์ที่สำคัญของภาพแมมโมแกรม

(2) ส่วนของโมเดลที่ใช้ในการจำแนก

(3) ส่วนการประเมินผลการจำแนกเพื่อวัดประสิทธิภาพและความแม่นตรงในการจำแนก

งานวิจัยนี้ได้พัฒนาเพิ่มเติมความสามารถด้านการนำฟีเจอร์หรือลักษณะที่สำคัญของภาพแมมโมแกรมและการใช้โมเดลการจำแนกที่มีประสิทธิภาพ โครงสร้างระบบแสดงได้ดังรูปที่ 3.2



รูปที่ 3.2 กรอบการวิจัยการพัฒนาเทคนิคการวิเคราะห์ภาพแมมโมแกรมเพื่อจำแนกมะเร็งเต้านม

แต่ละส่วนประกอบของระบบอธิบายได้ดังนี้

- โมดูล Image Pre-processing ทำหน้าที่ปรับปรุงภาพโดยใช้วิธีการกำจัดสัญญาณรบกวนภายในภาพด้วยเทคนิค median filter และทำการปรับปรุงความชัดเจนโดยเพิ่มความเข้มสีในบริเวณที่คาดว่าเป็นก้อนเนื้อเพื่อทำให้เห็นภาพบริเวณก้อนเนื้อได้ชัดเจนยิ่งขึ้น ในขณะเดียวกันก็ทำการปรับลดความเข้มสีในบริเวณที่เป็นพื้นหลังลงด้วย โดยใช้เทคนิค gamma correction
- โมดูล Segmentation เป็นส่วนที่ทำหน้าในการดึงเฉพาะส่วนหรือบริเวณที่คาดว่าจะเป็นก้อนเนื้อออกมา (region of interest: ROI) โดยทำการตัด (cropped) เฉพาะบริเวณก้อนเนื้อ โดยขั้นตอนนี้ใช้วิธีการดูค่าความเข้มสีและเปรียบเทียบกับพิกเซลข้างเคียง โดยใช้เทคนิค region growing
- โมดูล Feature extraction ทำหน้าที่ในการดึงเอา feature ต่าง ๆ ที่สนใจออกมาจาก ROI ที่ได้ตัดออกมาจากภาพแมมโมแกรมแล้ว โดยในโมดูลนี้จะทำการดึง feature มาทั้งหมด 3 ประเภท คือ texture feature, shape feature และ intensity histogram feature
- โมดูล Classification เป็นส่วนที่ทำหน้าที่ในการจำแนกภาพแมมโมแกรมว่าเป็น benign หรือ malignant โดยใช้ feature ทั้ง 3 ประเภทจากโมดูล Feature extraction นำมาสร้างโมเดลสำหรับการจำแนก สำหรับโมดูลนี้จะใช้เทคนิค support vector machine ในการจำแนก
- โมดูล Evaluation ทำหน้าที่ ประเมินความแม่นตรงของโมเดล ในการจำแนกภาพแมมโมแกรม โดยทำการแสดงผลทั้งในแบบตาราง และ กราฟ โดยทำการเปรียบเทียบกับข้อมูลทดสอบ

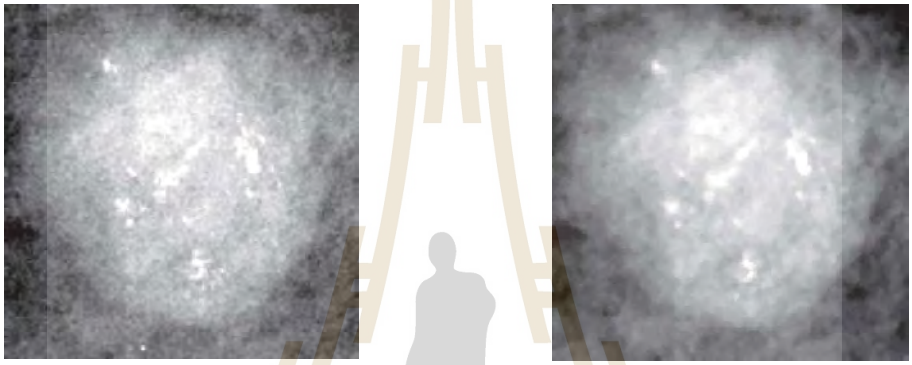
3.3 ขั้นตอนการดำเนินงานวิจัย

จากกรอบแนวคิดหลัก สามารถจำแนกการดำเนินงานออกเป็นขั้นตอนต่าง ๆ ได้ดังนี้

3.3.1 การปรับปรุงภาพแมมโมแกรม

โดยปกติแล้วภาพแมมโมแกรมมักจะมีสัญญาณรบกวนซึ่งเกิดจากคุณภาพของอุปกรณ์ในการรับภาพทำให้ภาพไม่ชัดเจน สัญญาณรบกวนในภาพแมมโมแกรมนั้นมี 2 แบบ คือ สัญญาณรบกวนแบบเกาเซียน (Gaussian noise) และสัญญาณรบกวนที่เป็นจุดดำหรือจุดขาวเล็ก ๆ หรือเรียกว่าสัญญาณรบกวนแบบเกลือและพริกไทย (salt and pepper noise) ในขั้นตอนนี้จะเป็นการ

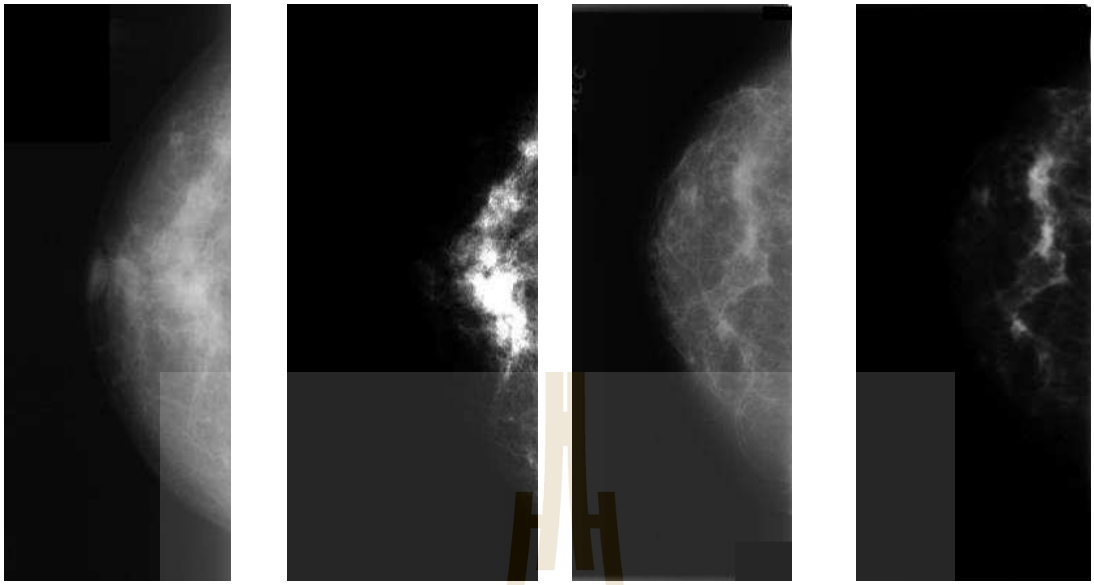
ปรับปรุงภาพโดยใช้วิธีการกำจัดสัญญาณรบกวนภายในภาพด้วยวิธีมีเดียนฟิลเตอร์ (median filter) หลักการของมีเดียนฟิลเตอร์คือการใช้หน้าต่างขนาดเล็กเช่น 3×3 พิกเซลหรือ 5×5 พิกเซล เลื่อนไปบนภาพที่ต้องการกำจัดสัญญาณรบกวน โดยในขณะที่เลื่อนนั้นหน้าต่างขนาดเล็กทำหน้าที่คำนวณและเปลี่ยนแปลงค่าพิกเซล ณ จุดใด ๆ โดยเรียงค่าความเข้มสีของบริเวณหน้าต่างที่ครอบภาพจากค่าน้อยไปมาก จากนั้นจึงคัดเลือกค่ามัธยฐาน แล้วนำค่ามัธยฐานแทนที่ลงในพิกเซลปัจจุบัน ดังนั้นเมื่อภาพผ่านกระบวนการของมีเดียนฟิลเตอร์แล้ว ภาพจะมีความชัดเจนขึ้นโดยยังรักษาความคมชัดและขอบของภาพไว้ได้ ดังแสดงตัวอย่างในรูปที่ 3.3



(ก) ภาพก่อนผ่านมีเดียนฟิลเตอร์ (ข) ภาพหลังผ่านการปรับปรุงด้วยมีเดียนฟิลเตอร์

รูปที่ 3.3 ภาพก่อนและหลังการปรับปรุงด้วยมีเดียนฟิลเตอร์

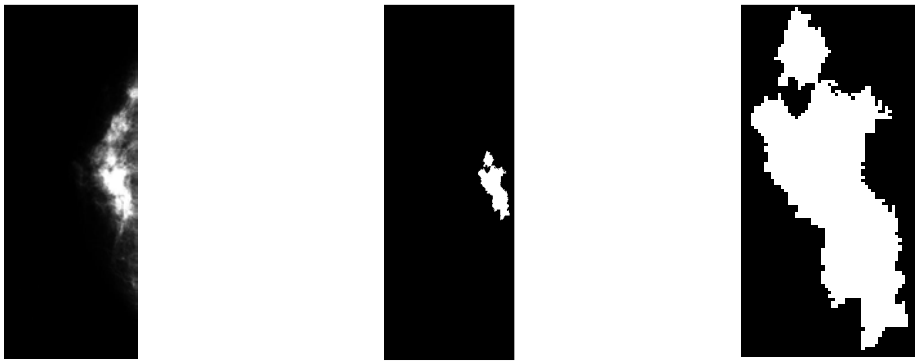
ขั้นตอนต่อไปจะเป็นการปรับปรุงความชัดเจนของภาพ โดยเพิ่มความเข้มสีในบริเวณที่คาดว่าเป็นก้อนเนื้อเพื่อทำให้เห็นภาพบริเวณก้อนเนื้อได้ชัดเจนยิ่งขึ้น ในขณะเดียวกันก็ทำการปรับลดความเข้มสีในบริเวณที่เป็นพื้นหลังลงด้วย ในขั้นตอนนี้จะใช้วิธีแก้ไขแกมมา (gamma correction) จากรูปที่ 3.4 แสดงภาพก้อนเนื้อในเต้านมทั้งในกรณีที่เป็นก้อนเนื้อร้ายและก้อนเนื้อไม่อันตราย โดยเมื่อเปรียบเทียบระหว่างรูป 3.4(ก) - 3.4(ข) และ 3.4(ค) - 3.4(ง) แล้วจะเห็นว่าการแก้ไขแกมมาช่วยให้ความเข้มสีในบริเวณที่สว่างยิ่งมีความเข้มสีที่มากขึ้น และในทางตรงกันข้ามบริเวณที่เป็นพื้นหลังที่มีความเข้มสีที่ค่อนข้างมืดก็จะถูกปรับลดความเข้มสีลง ส่งผลให้บริเวณที่เป็นก้อนเนื้อมีความชัดเจนมากขึ้น ลักษณะนี้สอดคล้องกับความคิดเห็นของนักรังสีวิทยาที่ได้กล่าวไว้ว่า บริเวณที่เป็นก้อนเนื้อที่มีความผิดปกติ ความเข้มสีบริเวณนั้นจะมีมากกว่าบริเวณอื่น ๆ ซึ่งวิธีการนี้จะเป็นประโยชน์ในการนำภาพเข้าสู่กระบวนการแบ่งขอบเขตภาพหรือ ROI ในลำดับต่อไป



(ก) (ข) (ค) (ง)

รูปที่ 3.4 ภาพก้อนเนื้อในเต้านม (ก) ภาพก้อนเนื้อร้าย (ข) ภาพก้อนเนื้อร้ายหลังจากปรับปรุงด้วยแกมมาคอเร็คชัน (ค) ภาพก้อนเนื้อไม่อันตราย (ง) ภาพก้อนเนื้อไม่อันตรายหลังจากปรับปรุงด้วยการแก้ไขแกมมา

ก่อนนำภาพไปเข้าสู่กระบวนการจำแนกแบบอัตโนมัติ ภาพนั้นจะต้องผ่านอีกขั้นตอนที่สำคัญคือขั้นตอนการแบ่งขอบเขตภาพ ขั้นตอนนี้มีความจำเป็นเนื่องจากภาพแมมโมแกรมมีขนาดใหญ่ประมาณ 3000×4000 พิกเซล แต่จุดที่แพทย์สนใจจะเป็นเพียงบริเวณก้อนเนื้อที่มีความผิดปกติเท่านั้น ดังนั้นพื้นหลังในภาพแมมโมแกรมและบริเวณที่เป็นส่วนไขมัน จึงไม่จำเป็นต้องนำเข้าสู่กระบวนการจำแนก กระบวนการแบ่งขอบเขตภาพนี้จะช่วยลดขนาดข้อมูล ทำให้การจำแนกทำได้รวดเร็วขึ้นและไม่เกิดปัญหาหน่วยความจำเต็ม ในขั้นตอนนี้เราจะใช้วิธีการการขยายพื้นที่ของส่วนภาพ (region growing) ซึ่งวิธีนี้เป็นวิธีที่นิยมใช้ในการแบ่งส่วนภาพ โดยในขั้นตอนแรกนั้นจะทำการคัดเลือกจุดกึ่งกลางของก้อนเนื้อ (seed point) โดยใช้วิธีการหาจุดเซนทรอยด์ (centroid) จากก้อนเนื้อในภาพ เมื่อพบจุดภาพที่เป็นบริเวณขอบของจุดกึ่งกลางหรือ seed region ก็จะพิจารณาจุดภาพข้างเคียง (neighbor) ด้วยการวางหน้าต่างขนาด 3×3 ครอบจุดภาพนั้น หากจุดภาพข้างเคียงใดมีค่าระดับสีเทาอยู่ในขอบเขตของการขยายพื้นที่ ก็จะทำการรวมหรือขยายพื้นที่ส่วนภาพไปยังจุดข้างเคียงนั้น แต่ถ้าค่าไม่ใกล้เคียง ก็จะพิจารณาจุดข้างเคียงถัดไป กระบวนการขยายส่วนพื้นที่ของภาพนี้ จะกระทำกับทุก ๆ seed region แบบวนซ้ำจนกระทั่งไม่สามารถขยายพื้นที่ได้ ผลจากการใช้กระบวนการขยายส่วนพื้นที่ในภาพแมมโมแกรมที่ผ่านการแก้ไขแกมมาแล้ว แสดงได้ดังรูปที่ 3.5



(ก) (ข) (ค)

รูปที่ 3.5 ผลลัพธ์ของการขยายส่วนพื้นที่ของภาพแมมโมแกรม (ก) ภาพที่ได้จากกระบวนการแก้ไขแกมมา (ข) ภาพหลังจากขยายส่วนพื้นที่ (ค) ภาพที่ตัดเฉพาะบริเวณก้อนเนื้อ

หลังจากได้ ROI จากการทำการขยายส่วนพื้นที่ของภาพมาแล้ว หากนำเพียงค่าความเข้มสีไปเข้ากระบวนการจำแนกอาจทำให้การจำแนกได้ผลไม่ดีนัก ดังนั้นการดึงลักษณะสำคัญในภาพโดยนำคุณสมบัติของ ROI ที่ตัดออกมาจากภาพแมมโมแกรมมาหาลักษณะสำคัญ จะเป็นวิธีการที่ช่วยให้ประสิทธิภาพในการจำแนกดีขึ้น สำหรับในงานวิจัยนี้ได้ใช้ลักษณะสำคัญทั้งหมด 3ลักษณะคือ ลักษณะสำคัญของลวดลายภายในก้อนเนื้อ (texture feature) ลักษณะสำคัญของรูปร่างก้อนเนื้อ (shape feature) และลักษณะสำคัญของความเข้มสีของก้อนเนื้อ (intensity histogram feature)

ลักษณะสำคัญของลวดลายหรือ texture feature เป็นการหารูปแบบของลวดลายจาก ROI โดยเทคนิคที่นิยมใช้คือ Gray-Level Co-occurrence Matrix (GLCM) ฟังก์ชันใน GLCM จะทำการคำนวณและเปรียบเทียบการเกิดขึ้นของลวดลายหรือแพทเทิร์นระหว่างพิกเซลในภาพ โดยใช้ความน่าจะเป็นในการแสดงผลของความชัดเจนของลวดลาย (contrast) การเกิดขึ้นร่วมกันของลวดลาย (correlation) รวมทั้งวัดค่าความเป็นเนื้อเดียวกัน (homogeneity) ของลวดลายในภาพ แสดงตัวอย่างค่าที่ได้จากการคำนวณในตารางที่ 3.1

ตารางที่ 3.1 ตัวอย่างลักษณะสำคัญของลวดลายในภาพแมมโมแกรม

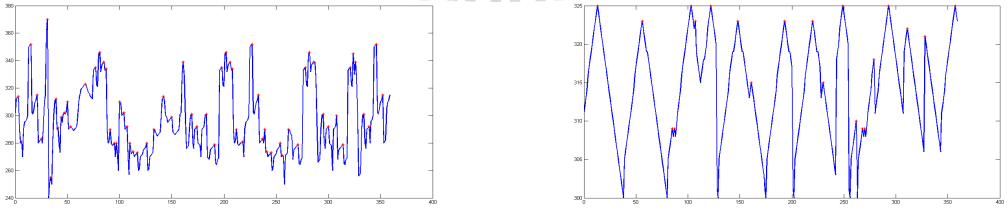
Direction	Homogeneity	Contrast	Correlation
0°	0.0125	3.0382	0.8074
45°	0.0082	4.0121	0.6369
90°	0.0075	4.0153	0.5988
135°	0.0069	4.7084	0.4613
Average	0.0087	3.9435	0.6261

ลักษณะสำคัญของรูปร่าง หรือ shape feature เป็นการวัดความโค้งหรือความหยักของ ROI เนื่องจากรังสีแพทย์ได้ให้ข้อสังเกตว่า รูปร่างของก้อนเนื้อไม่เป็นอันตรายนั้นจะมีรูปร่างที่ค่อนข้างเป็นวงกลมหรือวงรีที่มีขอบค่อนข้างเรียบหรือมีรอยหยักน้อย และในทางกลับกันรูปร่างของเนื้อร้ายที่จะก่อตัวเป็นมะเร็งหรือเนื้อร้ายที่เป็นมะเร็งแล้วนั้น จะมีรูปร่างที่ค่อนข้างบิดเบี้ยวและขอบของก้อนเนื้อจะมีรอยหยักค่อนข้างมาก ดังนั้นการนำเอาลักษณะสำคัญของรูปร่างขอบของ ROI มาเป็นลักษณะสำคัญหรือ feature ในการจำแนกจึงคาดว่าจะส่งผลดีต่อโมเดลในการจำแนกภาพ

รูปที่ 3.6 แสดงการวัดความโค้งของ ROI ที่เป็นก้อนเนื้อไม่อันตรายและก้อนเนื้อร้าย โดยทำการโยงเส้นจากจุดเซนทรอยด์ของ ROI และแสดงออกมาเป็นกราฟแสดงความโค้งงอของเส้นขอบ ROI โดยจะสังเกตเห็นว่าก้อนเนื้อร้ายนั้นจะมีความหยักหรือการเปลี่ยนแปลงความโค้งของขอบภาพมากกว่าก้อนเนื้อไม่อันตราย และรูปที่ 3.7 แสดงความแตกต่างของกราฟแสดงความหยักของเส้นขอบแสดงรูปร่างของก้อนเนื้อร้ายและก้อนเนื้อไม่อันตราย แกน X แทนองศา ตั้งแต่ 0 องศา ถึง 360 องศา และแกน Y แทนระยะทางที่วัดจากจุดกึ่งกลางไปเส้นขอบ โดยพิจารณาเฉพาะจำนวนจุดยอดของกราฟที่มีการเปลี่ยนแปลงให้มีค่าเป็น 1 ตามองศาการวัดระยะทางจากจุดเซนทรอยด์ไปยังเส้นขอบตั้งแต่ 0 ถึง 360 องศา ลักษณะของกราฟในรูปที่ 3.7 จะเป็น background knowledge ที่สำคัญในการจำแนกภาพมะเร็งเต้านม

(ก) (ข)

รูปที่ 3.6 การวัดความหยักของเส้นขอบโดยวัดจากจุดเซนทรอยด์ (ก) เส้นขอบแสดงรูปร่างของก้อนเนื้อร้าย (ข) เส้นขอบแสดงรูปร่างของก้อนเนื้อไม่อันตราย

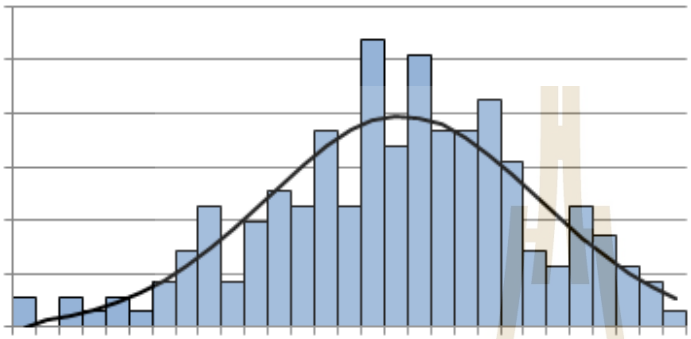


...10010011001010100100011011100 01… …10000000000010000000000001000000...

(ก) (ข)

รูปที่ 3.7 กราฟแสดงความหยักของ (ก) ก้อนเนื้อร้าย (ข) ก้อนเนื้อไม่อันตราย

ลักษณะสำคัญของภาพอีกอย่างหนึ่งคือลักษณะสำคัญของฮิสโตแกรม โดยเป็นค่าที่ใช้วัดลักษณะของเส้นโค้งแจกแจงความถี่ของความเข้มสีว่ามีลักษณะของเส้นโค้งเป็นลักษณะใด รูปที่ 3.8 แสดงตัวอย่างกราฟฮิสโตแกรมที่พิจารณาลักษณะสำคัญ 4 ค่า เพื่อใช้ในการอธิบายข้อมูลสถิติของความเข้มสีที่เกิดขึ้นในภาพ ได้แก่ ค่าเฉลี่ย ค่าความแปรปรวน ความเบ้ และ ความโด่ง



รูปที่ 3.8 ตัวอย่างกราฟฮิสโตแกรมที่พิจารณาลักษณะสำคัญ 4 ค่า

3.3.2 การจำแนกภาพด้วยซัพพอร์ตเวกเตอร์แมชชีน

ภาพแมมโมแกรมที่ผ่านการปรับปรุงภาพด้วยเทคนิคการประมวลผลภาพเพื่อดึงลักษณะสำคัญ 3 แบบมาจากภาพ (ลักษณะสำคัญของลวดลาย ลักษณะสำคัญของรูปร่าง และลักษณะสำคัญของความเข้มสีจากฮิสโตแกรม) จะถูกนำมาจำแนกเพื่อวินิจฉัยว่าเป็นภาพที่มีเนื้อร้ายหรือเป็นภาพปกติ การจำแนกจะใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนที่ใช้เคอร์เนลฟังก์ชันแบบเรเดียลเบสิส และในการเปรียบเทียบประสิทธิภาพจะใช้อัลกอริทึมนาอีฟเบย์และโครงข่ายประสาทเทียมมาเปรียบเทียบกับอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน โดยผลการทดลองจะนำเสนอในบทที่ 4

บทที่ 4
การทดสอบประสิทธิภาพของการวิเคราะห์ภาพแมมโมแกรม

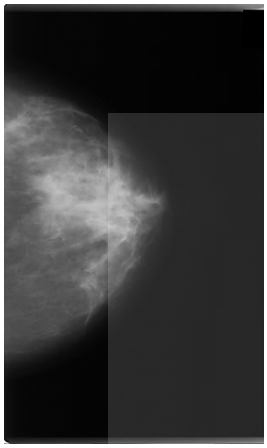
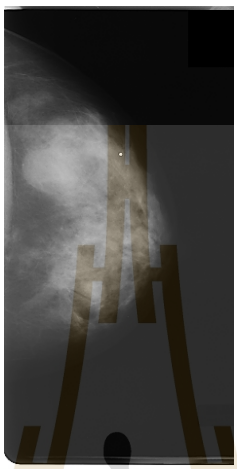
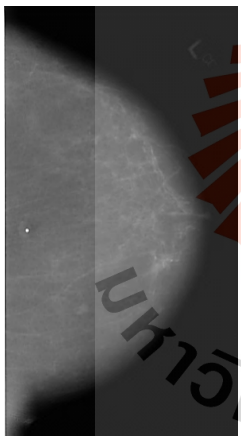
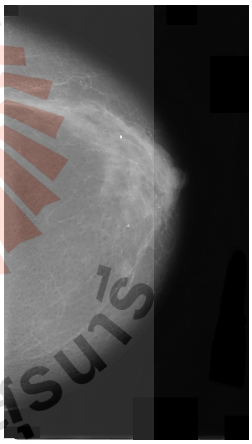
การทดสอบประสิทธิภาพของงานวิจัยนี้ ใช้การทดสอบด้วยค่าความแม่น (accuracy) ค่า sensitivity ค่า specificity ค่า F-measure และพื้นที่ใต้กราฟ ROC ในการจำแนกภาพแมมโมแกรมที่มีก้อนเนื้อร้ายและภาพที่มีก้อนเนื้อไม่อันตราย โดยผลการจำแนกด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนจะถูกเปรียบเทียบกับผลที่ได้จากอัลกอริทึมโครงข่ายประสาทเทียมและนาอีฟเบย์

4.1 ข้อมูลที่ใช้ในการทดสอบ

การทดสอบการจำแนกภาพแมมโมแกรมด้วยเทคนิคการประมวลผลภาพร่วมกับซัพพอร์ตเวกเตอร์แมชชีน ใช้ข้อมูลมาตรฐานภาพแมมโมแกรม (Digital Database for Screening Mammography: DDSM) จากเว็บไซต์ของมหาวิทยาลัยเซาท์ฟลอริดา (University of South Florida) ซึ่งเป็นภาพระดับสีเทา (grey scale image) มีข้อมูลจากคนไข้ทั้งหมด 2,500 ราย โดยมีทั้งข้อมูลของคนไข้ที่มีก้อนเนื้ออันตรายและก้อนเนื้อไม่อันตราย ซึ่งประกอบด้วยข้อมูลภาพแมมโมแกรมใน 2 มุมมอง คือภาพแมมโมแกรมในมุมมองแบบ MLO และ ภาพแมมโมแกรมในมุมมองแบบ CC โดยในการทดสอบนั้นจะคัดเลือกเฉพาะภาพแมมโมแกรมในมุมมองแบบ CC มาทั้งหมด 190 ภาพ เนื่องจากภาพในมุมมองแบบ CC ไม่มีส่วนพื้นที่สีขาวในมุมบนด้านซ้ายและขวาทำให้สะดวกต่อการประมวลผลภาพ ภาพในมุมมองแบบ CC ที่เลือกมานั้นมีคลาสเป้าหมายสองกลุ่ม คือ Malignant (ก้อนเนื้อร้าย) และ Benign (ก้อนเนื้อไม่อันตราย) โดยแต่ละภาพจะมีความกว้างอยู่ในช่วง 2,000-3,600 พิกเซล และความสูงอยู่ในช่วง 4,000-6,000 พิกเซล ดาวน์โหลดภาพแมมโมแกรมได้ที่เว็บไซต์ http://marathon.csee.usf.edu/Mammography/Database.html โดยตัวอย่างข้อมูลแสดงดังตารางที่ 4

เนื่องจากภาพแมมโมแกรมที่ใช้มีค่อนข้างขนาดใหญ่ดังนั้นจึงต้องผ่านกระบวนการการประมวลผลภาพก่อนด้วยโปรแกรม MATLAB R2013b โดยใช้วิธีการมีเดียนฟิลเตอร์ แกมมาคอเรคชัน และการขยายพื้นที่ของภาพเพื่อคัดเลือกเฉพาะบริเวณภาพก้อนเนื้อที่สนใจ หลังจากนั้นจึงทำการดึงลักษณะสำคัญของภาพ 3 ประเภท ออกมาเป็นข้อมูลตัวเลขซึ่งประกอบด้วยข้อมูลจำนวน 21 คอลัมน์ โดยคอลัมน์ที่ 1 –15 เป็นลักษณะสำคัญของลวดลาย คอลัมน์ที่ 16 – 19 เป็นลักษณะสำคัญของกราฟฮิสโตแกรม และ คอลัมน์ที่ 20 เป็นลักษณะสำคัญของรูปร่าง และคอลัมน์ที่ 21 เป็นหมายเลขคลาส รายละเอียดคอลัมน์สรุปได้ดังตารางที่ 4.2 และแสดงตัวอย่างข้อมูลลักษณะสำคัญของภาพแมมโมแกรมจำนวน 16 ตัวอย่างได้ดังตารางที่ 4.3

ตารางที่ 4.1 ตัวอย่างข้อมูลภาพแมมโมแกรมจาก DDSM ในมุมมองแบบ CC

Class: Malignant Size: 3520 × 5970 	Class: Malignant Size: 2360 × 4730 	Class: Malignant Size: 2210 × 4430 
Class: Benign Size: 2920 × 5370 	Class: Benign Size: 3080 × 4600 	Class: Benign Size: 2700 × 4800 

ตารางที่ 4.2 ชื่อคอลัมน์และความหมายของลักษณะสำคัญของภาพแมมโมแกรม

ลำดับที่	ชื่อคอลัมน์	คำอธิบาย
1	Cont0	ความชัดเจนของลวดลายในทิศทาง 0 องศา
2	Cont45	ความชัดเจนของลวดลายในทิศทาง 45 องศา
3	Cont90	ความชัดเจนของลวดลายในทิศทาง 90 องศา
4	Cont135	ความชัดเจนของลวดลายในทิศทาง 135 องศา
5	Homo0	ความเป็นเนื้อเดียวกันของลวดลายในทิศทาง 0 องศา
6	Homo45	ความเป็นเนื้อเดียวกันของลวดลายในทิศทาง 45 องศา
7	Homo90	ความเป็นเนื้อเดียวกันของลวดลายในทิศทาง 90 องศา
8	Homo135	ความเป็นเนื้อเดียวกันของลวดลายในทิศทาง 135 องศา
9	Corr0	การเกิดขึ้นร่วมกันของลวดลายในทิศทาง 0 องศา
10	Corr45	การเกิดขึ้นร่วมกันของลวดลายในทิศทาง 45 องศา
11	Corr90	การเกิดขึ้นร่วมกันของลวดลายในทิศทาง 90 องศา
12	Corr135	การเกิดขึ้นร่วมกันของลวดลายในทิศทาง 135 องศา
13	Avg_Cont	ค่าเฉลี่ยความชัดเจนของลวดลาย
14	Avg_Corr	ค่าเฉลี่ยความเป็นเนื้อเดียวกันของลวดลาย
15	Avg_Homo	ค่าเฉลี่ยการเกิดขึ้นร่วมกันของลวดลาย
16	Hist_Avg	ค่าเฉลี่ยของกราฟฮิสโตแกรม
17	Hist_Var	ค่าความแปรปรวนของกราฟฮิสโตแกรม
18	Hist_Skew	ค่าความเบ้ของกราฟฮิสโตแกรม
19	Hist_Kur	ค่าความโด่งของกราฟฮิสโตแกรม
20	Peak_No	จำนวนจุดยอดของกราฟที่มีการเปลี่ยนโค้ง
21	Class_No	หมายเลขคลาส 0 หมายถึง Benign และ 1 หมายถึง Malignant

ตารางที่ 4.3 ข้อมูลที่เป็นลักษณะสำคัญของภาพแมมโมแกรมจำนวน 16 ตัวอย่าง

Cont0	Cont45	Cont90	Cont135	Homo0	Homo45	Home90	Homo135	Corr0	Corr45	Corr90
0.18938	0.18938	0.16691	0.16691	0.16411	0.16411	0.17065	0.17065	0.90531	0.90531	0.91654
0.19237	0.19237	0.16828	0.16828	0.14315	0.14315	0.14983	0.14983	0.90382	0.90382	0.91586
0.19345	0.19345	0.17137	0.17137	0.1722	0.1722	0.17933	0.17933	0.90328	0.90328	0.91432
0.18832	0.18832	0.1646	0.1646	0.18784	0.18784	0.19516	0.19516	0.90584	0.90584	0.9177
0.28341	0.28341	0.23972	0.23972	0.14308	0.14308	0.15062	0.15062	0.8583	0.8583	0.88014
0.34413	0.34413	0.29337	0.29337	0.12725	0.12725	0.13512	0.13512	0.82793	0.82793	0.85332
0.22431	0.22431	0.19614	0.19614	0.12336	0.12336	0.13028	0.13028	0.88785	0.88785	0.90194
0.22187	0.22187	0.19291	0.19291	0.16631	0.16631	0.17544	0.17544	0.88906	0.88906	0.90355
0.22003	0.22003	0.18898	0.18898	0.16183	0.16183	0.1706	0.1706	0.88999	0.88999	0.90552
0.19232	0.19232	0.16818	0.16818	0.13818	0.13818	0.14508	0.14508	0.90384	0.90384	0.91591
0.16097	0.16097	0.14563	0.14563	0.1625	0.1625	0.16746	0.16746	0.91951	0.91951	0.92719
0.14913	0.14913	0.13389	0.13389	0.18883	0.18883	0.1927	0.1927	0.92547	0.92547	0.93308
0.18339	0.18339	0.1586	0.1586	0.1211	0.1211	0.12644	0.12644	0.9083	0.9083	0.9207
0.22083	0.22083	0.19004	0.19004	0.10329	0.10329	0.10928	0.10928	0.88959	0.88959	0.90549
0.20889	0.20889	0.17947	0.17947	0.11775	0.11775	0.12475	0.12475	0.89556	0.89556	0.91027
0.19323	0.19323	0.1674	0.1674	0.16154	0.16154	0.16958	0.16958	0.90338	0.90338	0.9163

Corr135	Avg_Cont	Avg_Homo	Avg_Corr	Hist_Avg	Hist_Var	Hist_Skew	Hist_Kur	Peak_No	Class_No
0.91654	0.17814	0.91093	0.16738	105.98	1026.5	-0.54017	1.7046	41	0
0.91586	0.18032	0.90984	0.14649	119.4	1102	-0.62523	1.8918	42	0
0.91432	0.18241	0.9088	0.17577	98.062	1247.7	-0.78261	2.349	41	0
0.9177	0.17646	0.91177	0.1915	104.4	1157.1	-0.71389	2.0602	44	0
0.88014	0.26156	0.86922	0.14685	107.29	1287.2	-0.62713	1.7038	41	0
0.85332	0.31875	0.84063	0.13119	120.69	1041.3	-0.96855	2.4706	37	0
0.90194	0.21023	0.89489	0.12682	93.071	1109.2	-0.27091	1.8367	41	0
0.90355	0.20739	0.89631	0.17088	84.89	1434	-0.19338	2.2212	41	0
0.90552	0.2045	0.89775	0.16622	125.3	1159.5	-0.82216	2.2815	37	0
0.91591	0.18025	0.90988	0.14163	127.31	1722.7	-0.72825	2.1301	83	1
0.92719	0.1533	0.92335	0.16498	132.22	1160.1	-0.72606	1.9087	84	1
0.93308	0.14151	0.92928	0.19077	93.419	1121.5	-0.15094	1.3695	86	1
0.9207	0.171	0.9145	0.12377	105.66	1294.3	-0.18212	1.5445	89	1
0.90549	0.20544	0.89754	0.10629	144.44	1150.8	-0.56714	1.8662	76	1
0.91027	0.19418	0.90292	0.12125	120.28	1112.8	-0.50497	2.0486	67	1
0.9163	0.18032	0.90984	0.16556	151.48	10528.1	-1.1487	3.2731	66	1

4.2 การเพิ่มชุดข้อมูลจากลักษณะสำคัญของรูปร่าง (Additional Data from Shape Feature: ADSF)

จากตารางที่ 4.3 จะเห็นว่าลักษณะสำคัญของรูปร่างจะมีเพียงแค่คอลัมน์เดียวคือคอลัมน์ชื่อ Peak_No โดยค่าตัวเลขเหล่านี้ได้มาจากวิธีการหาลักษณะสำคัญของรูปร่าง ซึ่งเป็นการวัดระยะห่างจากจุดเซนทรอยด์ไปยังเส้นขอบของรูปภาพ หลังจากนั้นจึงทำการนับจุดเปลี่ยนโค้งของเส้นขอบแสดงรูปร่างออกมาเป็นตัวเลขดังแสดงในตารางที่ 4.3 เนื่องจากลักษณะสำคัญของรูปร่างในคอลัมน์ Peak_No เมื่อนำไปเข้ากระบวนการจำแนกร่วมกับลักษณะสำคัญอื่น ๆ แล้วประสิทธิภาพการจำแนกยังไม่ดีเท่าที่ควร ดังนั้นจึงทำการเพิ่มเติมลักษณะสำคัญของรูปร่างโดยหาค่า threshold ที่

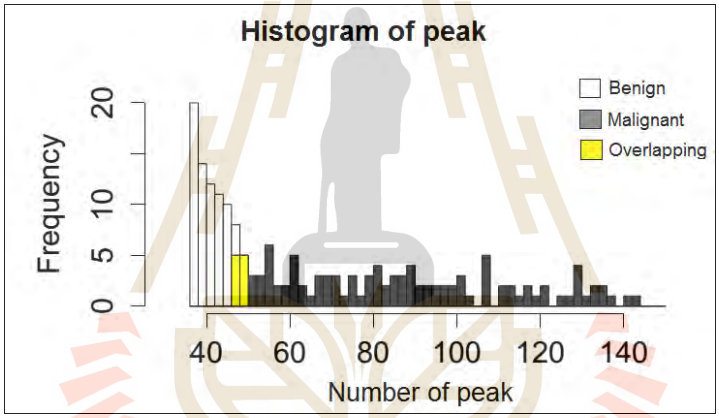
เหมาะสมของข้อมูลในคอลัมน์ Peak_No ระหว่างก้อนเนื้อไม่อันตรายกับก้อนเนื้ออันตราย วิธีการหาค่า Threshold ที่เหมาะสมทำได้ดังนี้

1) นำข้อมูลจากคอลัมน์ Peak_No จากคลาส Benign มาพล็อตกราฟฮิสโตแกรม

2) นำข้อมูลจากคอลัมน์ Peak_No จากคลาส Malignant มาพล็อตกราฟฮิสโตแกรม

3) หาค่า threshold ที่เหมาะสมโดยดูจากบริเวณที่กราฟฮิสโตแกรมทั้งสองฝั่งมีการซ้อนทับกัน (Overlapping) ดังแสดงในรูปที่ 4.1

4) เมื่อได้ค่า threshold แล้วนำค่านี้ไปเป็นค่าในการเพิ่มชุดข้อมูลลักษณะสำคัญของรูปร่าง โดยพิจารณาจากคอลัมน์ Peak_no เทียบกับค่า threshold หากค่า Peak_no มีค่าน้อยกว่าค่า threshold ให้ทำการเพิ่มตัวเลข 1 แทนที่คอลัมน์ Peak_no ต่อท้ายไปอีกเป็นจำนวน 20 คอลัมน์ และหากค่า Peak_no มีค่ามากกว่าหรือเท่ากับค่า threshold ให้ทำการเพิ่มตัวเลข 100 แทนที่คอลัมน์ Peak_no ต่อท้ายไปอีกเป็นจำนวน 20 คอลัมน์ดังแสดงในรูปที่ 4.2



รูปที่ 4.1 กราฟฮิสโตแกรมแสดงความถี่ของจุดพีคระหว่างก้อนเนื้ออันตรายและก้อนเนื้อไม่อันตราย

Threshold = 50

Peak_no	1	2	3	4	.....	17	18	19	20
48	1	1	1	1	.....	1	1	1	1
39	1	1	1	1	.....	1	1	1	1
42	1	1	1	1	.....	1	1	1	1
50	100	100	100	100	.....	100	100	100	100
63	100	100	100	100	.....	100	100	100	100
114	100	100	100	100	.....	100	100	100	100

รูปที่ 4.2 แสดงการเพิ่มชุดข้อมูลจากการพิจารณาฮิสโตแกรมลักษณะสำคัญของรูปร่าง

4.3 ผลการทดสอบประสิทธิภาพการจำแนกภาพแมมโมแกรม

การทดสอบประสิทธิภาพการจำแนกนี้จะเปรียบเทียบด้วยค่าความแม่น (accuracy) ค่า sensitivity ค่า specificity ค่า F-measure และพื้นที่ใต้กราฟ ROC ในการจำแนกข้อมูลภาพแมมโมแกรมของอัลกอริทึม 3 แบบ ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (ใช้เคอร์เนลฟังก์ชันชนิดเรเดียลเบสิส) โครงข่ายประสาทเทียม และ นาอีฟเบย์ ข้อมูลที่ใช้ทดสอบแบ่งเป็น ข้อมูลที่ใช้ในการฝึกสอน (train data) จำนวน 133 ข้อมูล (70% จาก 190 ภาพ) โดยแบ่งเป็นภาพก้อนเนื้ออันตรายจำนวน 77 ภาพ (คลาส Malignant) และก้อนเนื้อไม่อันตรายจำนวน 56 ภาพ (คลาส Benign) และข้อมูลที่ใช้ทดสอบ (test data) จำนวน 57ข้อมูล (30% จาก 190 ภาพ) โดยแบ่งเป็นภาพก้อนเนื้ออันตรายจำนวน 33 ภาพ และก้อนเนื้อไม่อันตรายจำนวน 24 ภาพ

ในการทดสอบประสิทธิภาพจะใช้ข้อมูลลักษณะสำคัญทั้ง 3 แบบคือลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม (ADSF, Texture, Histogram หรือ ADSF-TH) เป็นข้อมูลนำเข้า ข้อมูลนี้ประกอบด้วย 39 คอลัมน์ (ลักษณะสำคัญของลวดลายจำนวน 15 คอลัมน์ ลักษณะสำคัญของกราฟฮิสโตแกรมจำนวน 4 คอลัมน์ และลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล 20 คอลัมน์) โดยสุ่มข้อมูลทดสอบจำนวน 57 ข้อมูลจากทั้ง 2 คลาส ผลการทดสอบประสิทธิภาพแสดงในลักษณะของตาราง confusion matrix ได้ดังตารางที่ 4.4-4.6 ซึ่งแสดงผลการทดสอบจำแนกตามอัลกอริทึมการจำแนกที่แตกต่างกัน 3 อัลกอริทึม

ตารางที่ 4.4 ผลการจำแนกภาพแมมโมแกรมด้วยซัพพอร์ตเวกเตอร์แมชชีน

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">ค่าจริง (Actual)</td><td rowspan="2"></td></tr>
<tr><td>Positive</td><td>Negative</td></tr>
<tr><td rowspan="2">ค่าทำนาย (Predict)</td><td>Positive</td><td>True Positive (TP) = 31</td><td>False Positive (FP) = 2</td><td rowspan="2">Accuracy
=((31+22) / (31+22+2+2))*100
= 92.98%</td></tr>
<tr><td>Negative</td><td>False Negative (FN) = 2</td><td>True Negative (TN) = 22</td></tr>
<tr><td>Precision
= 31 / (31+2)
= 0.9394</td><td></td><td>Sensitivity
=(31/(31+2))*100
= 93.94%</td><td>Specificity
=(22/(2+22))*100
= 91.67%</td><td>F-measure
= (2×0.9394×0.9394) / (0.9394+0.9394)
= 0.9394</td></tr>
</table>

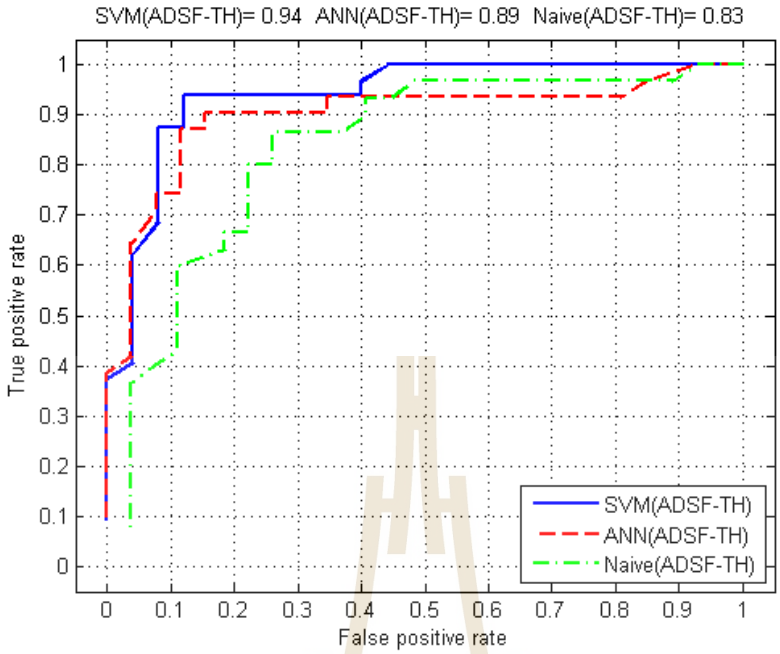
ตารางที่ 4.5 ผลการจำแนกภาพแมมโมแกรมด้วยโครงข่ายประสาทเทียม

		ค่าจริง (Actual)		
		Positive	Negative	
ค่าทำนาย (Predict)	Positive	True Positive (TP) = 30	False Positive (FP) = 3	**Accuracy =((30+21) / (30+21+3+3))*100 = 89.47%**
	Negative	False Negative (FN) = 3	True Negative (TN) = 21	
Precision = 30 / (30+3) = 0.9091		**Sensitivity =(30/(30+3))*100 = 90.91%**	**Specificity =(21/(3+21))*100 = 87.50%**	**F-measure = (2×0.9091×0.9091) / (0.9091+0.9091) = 0.9091**

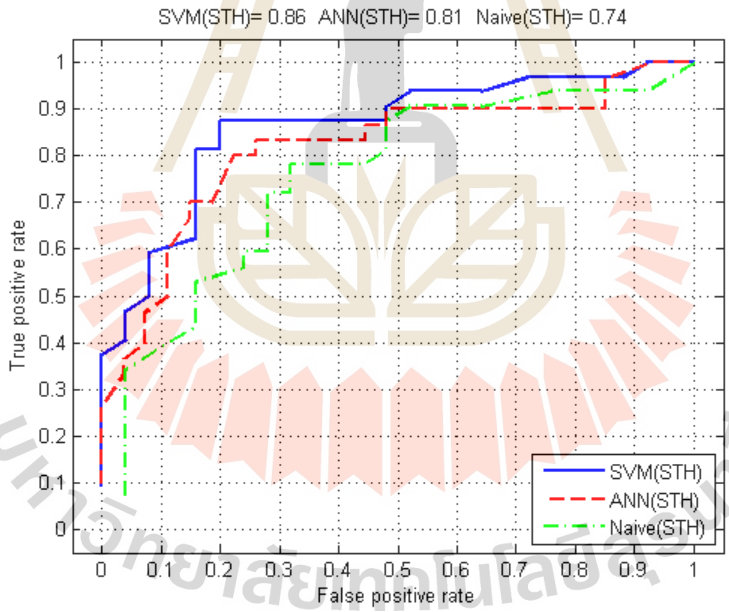
ตารางที่ 4.6 ผลการจำแนกภาพแมมโมแกรมด้วยนาอีฟเบย์

		ค่าจริง (Actual)		
		Positive	Negative	
ค่าทำนาย (Predict)	Positive	True Positive (TP) = 27	False Positive (FP) = 6	**Accuracy =((27+20) / (27+20+6+4))*100 = 82.45%**
	Negative	False Negative (FN) = 4	True Negative (TN) = 20	
Precision = 27 / (27+6) = 0.8182		**Sensitivity =(27/(27+4))*100 = 87.10%**	**Specificity =(20/(6+20))*100 = 79.92%**	**F-measure = (2×0.8182×0.8710) / (0.8182+0.8710) = 0.8438**

ผลการทดลองที่แสดงดังตารางที่ 4.4-4.6 เป็นการทดสอบกับชุดข้อมูลที่เรียกชื่อว่า ADSF-TH ซึ่งเป็นข้อมูลที่รวมลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล (Additional Data from Shape Feature: ADSF) ลักษณะสำคัญของลวดลาย (texture) และลักษณะสำคัญของกราฟฮิสโตแกรม (histogram) เพื่อเป็นการยืนยันประสิทธิภาพของแนวคิดการปรับปรุงข้อมูลด้วยเทคนิคการประมวลผลภาพ งานวิจัยนี้ได้ทดลองเปรียบเทียบกับชุดข้อมูลนำเข้าเดียวกันแต่ใช้ฟีเจอร์ข้อมูลที่ประกอบด้วย ลักษณะสำคัญของรูปร่างแบบไม่เพิ่มชุดข้อมูล (shape) ลักษณะสำคัญของลวดลาย (texture) และลักษณะสำคัญของกราฟฮิสโตแกรม (histogram) และเรียกชื่อชุดข้อมูลนี้ว่า STH ผลการทดสอบประสิทธิภาพการจำแนกภาพแมมโมแกรมด้วยค่าพื้นที่ใต้กราฟ ROC ระหว่างชุดข้อมูล ADSF-TH และชุดข้อมูล STH แสดงได้ดังรูปที่ 4.3 และ 4.4 ตามลำดับ



รูปที่ 4.3 พื้นที่ใต้กราฟ ROC โดยใช้ลักษณะสำคัญแบบ ADSF-TH



รูปที่ 4.4 พื้นที่ใต้กราฟ ROC โดยใช้ลักษณะสำคัญแบบ STH

ในกรณีใช้ฟีเจอร์แบบ ADSF-TH (รูปที่ 4.3) อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนให้ค่าพื้นที่ใต้กราฟROC มากที่สุด คือ 0.94 รองลงมาเป็นอัลกอริทึมโครงข่ายประสาทเทียมให้ค่าพื้นที่ใต้กราฟ 0.89 และลำดับสุดท้ายคืออัลกอริทึมนาอีฟเบย์ ให้ค่าพื้นที่ใต้กราฟ 0.83 สำหรับกรณีใช้ข้อมูลลักษณะสำคัญแบบ STH (รูปที่ 4.4) อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าพื้นที่ใต้กราฟ ROC

มากที่สุดเช่นเดียวกันคือ 0.86 รองลงมาเป็นอัลกอริทึมโครงข่ายประสาทเทียมให้ค่าพื้นที่ใต้กราฟ 0.81 และลำดับสุดท้ายคืออัลกอริทึมนาอีฟเบย์ ให้ค่าพื้นที่ใต้กราฟ 0.74 ดังนั้นจึงสรุปได้ว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพในการจำแนกดีที่สุดคือได้พื้นที่ใต้กราฟ ROC มากที่สุด และให้ค่า False Positive Rate ต่ำสุดซึ่งสังเกตได้จากการที่กราฟ ROC ชิดมุมซ้ายบนมากที่สุด และกรณีใช้ข้อมูลลักษณะสำคัญแบบ ADSF-TH ให้ผลลัพธ์ที่ดีกว่าแบบ STH

ตารางที่ 4.7 แสดงการเปรียบเทียบค่าความแม่น ค่า Sensitivity ค่าSpecificity ค่า F-measure และ พื้นที่ใต้กราฟ ROC ระหว่าง 3 อัลกอริทึมจำแนกตามลักษณะฟีเจอร์แบบ ADSF-TH และแบบ STH จากค่าที่แสดงในตารางจะให้เห็นว่าการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ข้อมูลจากลักษณะสำคัญแบบ ADSF-TH ให้ค่าสูงที่สุดในทุกด้าน

ตารางที่ 4.7 เปรียบเทียบค่า Accuracy Sensitivity Specificity F-measure และ AUC

Algorithm	Features	Accuracy	Sensitivity	Specificity	F-measure	AUC
SVM	**ADSF-TH**	**92.98%**	**93.94%**	**91.67%**	**93.94%**	**0.94**
	STH	87.72%	90.63%	84.00%	89.23%	0.86
ANN	ADSF-TH	89.47%	90.91%	87.50%	90.91%	0.89
	STH	84.21%	87.50%	80.00%	86.15%	0.81
Naïve Bayes	ADSF-TH	82.45%	87.10%	79.92%	84.38%	0.83
	STH	78.95%	83.87%	73.08%	81.25%	0.74

4.4 อภิปรายผล

จากผลการทดสอบประสิทธิภาพการจำแนกภาพแมมโมแกรมโดยใช้การประมวลผลภาพร่วมกับซัพพอร์ตเวกเตอร์แมชชีน ได้ทำการทดสอบกับข้อมูลภาพแมมโมแกรมจำนวน 190 ภาพโดยประกอบด้วยข้อมูลภาพ 2 คลาส คือ ก้อนเนื้อร้ายแรง (malignant) และก้อนเนื้อไม่อันตราย (benign) กระบวนการประมวลผลภาพได้ถูกนำมาใช้ เพื่อทำการดึงเฉพาะลักษณะสำคัญของภาพก่อนนำข้อมูลไปเข้ากระบวนการจำแนกและประเมินประสิทธิภาพ สามารถสรุปผลการทดสอบเปรียบเทียบได้ดังนี้

1) การประมวลภาพ โดยใช้วิธีการกำจัดสัญญาณรบกวนในภาพด้วยตัวกรองมัธยฐาน การแก้ไขค่าแกมมา และการขยายส่วนของพื้นที่ มีผลทำให้ข้อมูลภาพมีขนาดเล็กลง และทำให้สามารถดึงลักษณะสำคัญของภาพออกมาได้ง่ายขึ้น ทั้งนี้เพื่อลดมิติข้อมูลก่อนเข้ากระบวนการจำแนก

เนื่องจากข้อมูลความเข้มสีของภาพเป็นข้อมูลที่มีขนาดมิติใหญ่มาก ซึ่งส่งผลต่อประสิทธิภาพในการจำแนกโดยตรง

2) ลักษณะสำคัญ 3 ประเภทที่นำมาใช้ในการจำแนก ได้แก่ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม เป็นข้อมูลที่สำคัญที่ทำให้การจำแนกมีประสิทธิภาพ โดยเฉพาะอย่างยิ่ง ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ซึ่งเป็นตัวบ่งบอกความหยักของก้อนเนื้อ และเป็นลักษณะสำคัญที่ทำให้การจำแนกมีประสิทธิภาพเพิ่มมากขึ้นอย่างมีนัยสำคัญ

3) การจำแนกภาพแมมโมแกรมด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนโดยใช้เคอร์เนลฟังก์ชันแบบเรเดียลเบสิส โดยใช้ลักษณะสำคัญทั้ง 3 แบบ คือ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของฮิสโตแกรม ให้ประสิทธิภาพในการจำแนกดีที่สุด เมื่อเปรียบเทียบกับอีก 2 อัลกอริทึม คือ โครงข่ายประสาทเทียม และ นาอีฟเบย์ โดยการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีนได้ค่าความแม่น=92.98% ค่า Sensitivity=93.94% ค่า Specificity=91.67% ค่า F-measure=93.94% และ พื้นที่ใต้กราฟ ROC=0.94

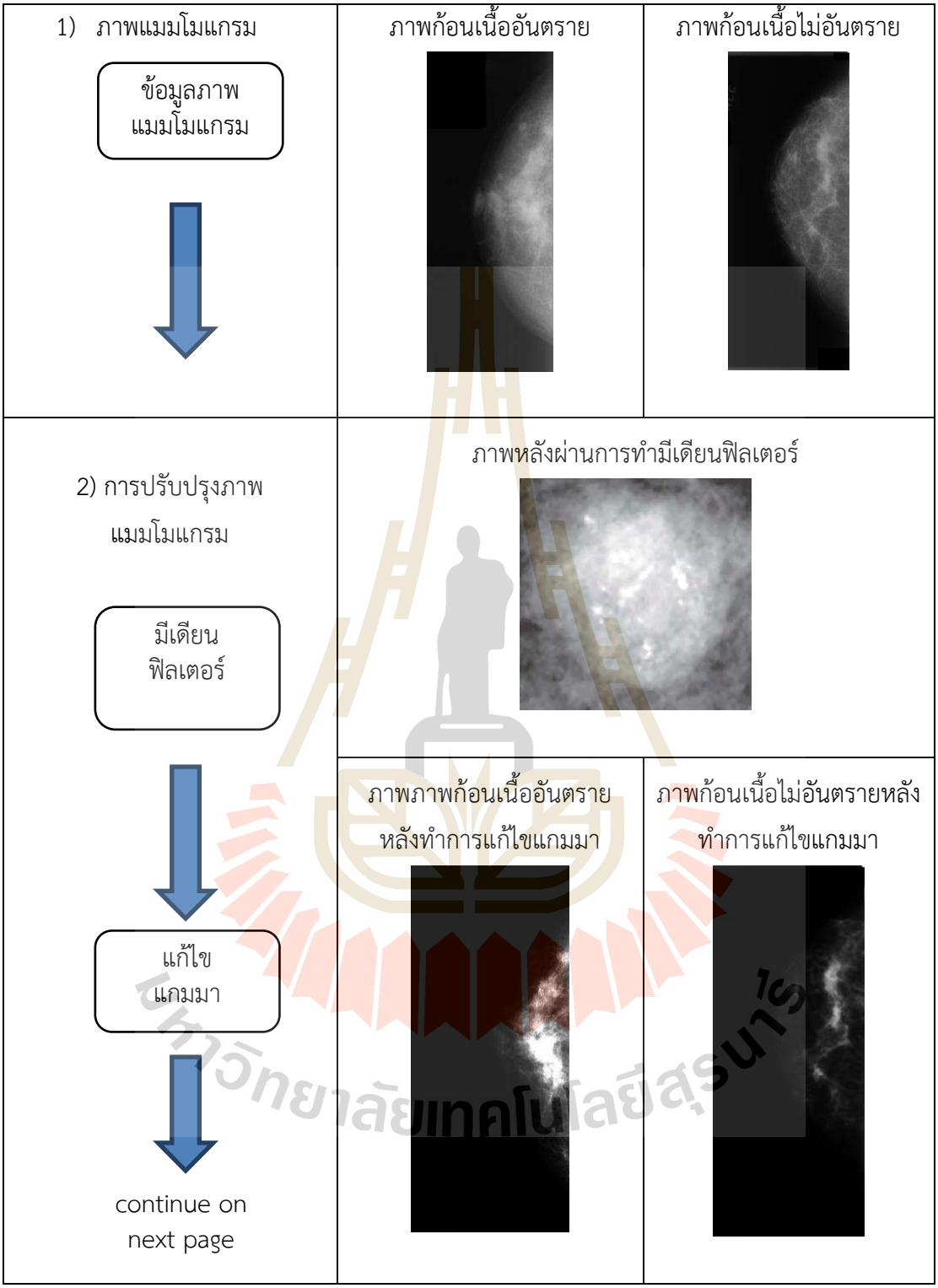
บทที่ 5
บทสรุป

5.1 สรุปผลการวิจัย

การตรวจวินิจฉัยมะเร็งเต้านมจากภาพแมมโมแกรม มีวัตถุประสงค์เพื่อจำแนกก้อนเนื้อภายในภาพแมมโมแกรมว่าเป็นก้อนเนื้อไม่อันตรายหรือก้อนเนื้ออันตราย ในปัจจุบันมีนักวิจัยจำนวนมากประยุกต์เทคนิคต่าง ๆ โดยเฉพาะเทคนิคเกี่ยวกับการประมวลผลภาพ เพื่อเพิ่มประสิทธิภาพของการจำแนกภาพแมมโมแกรมให้มีความแม่นยำมากขึ้น การปรับปรุงคุณภาพของภาพก่อนการนำไปจำแนกด้วยเทคนิคการเรียนรู้ของเครื่อง เป็นขั้นตอนที่สำคัญเนื่องจากภาพแมมโมแกรมอาจมีความไม่ชัดเจนหรือมีสัญญาณรบกวนในภาพทำให้การจำแนกได้ผลที่ไม่ดีนัก

ดังนั้นงานวิจัยนี้จึงได้เสนอวิธีการปรับปรุงคุณภาพของภาพแมมโมแกรมด้วยแนวคิดหลัก คือ การกำจัดหรือลดสัญญาณรบกวนภายในภาพออกไป แล้วจึงทำการปรับปรุงภาพโดยทำให้ความเข้มสีบริเวณก้อนเนื้อในภาพชัดเจนขึ้น จากนั้นจึงใช้เทคนิคการประมวลผลภาพด้วยวิธีการหาขอบเขตที่น่าสนใจ โดยใช้ขั้นตอนวิธีในการตัดเฉพาะบริเวณก้อนเนื้อในภาพแมมโมแกรมเพื่อนำมาประมวลผล หลังจากได้บริเวณขอบเขตที่สนใจแล้ว ขั้นตอนการปรับปรุงภาพที่สำคัญในขั้นสุดท้ายคือการหาลักษณะสำคัญภายในบริเวณขอบเขตที่สนใจ โดยงานวิจัยนี้จะพิจารณาลักษณะสำคัญ 3 ลักษณะ คือ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูลเพื่อเป็นการเพิ่มน้ำหนักให้กับบริเวณสำคัญในภาพแมมโมแกรม ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของฮิสโตแกรม ในขั้นตอนของการจำแนกภาพระหว่างก้อนเนื่ออันตราย (malignant) และก้อนเนื้อไม่อันตราย (benign) ลักษณะสำคัญทั้ง 3 แบบจะถูกนำไปใช้เป็นข้อมูลประกอบการจำแนก

งานวิจัยนี้ใช้การจำแนกข้อมูลแบบมีผู้สอน โดยเลือกใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เนื่องจากวิธีซัพพอร์ตเวกเตอร์แมชชีนนิยมใช้ในการจำแนกข้อมูลภาพ และจากการสำรวจวรรณกรรมพบว่าเทคนิคซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความแม่นยำสูงกว่าเทคนิคอื่น ๆ ความแม่นยำของซัพพอร์ตเวกเตอร์แมชชีน เกิดจากการเลือกใช้เคอร์เนลฟังก์ชันได้หลายแบบเพื่อแปลงข้อมูลไปสู่ระนาบที่เหมาะสมกับการจำแนกข้อมูลแต่ละประเภท งานวิจัยนี้ใช้เคอร์เนลฟังก์ชันแบบเรเดียลเบสิสซึ่งจากผลการทดลองเบื้องต้นพบว่าให้ความแม่นยำในการจำแนกภาพแมมโมแกรมได้ดีที่สุด และได้ทำการเปรียบเทียบประสิทธิภาพการจำแนกระหว่างเทคนิคซัพพอร์ตเวกเตอร์แมชชีนกับเทคนิคการจำแนกแบบอื่น ๆ เช่น โครงข่ายประสาทเทียมและนาอีฟเบย์ โดยพบว่าการนำลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูลเข้าไปใช้ในการจำแนก ทำให้ประสิทธิภาพในการจำแนกดีขึ้นในทั้ง 3 อัลกอริทึม ทั้งนี้อัลกอริทึมที่ให้ผลลัพธ์ดีที่สุดคือซัพพอร์ตเวกเตอร์แมชชีน การทดสอบประสิทธิภาพของการจำแนกภาพแมมโมแกรมด้วยชุดข้อมูลมาตรฐาน DDSM สามารถสรุปขั้นตอนทั้งหมดได้ดังรูปที่ 5.1



รูปที่ 5.1 สรุปขั้นตอนการดำเนินงานวิจัย

3) การแบ่งขอบเขตภาพ

การขยายพื้นที่ขอบเขตภาพ

ภาพหลังทำการขยายพื้นที่ขอบเขตภาพ

4) การดึงลักษณะสำคัญในภาพ

ลักษณะสำคัญของลวดลาย

Direction	Homogeneity	Contrast	Correlation
0°	0.0125	3.0382	0.8074
45°	0.0082	4.0121	0.6369
90°	0.0075	4.0153	0.5988
135°	0.0069	4.7084	0.4613
Average	0.0087	3.9435	0.6261

ลักษณะสำคัญของฮิสโตแกรม

Average	Variance	Skewness	Kurtosis
0.002196	0.020183	-0.21837	-0.00766

ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล

Threshold = 50

Peak_no	1	2	3	4	.....	17	18	19	20
48	1	1	1	1	.....	1	1	1	1
39	1	1	1	1	.....	1	1	1	1
42	1	1	1	1	.....	1	1	1	1
50	100	100	100	100	.....	100	100	100	100
63	100	100	100	100	.....	100	100	100	100
114	100	100	100	100	.....	100	100	100	100

5) จำแนกโดย SVM(RBF)

ข้อมูลที่ใช้ในการฝึกสอน (Train) จำนวน 133 ภาพ
ข้อมูลที่ใช้ในการทดสอบ (Test) จำนวน 157 ภาพ

6) ประเมินประสิทธิภาพ

Accuracy, Sensitivity, Specificity, F-measure และ ROC

รูปที่ 5.1 สรุปขั้นตอนการดำเนินงานวิจัย (ต่อ)

5.2 ข้อจำกัดและข้อเสนอแนะ

ภาพแมมโมแกรมเป็นภาพที่มีขนาดใหญ่มาก ทำให้การประมวลผลในส่วนของการปรับปรุงคุณภาพของภาพใช้เวลาในการประมวลผลนาน เนื่องจากในงานวิจัยนี้ได้ใช้การประมวลผลภาพหลากหลายวิธี เพื่อทำการปรับปรุงภาพให้มีคุณภาพที่ดีขึ้นก่อนนำไปเข้ากระบวนการจำแนก ซึ่งข้อเสนอแนะสำหรับการพัฒนางานวิจัยในอนาคต อาจพิจารณานำเทคนิคการปรับปรุงภาพแบบอื่น ๆ ที่มีขั้นตอนการประมวลผลรวดเร็วกว่านี้มาใช้ เพื่อทำให้กระบวนการปรับปรุงภาพมีความรวดเร็วมากยิ่งขึ้น

งานวิจัยนี้ในขั้นตอนการคัดเลือกลักษณะสำคัญนั้น เป็นการทดสอบการผสมผสานลักษณะสำคัญ 3 แบบ และนำมาเปรียบเทียบประสิทธิภาพเพื่อสรุปว่าลักษณะสำคัญใด มีผลทำให้ประสิทธิภาพในการจำแนกดีขึ้น ซึ่งการนำลักษณะสำคัญมาผสมผสานกันนั้นยังไม่เป็นกระบวนการอัตโนมัติ ทำให้หากมีลักษณะสำคัญหลาย ๆ แบบจะทำให้สิ้นเปลืองเวลาในการทดสอบและเปรียบเทียบ ดังนั้นการพัฒนางานวิจัยนี้ให้ใช้เวลาประมวลผลสั้นลง ทางผู้วิจัยจึงมีข้อเสนอแนะว่าควรนำเทคนิคในการคัดเลือกลักษณะสำคัญแบบอัตโนมัติมาใช้ เพื่อเพิ่มประสิทธิภาพในการคัดเลือกลักษณะสำคัญของภาพแมมโมแกรม

บรรณานุกรม

R. Campanini, D. Donggiovanni, E. Iampieri, N. Lanconelli, M. Masotti, G. Palermo, A. Riccardi, and M. Roffilli (2004), A novel featureless approach to mass detection in digital mammograms based on support vector machines, Physics in Medicine and Biology, Vol.49, No.6, pp.961-975.

H.D. Cheng, X.J. Shi, R Min L.M. Hu, X.P. Cai, and H.N. Du (2006), Approaches for automated detection and classification of masses in mammograms, Pattern Recognition, Vol.39, No.4, pp.646-668.

I. Christoyianni, E. Dermatal, and G. Kokkinakis (2000), Fast detection of masses in computer-aided mammography, IEEE Signal Processing Magazine, Vol.17, No.1, pp.54-64.

M. Elter and A. Horsch (2009), CADx of mammographic masses and clustered microcalcifications: a review, Medical Physics, Vol.36, No.6, pp.2052-2068.

Z. Huo, M.L. Giger, C.J. Vyborny, U. Bick, P. Lu, D.E. Wolverton, and R.A. Schmidt (1995), Analysis of speculation in the computerized classification of mammographic masses, Medical Physics, Vol.22, No.10, pp.1569-1579.

V. Jackson, R. Hendrick, S. Feig, and D. Kopans (1993), Imaging of the radiographically dense breast, Radiology, Vol.188, pp.297-301.

H. Jiang, W. Tiu, S. Yamamoto, and S. lisaku (1998), A method for automatic detection of spicules in mammograms, Journal of Computation Aided Diagnosis of Medical Information, Vol.2, No.4, pp.1-8.

W.P. Kegelmeyer, J.M. Pruneda, P.D. Bourland, A. Hillis, M.W. Riggs, and M.L. Nipper (1994), Computer-aided mammographic screening for speculated lesions, Radiology, Vol.191, No.2, pp.331-337.

R.M. Rangayyan, F.J. Ayres, and J.E.L. Desautels (2007), A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs, Journal of the Franklin Institute, Vol.334, No.3-4, pp.312-348.

H.D. Cheng, X.J. Shi, R Min L.M. Hu, X.P. Cai, and H.N. Du (2006), Approaches for automated detection and classification of masses in mammograms, Pattern Recognition, Vol.39, No.4, pp.646-668.

R. Sivaramakrishna, K.A. Powell, M.L. Lieber, W.A. Chilcote, and R. Shekhar (2002), Texture analysis of lesions in breast ultrasound images, Computerize Medical Imaging and Graphics, Vol.26, No.5, pp.303-307.

R. Stathaki and A.G. Constantinides (1994), Neural networks and higher order spectra for breast cancer detection, IEEE Workshop on Neural Networks and Signal Processing, pp.473-481.

ภาคผนวก

ผลผลิตของงานวิจัย

ภาคผนวก ก

บทความวิจัยตีพิมพ์ในวารสารและเอกสารสืบเนื่องจากการประชุมวิชาการ

1. K. Chaiyakhan, N. Kerdprasop, K. Kerdprasop (2016). Mammography image classification and clustering using support vector machine and k-means. ICIC Express Letters, Part B: Applications, vol.7, no.5, May, pp. 961-967.

2. K. Suksut, R. Chanklan, N. Kaoungku, K. Chaiyakhan, N. Kerdprasop, K. Kerdprasop (2017). Parameter optimization for mammogram image classification with support vector machine. Proceedings of the 25th International MultiConference of Engineers and Computer Scientists (IMECS2017), Hong Kong, 15-17 March, pp. 337-341.

3. K. Chaiyakhan, N. Kerdprasop, K. Kerdprasop (2016). Feature selection techniques for breast cancer image classification with support vector machine. Proceedings of the 24th International MultiConference of Engineers and Computer Scientists (IMECS2016), Hong Kong, 16-18 March, pp.237-232.

4. K. Chaiyakhan, N. Kerdprasop, K. Kerdprasop (2015). Mammography images categorization with k-means clustering. Proceedings of the 9th South East Asia Technical University Consortium (SEATUC) Symposium, Suranaree University of Technology, Thailand, 27-30 July, pp.111-114.

ICIC Express Letters
Part B: Applications
Volume **7**, Number **5**, May **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **961–967**

MAMMOGRAPHY IMAGE CLASSIFICATION AND CLUSTERING USING SUPPORT VECTOR MACHINE AND K-MEANS

Kedkarn Chaiyakhan, Nittaya Kerdprasop and Kittisak Kerdprasop

School of Computer Engineering
Suranaree University of Technology
111 University Avenue, Nakhon Ratchasima 30000, Thailand
kedkarnc@hotmail.com; {nittaya; kerdpras}@sut.ac.th

Received October 2015; accepted January 2016

Abstract. *Mammography is an extraordinary type of low-powered x-ray process that provides detailed images of the internal structure of the breast. An early detection of breast cancer by means of mammography results in a successful treatment. Many researches show that the dense masses in the breast density are one of the strongest indicators of breast cancer developing. In this paper, we propose an approach to automatically appraise the density and contrast of breast images using gamma correction to increase the intensity of dense pixels with light intensity and vice versa to decrease the sparse intensity pixels showing dark intensity. In the segmentation process, we use region growing technique to get region of interest. We also extract three important features including texture, shape, and intensity histogram. In the classification process, we use SVM to classify tumor into two classes: malignant and benign. Moreover, we also compare the SVM classification result to the Naïve Bays and artificial neural network techniques. In clustering process, we use the k-means algorithm to cluster image into 2 categories: malignant and benign. The results of classification and clustering show that our proposed work can classify and cluster two types of mammography images after the appropriate application of gamma correction feature extraction process.*

Keywords: Image segmentation, Image classification, Image clustering, k-means, Support vector machine

1. **Introduction.** Breast cancer is a dangerous type of tumor originated from breast tissue. The most effective way to detect breast cancer is through the breast mammogram screening. However, the major limitation for mammography diagnosis is its sensitivity because interpreting mammography is a labor-intensive task for radiologists who cannot always offer stable results during interpreting. Many methodologies have thus been proposed to solve this uncertain interpretation problem by providing assistance to the advanced cancer detection and diagnosis tools.

The statistical approach has been proposed [1]. The authors provide connected density clusters taking the spatial information of the breast tissue into account. Quantitative and qualitative results show that their approach is able to correctly detect dense breasts apart from other tissue types. A methodology that is based on modeling a set of patched of either fatty or dense parenchyma using statistical analysis has been presented [2]. The two strategies, PCA and linear-discriminant analysis, are applied in the modeling process. In the work of [3], they use mixtures of Gaussian for modeling and segmenting the breast into four and five regions, respectively. However, these approaches do not take spatial information into account resulting in too many disconnected regions. Thus, the work of [4] has included a fuzzy affinity function in their proposed method, while [5] employs textural features to take the spatial distribution of the pixel and its neighborhood. Some researchers [6,7] use region growing, which is the region-based segmentation method. In the work of [8], they use region growing method based on the gradients and variances along

and inside of the boundary curve. Some researchers use edge and smoothness factors as criteria to determine initial seed points and then seeded region growing method is used to segment images based on seed regions [9].

In our proposed method, we use gamma correction to enhance the image contrast. In segmentation process, we use a well-known region growing method to find the ROI and then crop the image to consider only the tumor region. The unnecessary background has been removed in this process. After that we extract three types of feature and input digital data to the classification and clustering process. The performance of the proposed image classification approach has been evaluated by comparing the accuracy with some state of the art classification algorithms.

2. **Proposed Work.** In the proposed work, we have divided our process into five main parts: image preprocessing, segmentation, feature extraction, classification, and clustering. Figure 1 shows the framework of this research.

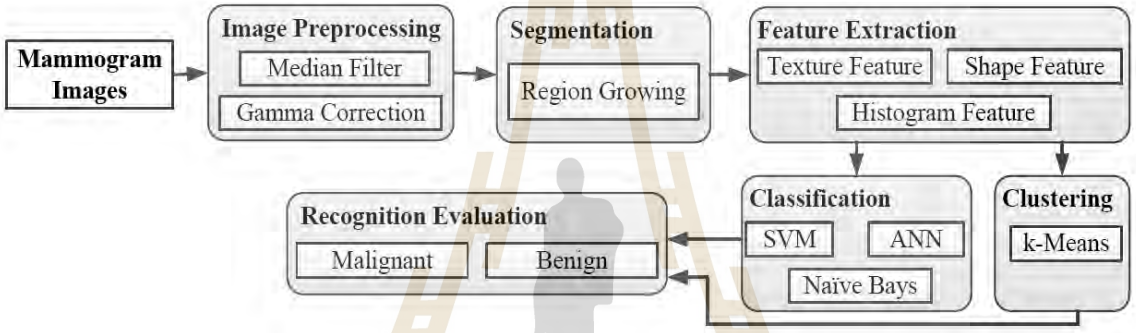


FIGURE 1. The framework of the proposed tumor recognition system

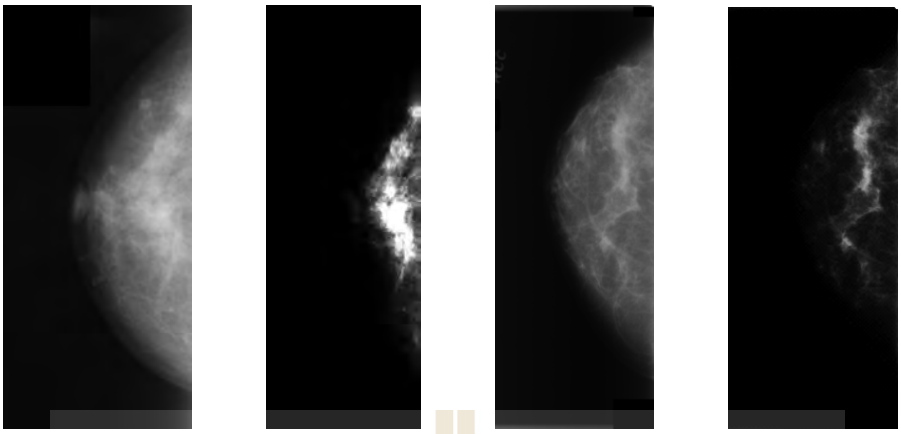
2.1. **Image preprocessing.** Mammogram images usually contain noises because of disturbances like Gaussian noise or some little darkness and brightness noise called salt and pepper noise. We use median filter to remove these noises. The output of this de-noising step is the clear images that are appropriate for further processing.

The next step of image preprocessing is image enhancement. We adjust the brightness and darkness of images using gamma correction algorithm. Figure 2 shows the original images of malignant and benign cases comparing to the improved results after applying the gamma correction technique. The gamma correction helps contrasting the tumor area from the fatty area.

2.2. **Segmentation.** This process separates the tumor areas from the background tissue in mammogram images. In this step, we apply the region growing segmentation method. Region growing is a region-based method starting with selecting seed points in the image, then propagating seeds until the specified stopping criteria are satisfied. Appropriate seed point selection is important. Therefore, in our proposed work, we select seed point using the centroid of object computed from area and position of object (centroid), as shown in Equation (1).

$$Centroid \quad \bar{x} = \frac{\sum_i \sum_j jW[i,j]}{Area} \qquad \bar{y} = \frac{\sum_i \sum_j iW[i,j]}{Area} \tag{1}$$

where W is the white pixel in the image, $Area$ is summation of white pixels, and i, j are the position of white pixel. After the region growing process, we will get the region of interest (ROI, white pixels) and then we crop only the ROI (Figure 3) removing background that may affect the classification and clustering process.



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

FIGURE 2. Breast tumor images: (a) original malignant case, (b) malignant image after gamma correction, (c) original benign case, (d) benign image after gamma correction

(a) (b) (c)

FIGURE 3. The result of region growing and the cropped image: (a) gamma corrected image, (b) the image after applying region growing technique, (c) cropped image

2.3. **Feature extraction.** In this work, we extract three types of features: texture, shape, and intensity histogram features.

1). Texture Features

Texture is one of the important features used in identifying objects in an image. Texture features are based on the gray-level co-occurrence matrix (GLCM). The GLCM function characterizes the texture of an image by calculating how often pairs of pixels with specific values and in a specified spatial relationship occur in an image. We create a GLCM, and then extract from the matrix statistical measures such as contrast, correlation, energy, and homogeneity.

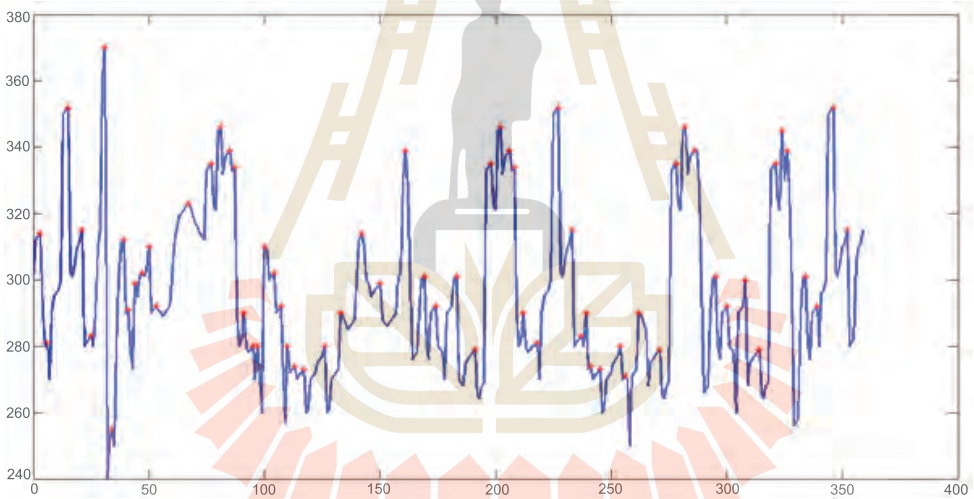
2). Shape Features

We extract shape feature using the percentage of curvature. First we drag lines from centroid to every edge pixel and then measure distance and angle from centroid to every edge pixel. After that, we plot the graph with angle along the x-axis and distance on the y-axis. From the graph, we can notice difference of curvature because of the distinct shape of malignant and benign tumors. We also do the normalization to find the percentage of

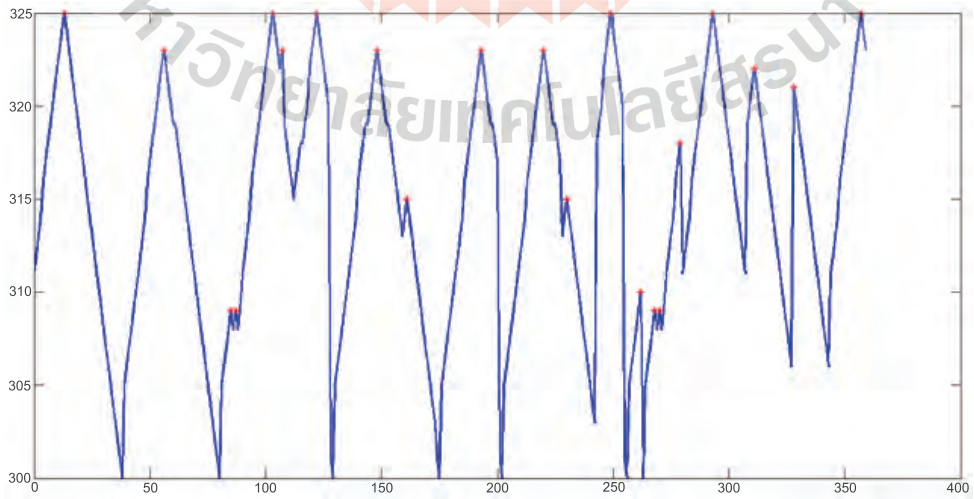
curvature. As a result, we get the different percentage of curvature between malignant and benign cases. We observe that malignant tumor shows many curves along its contour and we can get the percentage of peak in this graph. On the contrary, benign tumor has less curves than the malignant contour. Figure 4 illustrates example of curvature measurement. Figure 5 shows the different graph of curvature between malignant and benign contours.

(a) (b)

FIGURE 4. Measuring the curvature: (a) malignant shape, (b) benign shape



(a)



(b)

FIGURE 5. Graph of curvature: (a) malignant contour, (b) benign contour

3). Intensity Histogram Features

The shape of the intensity histogram features provides much information to describe the properties of the image. Six statistic features obtained from the histogram are mean, variance, skewness, kurtosis, energy, and entropy. The mean is the average intensity level, whereas the variance is the variation of intensities around the mean. The skewness shows whether the histogram is symmetric. The histogram is symmetrical if the skewness is zero.

2.4. **Classification.** We use support vector machine (SVM) with radial basis function (RBF) kernel to classify the mammogram images using three features including texture, shape (percentage of curvature), and intensity histogram. In the SVM training process, we train SVM with 56 images (70% of 80 images selected from the DDSM database). In the evaluation process, the rest 24 images are used for testing. Training and testing images have been preprocessed through the same steps. We also use Naïve Bays and artificial neural network (ANN) in the classification process to compare the performance with SVM.

2.5. **Clustering.** In the clustering process, we use k-means algorithm (k = 2) to cluster the mammogram images. We also use the same three features (texture, shape, intensity histogram) as in the classification process. By means of this feature section process, we have noticed that k-means can accurately cluster images into the correct class.

3. **Experimental Results.** In this proposed work, we use data set from DDSM (digital database for screening mammography). We have selected from DDSM 80 images that include both cases of tumor, that is, malignant and benign (each case containing 40 images). This work has been implemented using MATLAB and R language. We run our experiments on a core i5/2.4 GHZ computer with 4 GB RAM. In the classification process, we compare our proposed method using SVM with Naïve Bays and ANN.

It can be noticed from the classification results summarized in Table 1 that the accuracy on recognizing the benign and malignant images of the SVM (with RBF kernel) shows the highest rate at 88.75%. In other two classification algorithms using Naïve Bays and ANN, the accuracy are 82.50% and 86.25%, respectively. We can conclude from this result that our proposed work using three types of feature and SVM classification has a higher accuracy than Naïve Bays and ANN. We also show in Figure 6 the area under curve (AUC) of the three classifiers: SVM, Naïve Bays, and ANN. As a result, SVM, Naïve Bays, and ANN show AUC value as 0.87, 0.83, and 0.85, respectively. The AUC closer to 1 is the better.

Table 1. Classification results for three learning algorithms

	Accuracy (%)	AUC
SVM (with RBF kernel)	**88.75**	**0.87**
Naïve Bays	82.50	0.83
ANN (artificial neural network)	86.25	0.85

From the result of clustering process using k-means with k = 2 (according to the two classes of images: benign and malignant) which is illustrated in Table 2, we obtain the image recognition accuracy as high as 90.00%. This means that k-means clustering can cluster the data to their actual class accurately. This good clustering result may be due to the effect of image preprocessing steps and the proper setting of cluster number.

Figure 7(a) demonstrates the plot of two cluster components: malignant and benign cases. The two-dimensional clustering plot of the two clusters and lines show the distance between clusters. Clustering shows a good result because it can clearly separate two

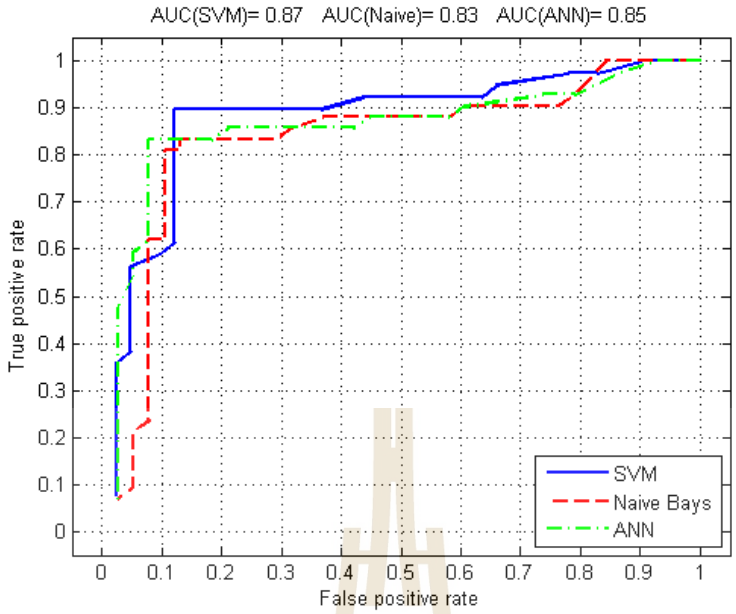
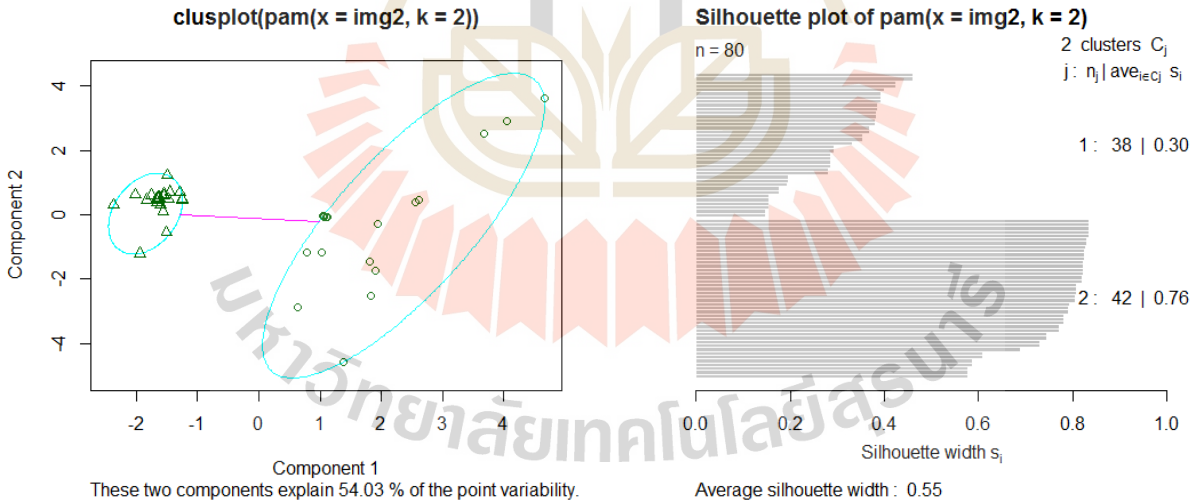


FIGURE 6. Area under curve of three classifiers

TABLE 2. Clustering result using k-means

	Benign	**Malignant**
Cluster 1 (Benign)	**35**	3
Cluster 2 (Malignant)	5	**37**



(a) (b)

FIGURE 7. k-means clustering results: (a) two components of clustering plot, (b) silhouette plot when k = 2

clusters, corresponding to the correct two classes. From the silhouette plot in Figure 7(b), the width of clusters, S_i, are 0.30 and 0.76. The average silhouette width is 0.55.

4. **Conclusions.** Mammography classification using support vector machine with image enhancement and three types of extracted features that we proposed in our framework is the main contribution of this paper. Mammography images are obtained from the well-known DDSM database. Image enhancement using gamma correction can improve

contrast of mammogram images to be seen clearly. We extract the region of interest (ROI) using region growing that can help the cropping of only the tumor object and at the same time eliminate the unnecessary background. After the ROI extraction, the three types of image features including texture, shape, and intensity histogram can be constructed. The processed images are then sent as input to the classification process using SVM with RBF kernel. The classification accuracy of SVM (88.75%) is higher than the ANN (86.25%) and Naïve Bays (82.50%) classifiers.

We also apply exactly the same image preprocessing steps but change from the classification algorithms to be the k-means clustering. We have found that k-means can cluster the mammography images correctly. It clusters images into a group of malignant and benign cases with the accuracy as high as 90.00%.

Acknowledgment. The authors would like to express grateful thanks to the reviewers for their useful comments for improving the content and readability of the paper. The first author has been supported by grant from Rajamangala University of Technology Isan.

REFERENCES

[1] A. Oliver, X. Llado, E. Perez, J. Pont, E. Denton, J. Freixener and J. Marti, A statistical approach for breast density segmentation, *Journal of Digital Imaging*, vol.23, no.5, pp.527-537, 2010.

[2] D. Brzakovic, N. Vujovic, M. Neskovic, P. Brzakovic and K. Fogarty, An approach to automated detection of tumors in mammograms, *IEEE Transactions on Medical Image*, vol.9, no.3, pp.233-241, 1990.

[3] S. R. Aylward, B. H. Hemminger and E. D. Pisano, Mixture modeling for digital mammogram display and analysis, *International Workshop in Digital Mammography*, pp.305-312, 1998.

[4] P. K. Saha, J. K. Udupa, E. F. Conant, P. Chakraborty and D. Sullivan, Breast tissue density quantification via digitized mammograms, *IEEE Transactions on Medical Image*, vol.20, no.8, pp.792-803, 2001.

[5] R. Zwiggelaar and E. Denton, Optimal segmentation of mammographic images, *International Workshop in Digital Mammography*, pp.751-757, 2004.

[6] C. H. Wei, S. Y. Chen and X. Liu, Mammogram retrieval on similar mass lesions, *Computer Methods and Programs in Biomedicine*, vol.106, no.3, pp.234-248, 2012.

[7] R. Adam and L. Bischof, Seeded region growing, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.16, pp.641-647, 1994.

[8] W. Deng, W. Xiao, H. Deng and J. Liu, MRI brain tumor segmentation with region growing method based on the gradients and variances along and inside of the boundary curve, *International Conference on Biomedical Engineering and Informatics*, vol.1, pp.393-396, 2010.

[9] C. Huang, Q. Liu and X. Li, Color image segmentation by seeded region growing and region merging, *International Conference on Fuzzy Systems and Knowledge Discovery*, vol.2, pp.533-536, 2010.

Parameter Optimization for Mammogram Image Classification with Support Vector Machine

Keerachart Suksut, Ratiporn Chanklan, Nuntawut Kaoungku, Kedkard Chaiyakhan,
Nittaya Kerdprasop, Kittisak Kerdprasop

Abstract— **Breast cancer is the malignant tumor occurred mostly in women. Even though breast cancer can be fatal, the patient's survival rate could be as high as 90% if it is detected at the early stage of development. Mammography, ultrasound, and magnetic resonance imaging are examples of screening test for breast cancer. However, to precisely and correctly interpret these images, the medical expertise of radiologists is essential. At present with the matured machine learning techniques, computerized methods can be applied to assist tumor diagnosis, such as the classification between benign and malignant types of tumor. We present in this paper the image-preprocessing and the optimized parametric techniques to help improving accuracy of benign-malignant classification from mammogram images. For the image-preprocessing, we used median filter for noise reduction and gamma correction for image brightness adjustment. We also used region growing technique to find the region of interest, then we extracted three groups of potentially discriminative features: texture feature, shape feature, and intensity histogram feature. After the image-preprocessing, we performed parameter optimization using genetic algorithm prior to the classification done by support vector machine. The results showed that with the appropriate feature selection and the optimal parameter adjustment, the support vector machine can improve its accuracy from 89.47% into 92.98% for mammogram image classification.**

Index Terms— **Parameter Optimization, Genetic Algorithm, Mammogram Images Classification, Support Vector Machine.**

Manuscript received September 26, 2016; revised January 16, 2017. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

K. Suksut is a doctoral student with the School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, NakhonRatchasima, Thailand (corresponding author: phone: +66879619062; e-mail: mikaiterng@gmail.com).

R. Chanklan is a doctoral student with the School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, NakhonRatchasima, Thailand (e-mail: arc_angle@hotmail.com).

N. Kaoungku is with the School of Computer Engineering, Suranaree University of Technology, Muang, Nakhon Ratchasima, Thailand (e-mail: nittaya.k@gmail.com).

K. Chaiyakhan is with the Computer Engineering Department, Rajamangala University of Technology Isan, Muang, Nakhon Ratchasima, Thailand (e-mail: kedkarnc@hotmail.com).

N. Kerdprasop is with the School of Computer Engineering, Suranaree University of Technology, Muang, Nakhon Ratchasima, Thailand (e-mail: nittaya.k@gmail.com).

K. Kerdprasop is with the School of Computer Engineering, Suranaree Universiyof Technology, Muang, Nakhon Ratchasima, Thailand (e-mail: kittisakThailand@gmail.com).

I. Introduction

Among diagnosed cancers in women, breast cancer is the most prominent type and it can be deadly. Usually, early tumor diagnosis can improve survival rate and help the preparation for appropriate treatment. Breast cancer detection can be done through the ultrasound screening [1], magnetic resonance imaging [2], and mammography [3]. The background knowledge for screening cancerous cases is that for the benign (or non-harmful) cases, tumor shapes are regularly round and smooth. On the contrary, for the malignant (or harmful) breast cancer cases, tumors tend to demonstrate irregular and undulated shapes [4].

During the last years, many researchers used mammogram images for breast cancer diagnosis. However, the mammogram images always have noise. The effect of noise is that it can blur some important parts in the images (some points or pixels in images that are normal tissue might look like tumor).

Currently, there are many techniques for de-noise (remove noise) such as image enhancement [5], image segmentation [6], and image feature extraction [7]. It can improve the accuracy for classifying between benign and malignant tumors.

At present, there are many efficient automatic techniques for classification such as decision tree learning, artificial neural network, support vector machine, and many more. Among the existing techniques, support vector machine is generally the most accurate one. If we apply techniques for de-noising and then adopt support vector machine algorithm with the optimized parameters for classification, it can intuitively improve performance of mammogram image classification.

In this paper, we thus propose parameter optimization for support vector machine to classify mammogram image. The goal of this research is to improve the breast cancer classification performance. We apply genetic algorithm for parameter optimization (parameters C, epsilon, and gamma to be used in the support vector machine). We pre-process the images by de-noising with the median filter technique, adjusting image intensity with the gamma correction technique, then finding the region of interest to choose only the potential area for cancerous cell detection with region growing technique, and finally performing feature extraction to contain texture feature, shape feature, and intensity histogram.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. Materials and Methods

A. Median Filter

The intuitive idea of median filter is that some pixels in the image may contain noise and this noise can be detected through its extreme value that does not get along with the surrounding pixels. The median filter method [8] to handle noisy pixel is thus to create a small window frame for normalizing a specific pixel value within that frame (in this work, we set the size of a window to be 3x3 pixels). During the filtration process, a small window is moved along the pixel grid within the image. At each position of a window frame, all the pixel values (i.e., nine values for our 3x3 frame) within the frame are sorted. The median pixel value is then used to replace the existing pixel value. Example of a median filter process is illustrated in fig 1.

B. Gamma Correction

Gamma correction [9] can enhance the contrast of the image. It has value between 0 to 1, where 0 means darkness (black color) and 1 means the brightness (white color). Given the parameter γ as the encoding or decoding value, we can compute the value of gamma correction with the formula given in equation (1).

$$Corrected = 255 * \left(\frac{Image}{255}\right)^{\left(\frac{1}{\gamma}\right)} \quad (1)$$

Note that if γ > 1, it is called a decoding gamma in which the shadow in that image will be set darker. For γ < 1, it is called an encoding gamma and will be used to make the dark region to be lighter.

C. Region Growing

Region growing [10] is applied to choose only specific are of interest by merging surrounding areas with similar intensity. The process starts by setting the seed point, which is normally the middle point (or middle pixel) in the image and then compare the intensity value of that point with the intensity values of the neighbor pixels. If the values are in the same class, we then increase the size of the region.

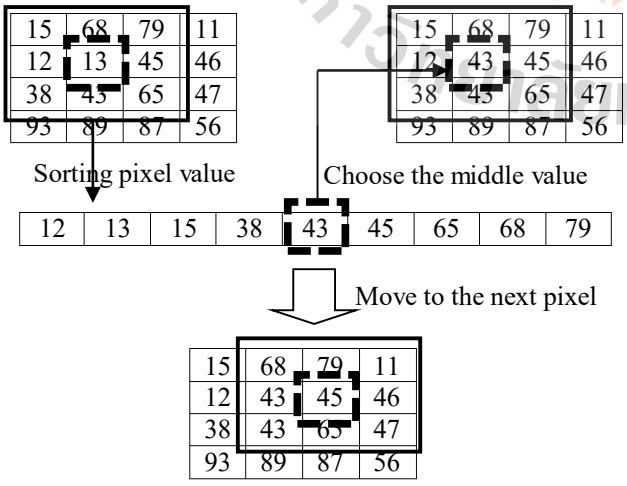


Fig 1. Demonstration of the median filter process

When the growth of one region stops, we then select another seed pixel outside the area previously processed.

D. Texture Feature

Texture feature [11] can help to identify the object in the image. Texture in the image can describe the physical properties (such as shape, curve) and can help to split different objects in an image. We can find texture feature with Grey Level Co-occurrence Matrix (GLCM).

E. Intensity Histogram Feature

Intensity histogram feature is used for describing the properties of the image. In this work, we consider four statistical features that can be obtained from the histogram. These statistics are mean, variance, skewness and kurtosis.

Mean is an average intensity level. Variance is the variation of intensities around the mean. Skewness is the indicator whether the histogram is symmetric, and kurtosis is a measure of whether the data are peak.

Given that G be the image gray scale level and P be the probability level of gray scale, the mean (μ), variance (σ^2), skewness (s), and kurtosis (k) can be computed with formulas given in equations (2) to (5), respectively.

$$\mu = \sum_{i=1}^{G-1} iP(i) \quad (2)$$

$$\sigma^2 = \sum_{i=1}^{G-1} (i-\mu)^2 P(i) \quad (3)$$

$$s = \sigma^{-3} \sum_{i=1}^{G-1} (i-\mu)^3 P(i) \quad (4)$$

$$k = \sigma^4 \sum_{i=1}^{G-1} (i-\mu)^4 P(i) \quad (5)$$

F. Shape Feature

Shape feature [12] can help to identify the object in the image by using shape of object within the image. Shape can differentiate between benign and malignant cases because benign tumors have smooth shapes and regularly round but malignant breast tumors tend to demonstrate irregular and undulated shapes. So, we can classify the object in image by compute the distance between center point in tumor and its edge. For a number of computed distances, if the values do not change or there is only a few change, we can predict that that image is a benign tumor. But if the distance values show much fluctuation, we can predict that the image is malignant tumor.

G. Genetic Algorithm

Genetic algorithm [13-14] is an algorithm to find the solution with adaptive heuristic search based on the evolutionary characteristic of nature. Genetic algorithm combines the concept of random search space and compares the randomly selected solutions based on some fitness function, and then selects the better solution. The simple genetic algorithm is shown in fig 2.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

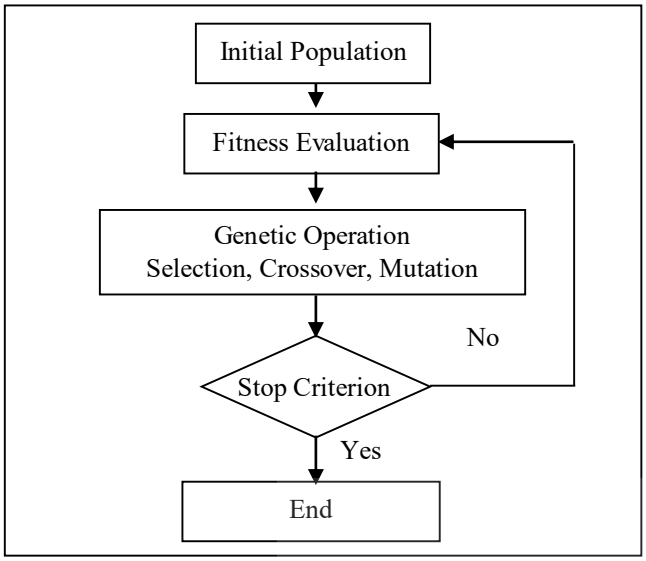


Fig 2. Flowchart for simple genetic algorithm.

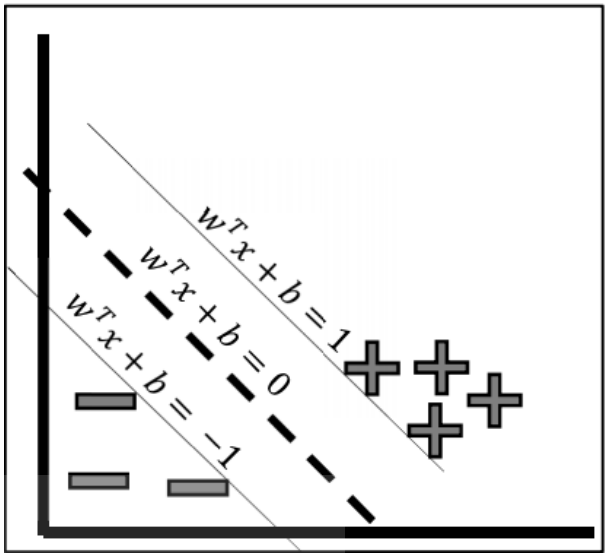


Fig 3. Optimal hyperplane

From fig 2, we can describe genetic algorithm with 5 main steps. Step 1 is setting the initial population; it is normally a random selection. Step 2 is defining the fitness function; it is used for evaluating the fitness of each population or chromosome. Step 3 is applying the genetic operation; the operation can be either selecting the chromosome or population with random selection, crossing over two parent chromosomes to create better offspring, or mutating a chromosome with randomly selected point. Step 4 is replacing individual in the population; it is the replacement of the old chromosome (parent chromosome or parent population) with the new generation. Step 5 is checking for stop criterion; it is a check point for whether to end the process such as stop the process when it has created the new generation over 3 generations.

H. Support Vector Machine

Support Vector Machine (SVM) [15] is a machine learning algorithm for classifying different classes of objects. SVM has been widely applied to many fields. SVM is a supervised learning machine in that it requires a class attribute for guiding the learning process to build a model that can classify objects with mixing classes correctly. The main concept of SVM is the generation of the optimal hyperplane that can separate the objects such that objects with the same class form themselves as a group, whereas objects in different classes should be in a different group. The hyperplane is called an optimal one if such plane can separate classes with the most distance between each class. Fig 3 shows an optimal hyperplane with a dashed line and the two classes in the figure are positive (represented as 1) and negative (-1). To use the hyperplane as a model to classify objects, the formula given in equation (6) can be applied.

$$w^T x + b \geq 1, when\ y_i = +1 \quad (6)$$
$$w^T x + b \leq 1, when\ y_i = -1$$

where
x is data vector,
w is weight vector,
b is bias, and
y is a class.

To apply support vector machine for the classification task, users have to set three important parameters (C, epsilon, and gamma). Parameter C is to control the cost for miss-classification. This parameter is used to control the influence of each individual support vector (i.e., the data points on the borderlines which are up and below the optimal hyperplane in fig 3). Setting the C parameter involves trading error penalty for stability. Parameter epsilon is used to fit the training data. It controls the width of the epsilon-insensitive zone. The value of epsilon can affect the number of support vectors that are used to find the optimal hyperplane. Parameter gamma is the kernel parameter of the Gaussian radial basis function.

The small gamma implies that the learned model will have the large margin; the hyperplane has large distance between two class borderlines and more flexibility in data classification. The large gamma means that learned model will have small margin; the hyperplane has small distance between two class borderlines and thus no flexible in new data classification (may cause overfitting).

III. Proposed Work

In the proposed work, we have designed the process of parameter optimization with genetic algorithm for mammogram image classification with the support vector machine as shown in fig 4.

From fig 4. We can describe our proposed framework as follows. For pre-processing images, we used median filter method for de-nosing, the output from this process is clearer image without noise. After that, we use gamma correction to enhance contrast of the image, the output from this step is sharp image such that the tumor area has lighter intensity and density than the original image. For segmentation process, we use region of interest technique for choosing only region of interest. The output of this process is the smaller image than the original one. A small size means the reduction in dimension to contain only discriminative regions. For feature extraction process, we extract feature

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

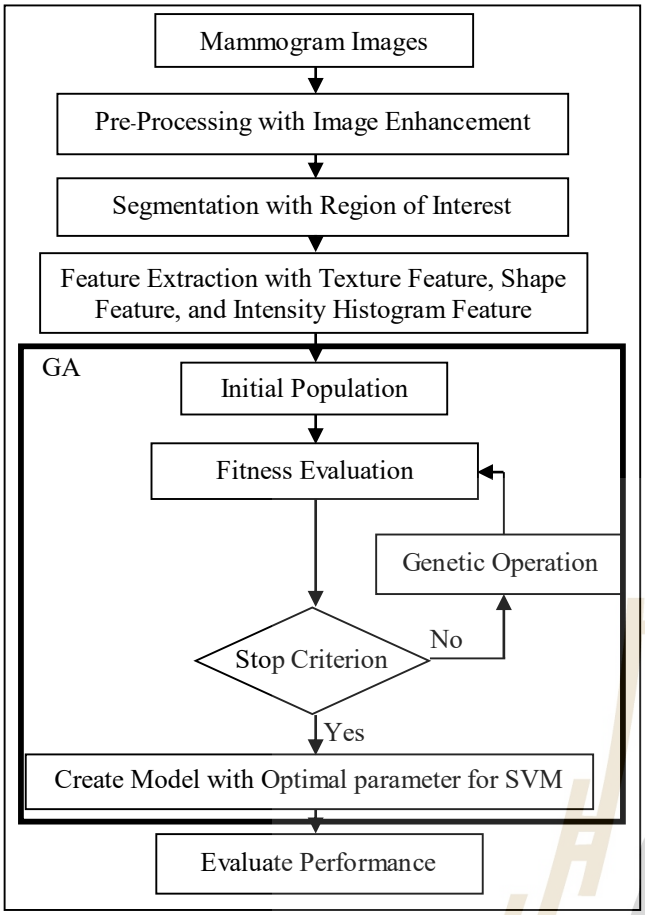


Fig 4. Flowchart of proposed framework for mammogram image classification.

with three techniques (texture feature, shape feature, and intensity histogram feature). The output of this process is the data that extract properties of images (shape, texture, mean, variance, etc.). Then we split the data from previous process into 2 parts. The first part (70% of all data) has been used to find parameter C, epsilon, and gamma with genetic algorithm. This first part of data is also used to create a classification model with support vector machine. The second part (30% of all data) has been used for performance evaluation of the learned model.

In genetic algorithm process, we define parameter control for genetic algorithm as follows:

Population size = 100
Iteration (number of generation) = 100
Probability of crossover = 0.8
Probability of mutation = 0.01
C in the range: $10^{-4} \le C \le 10$
Epsilon in the range: $10^{-2} \le \text{epsilon} \le 2$
Gamma in the range: $10^{-3} \le \text{gamma} \le 3$

Fitness function = $\text{Accuracy} = \frac{TP+TN}{N}$

where
TP is number of true predicted benign cases,
TN is number of true predicted malignant cases, and
N is number of all data that are used to test model.

The output of genetic algorithm is the three parameters that are optimal ones for SVM. After that, we use the optimal parameter to create model with SVM. Finally, we evaluate performance model to assess its accuracy by using the test data. We finally compare the SVM performance with different set of input features.

IV. Experimental Results

For experimentation, we use data set from the Digital Database for Screening Mammography (DDSM) with 190 images (benign 80 images, malignant 110 images) and split data into two parts with 133 images (70% of all data) used for creating a model and finding optimized parameters; we call this data set as “training set”. We use 57 images (30% of all data) for evaluating the performance of classification model; we call this data set as “testing set”. This work has been implemented with MATLAB and RStudio. We run our experiments on a core i3/3.50 GHZ computer with 12 GB of RAM. The details of data after extracting features by using texture feature, shape feature, and intensity histogram are shown in Table 1.

In the classification process, we also compare between different sets of input features that used as input to the support vector machine. We test different combinations of texture feature, shape feature, intensity histogram feature, and the optimized parameter with genetic algorithm for support vector machine. The accuracies of SVM after applying different combinations of input features are shown in Table 2.

From table 2, it can be seen that the adjusted optimal parameters for support vector machine combined with techniques to extract only important features including texture feature, shape feature, and intensity histogram altogether can improve the performance for mammogram

Table 1. Detail of data set

Feature Extraction Techniques	# Training set	# Testing set	# Features
Texture + Shape + Intensity Histogram	133	57	21
Shape + Intensity Histogram	133	57	6
Texture + Shape	133	57	17
Texture + Intensity Histogram	133	57	20

Table 2. Classification results

Feature Extraction Techniques	Accuracy
Texture + Intensity Histogram	81.58%
Texture + Shape	85.26%
Shape + Intensity Histogram	87.37%
Texture + Shape + Intensity Histogram	89.47%
Texture + Shape + Intensity Histogram + Optimized Parameter for SVM with Genetic Algorithm	**92.98%**

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

image classification from the 81.58% accuracy level at 81.58% up to the 92.98%. The classification by SVM using only the extracted features (i.e., the texture feature, shape feature, and intensity histogram) can obtain the highest accuracy at 89.47%. The experimental results show that with an extra steps of optimal parameter adjustment through genetic algorithm, the support vector machine shows an improve performance (from 89.47% to 92.98%) for classification mammogram images.

V. Conclusion

Breast cancer is the major type of dangerous tumors mostly occurred in women and causes numerous deaths in the developing countries. Early detection of malignant breast cancer cases is, more or less, expected to help the appropriate preparation for successful treatment. Breast cancer can be screened with ultrasound imaging, magnetic resonance, or mammogram imaging.

In this work, we propose a framework for automatic classification of malignant breast cancer, the harmful one, from the benign cases, the non-harmful. According to our framework of breast cancer classification with mammogram image, the first step is the noise removal from the mammogram image and the image intensity enhancement. Median filter and gamma correction are the two techniques to de-noise and to enhance contrast of the image, respectively. Region growing technique is then applied to select only area or region of interest. In our work, it is the image regions that are anticipated to contain tumor cells.

We then apply image feature extraction to obtain only important features suitable for the subsequent classification model learning step. The prominent features are texture feature, shape feature, and intensity histogram containing statistical information including mean, variance, skewness, and kurtosis. Another important step in our framework is the application of the genetic algorithm to find the optimal parameters (cost, epsilon, and gamma) for training the support vector machine. The experimental results show that the parameter optimization through genetic algorithm technique can obviously improve the SVM performance for mammogram image classification; it is better than using the default parameters.

References

[1] X. Shi, H.D. Cheng, L. Hu, W. Ju, and J. Tian, “Detection and classification of masses in breast ultrasound images,” Digital Signal Processing, vol. 20, no. 1, pp.824-836, 2010.

[2] M.J. Collins, J. Hoffmeister, and S.W. Worrell, “Computer-aided detection and diagnosis of breast cancer,” Seminars in Ultrasound, CT and MRI, vol. 27, no. 4, pp.351-355, 2006.

[3] A. Oliver, X. Llado, E. Perez, J. Pont, E. Denton, J. Freixenet, and J. Marti, “A statistical approach for breast density segmentation,” Journal of Digital Imaging, vol. 23, no. 5, pp.527-537, 2010.

[4] H. Lee, and Y. Chen, “Image based computer aided diagnosis system for cancer detection,” Expert Systems with Applications, vol. 42, no. 1, pp.5356-5365, 2015.

[5] R. Berannek, W. Jakubowski, A. Mazurczak, M. Postolski, and W. Wiazel, “Contrast enhanced evaluation of the solid lesions in the breast-own experience,” European Journal of Ultrasound, vol. 7, no. 1, pp. S13, 1998.

[6] R. Szeliski, “Computer Vision Algorithms and Applications,” Springer, 2010.

[7] K. Chaiyakhan, N. Kerdprasop, K. Kerdprasop, “Feature selection techniques for breast cancer image classification with support vector machine,” Proceedings of the 24th International Multi Conference of Engineers and Computer Scientists (IMECS2016), Hong Kong, pp.237-232, March 2016.

[8] T. Chen, K. K. Ma, and L. H. Chen. “Tri-state median filter for image denoising. Image Processing,” IEEE Transactions on, vol. 8, no. 12, pp.1834-1838. 1999.

[9] H. Farid, “Blind inverse gamma correction,” Image Processing, IEEE Transactions on, vol. 10, no. 10, pp.1428-1433, 2001.

[10] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, “Benign and malignant breast tumors classification based on region growing and CNN segmentation,” Expert Systems with Applications, vol. 42, no. 1, pp.990-1002, 2015.

[11] A. V. Alvarenga, W. C. A. Pereira, A. F. C. Infantosi, and C. M. Azevedo, “Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images,” Medical Physics, vol. 34, no. 2, pp.379-387, 2007.

[12] W. C. Pereira, A.V. Alvarenga, A. F. Infantosi, L. Macrini, and C. E. Pedreira, “A non-linear morphometric feature selection approach for breast tumor contour from ultrasonic images,” Computer in Biology and Medicine, vol. 40, 2010.

[13] H. Holland, “Adaptation in Natural and Artificial Systems,” Ann Arbor: the University of Michigan Press, Michigan, 1975.

[14] R. A. C. Yang, Z. Zhou, L. Wang, and Y. Pan, “Comparison of Different Optimization Methods with Support Vector Machine for Blast Furnace Multi-Fault Classification,” IFAC-Papers Online, vol. 48, no. 21, pp.1204-1209, 2015.

[15] C. Cortes, and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

Feature Selection Techniques for Breast Cancer Image Classification with Support Vector Machine

Kedkarn Chaiyakhan, Nittaya Kerdprasop, and Kittisak Kerdprasop

Abstract— **Mammography is a special type of low-powered x-ray method that has been used to improve diagnostic and decrease the number of unneeded biopsies. Detection breast cancer in early stage can help treatment successful. Many researches show that malignant breast tumors tend to demonstrate irregular and undulated shapes, whereas benign breast tumors are regularly round and smooth shapes. Consequently, many researches about tumor shape may help in maintaining diagnosis. Thus, the contour feature of tumor contour is very significant feature to distinguish between malignant and benign tumor. In this paper, we propose an approach to automatically appraise the density and contrast of breast images using gamma correction to increase the intensity of dense pixels with light intensity and vice versa to decrease the sparse intensity pixels showing dark intensity. In the segmentation process, we use region growing technique to get region of interest. We also extract three important features including texture, shape, and intensity histogram. In the classification process, we use SVM to classify tumor into two classes: malignant and benign. Moreover, we also compare between three features by combines and separate these features for SVM classification. The results of classification shows that when we combine the shape feature in the classification process, it can be able to correctly classify two types of mammography images and we obtained the high accuracy more than using only texture features and intensity features.**

Index Terms—**feature selection, image classification, mammography, support vector machine.**

I. INTRODUCTION

Breast cancer is a dangerous type of tumor originated from breast tissue, and it accounts for 23% of all cancers in women. The most effective way to detect breast cancer is through the breast mammogram screening, ultrasound images [1]-[7], and also magnetic resonance [5]-[7]. Mammography is the most common imaging technique to detect breast cancer. However, the major limitation for mammography diagnosis is sensitivity due to interpreting mammography is a labor-intensive task for radiologists who cannot always offer stable results during interpreting [8].

K. Chaiyakhan is with the Computer Engineering Department, Rajamangala University of Technology Isan, Muang, Nakhon Ratchasima, Thailand (corresponding author to provide phone: +66868129127; e-mail: kedkarnc@hotmail.com).

N. Kerdprasop is with the School of Computer Engineering, Suranaree University of Technology, Muang, Nakhon Ratchasima, Thailand (e-mail: nittaya.k@gmail.com).

K. Kerdprasop is with the School of Computer Engineering, Suranaree Universiyof Technology, Muang, Nakhon Ratchasima, Thailand (e-mail: kittisakThailand@gmail.com).

The interpreting depends on experience, training, and subjective criteria. Actually, about ten percent of all malignant tumors in mammography are missed by radiologists, and ninety percent of the missed tumors are dense area of breast tissue. It is also admitted that expert radiologists can miss a significant proportion of abnormal tumors. On the contrary, a large number of diagnosed abnormal tumors turn out to be benign after biopsy. Many methodologies have thus been proposed to solve this uncertain interpretation problem by providing assistance to the advanced cancer detection and diagnosis tools

During the last year, several algorithms have been proposed for breast density segmentation. The statistical approach has been proposed by [9]. They provide connected density clusters taking the spatial information of the breast tissue into account. Quantitative and qualitative results show that their approach is able to correctly detect dense breasts apart from other tissue types. A methodology that based on modeling a set of patched of either fatty or dense parenchyma using statistical analysis has been presented by [10]. They analyze two different strategies to perform this modeling process such as principal component analysis and linear-discriminant based model. Once the tissue models have been learned, each pixel of a new mammogram is classified based on neighborhood information as being fatty or dense tissue.

Malignant breast tumors are characterized by cluster of cells indicating uncontrolled outgrowth that leads to penetrate surrounding tissue [11]. The penetration of malignant tumors tends to spread an irregular tumor contour, which will be displayed in mammography as irregular, undulated and ill-defined contour, whereas benign tumors have a uniform outgrowth, round and smooth contour. Hence, it is significant that the contour feature will affect better result of classification.

In our proposed method, we use gamma correction to enhance the image contrast. In segmentation process we use a well-known region growing method to find the ROI and then crop the image to consider only the tumor region. This process will speed up the subsequent classification process because unnecessary background has been removed. After that we extract three types of feature such as texture [12], intensity histogram and shape feature [13]. After that we input digital data to the classification process. The performance of the proposed image classification approach has been evaluated by comparing the accuracy between three features that we extracted after preprocessing image.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. MATERIALS AND METHODS

A. Gamma Correction

Gamma correction is the name of nonlinear operation used to code and decode luminance (or brightness level) on an image. It can also enhance contrast of the image. The luminance value is between 0 and 1, where 0 means absolute darkness (black), and 1 is the brightest (white). Different camera devices do not correctly capture luminance and do not display luminance precisely. Therefore, we need to correct them using gamma correction function. Images which are not corrected can look either light region darker or dark region lighter. Suppose a computer monitor has 2.2 power function as intensity to voltage response. This just means that if we send a message to the monitor that a certain pixel should have intensity equal to x, it will actually display a pixel with intensity equal to $x^{2.2}$. Because the range of voltages sent to monitor is between 0 and 1, it means that the intensity value displayed will be less than what we want it to be. Fig. 1 illustrates the gamma correction model which has been computed from a formula given in (1).

$$Corrected = 255 * (\frac{Image}{255})^{(\frac{1}{\gamma})} \quad (1)$$

where γ is the encoding or decoding value. If value of γ < 1, it is called and encoding gamma or gamma compression, conversely if γ > 1, it is called a decoding gamma or gamma expansion. The effect of gamma correction on an image if γ >1 is that shadow in that image will be darker because the mapping weighs toward lower (darker) output values. If γ < 1, dark region will be lighter because the mapping biases toward higher (brighter) output values. Fig. 2 illustrates this relationship. The two transformation curves show how values are mapped when gamma that is less than and greater than 1. In each graph, the x-axis demonstrates the intensity values of the input image, and the y-axis is the intensity values in the output image.

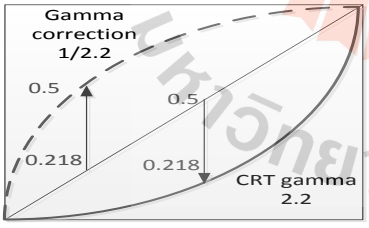


Fig. 1. Gamma correction model.

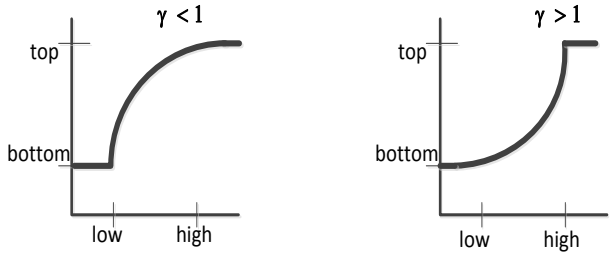


Fig. 2. Two different gamma correction settings.

B. Region Growing

Region growing is a simple region-based image segmentation method using pixel information to adjust the seed point initialization. Small areas in an initial set are iteratively merged according to similarity constraints. The seed point selection starts by choosing an arbitrary pixel and compare it with neighboring pixels that have similar value. After that, increase the size of the region. When the growth of one region stops, then simply choose another seed pixel that does not yet belong to any region and start the process again. The process stops when all pixels belong to some region. Fig. 3 shows the example of region growing.

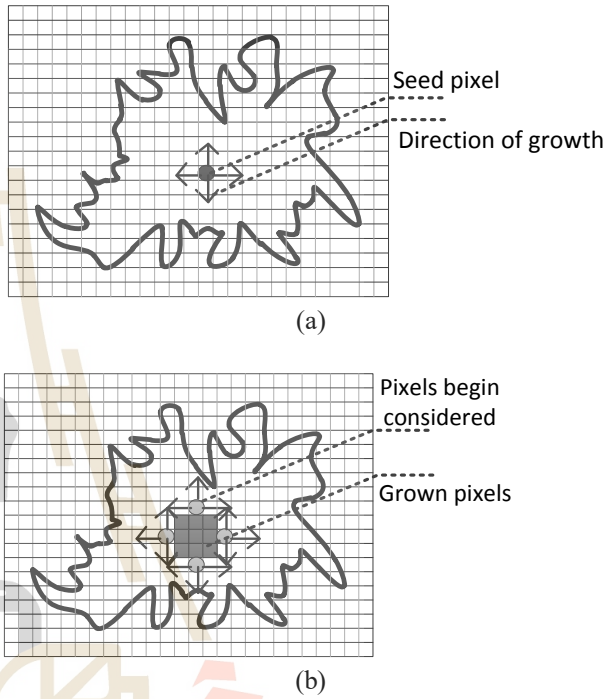


Fig. 3. The example of region growing.

Region growing determines the region of object directly. The basic formulation is shown in (2). This equation states that the segmentation completes when every pixel is in a region and the points in the regions must be disjoint. Equation (3) states the property that the pixels must be in a segmented region. Equation (4) constrains that regions R_i and R_j are different in the sense of predicate H.

$$R = \cup_{i=1}^{S} R_i \qquad R_i \cap R_j = 0 \qquad i \neq j \quad (2)$$

$$H(R_i) = TRUE \qquad i = 1, 2, \ldots, S \quad (3)$$

$$H(R_i \cup R_i) = FALSE \quad i \neq j, \quad R_i \ \ adjacent\ to\ R_j \quad (4)$$

C. Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm used for classification and regression problems. SVM classifies objects by generating the optimal separation in a multi-dimensional space called a hyperplane. In Fig. 4, two parallel separation lines are constructed on each side of the datasets. The optimal hyperplane is the one

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

that maximizes the distance between the two parallel hyperplanes. An assumption is made that the larger of this margin, the better of data classification.

We consider 2 datasets of the form in (5).

$$D = \{ (x_1, y_1), \dots, (x_l, y_l) \}, x_i \in R^m, y_i \in \{-1, 1\} \quad (5)$$

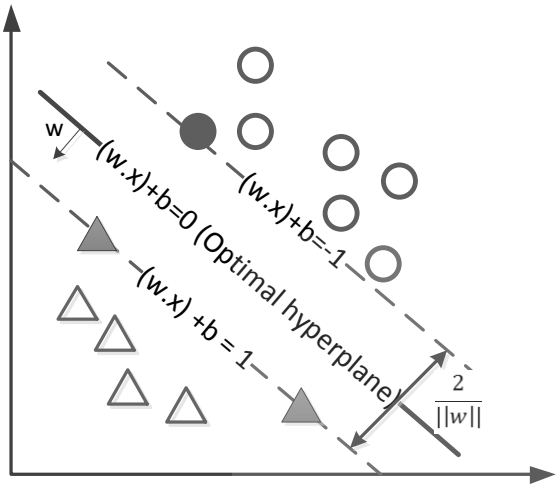


Fig. 4. Optimal hyperplane with maximum margin.

where l denotes the total data instances, i denotes the sequence of data, m is number of dimensions, and y is a class label (+1 or -1) to denote each group of data after separation process. If the training data are linearly separable, we classify each data instance as either positive, or negative based on the computation given in (6). In this equation, w denotes weight of data vector on the separation line, x_1 is positive data vector, and x_2 is negative data vector.

$$(w * x_1) + b > 0 \text{ where, } y_i = +1 \quad (6)$$
$$(w * x_2) + b < 0 \text{ where, } y_i = -1$$

III. Proposed Work

In the proposed work, we have divided our process into five main parts: image preprocessing, segmentation, feature extraction and classification. Fig. 5 shows the framework of this research.

A. Image Preprocessing

Mammogram images usually have noises due to disturbances like Gaussian noise or some little darkness and brightness noise called salt and pepper noise. In this paper, we use median filter to remove these noises. Median filter is a nonlinear method effectively used for removing noise while retaining edges. It works by moving the little window called filter that moves pixel by pixel through the image and changes the pixel value to be the median of neighboring pixels. The median is calculated by first sorting all the pixel values from the filter into numerical order, and then picking the middle pixel value. The output of this de-noising step is the clearer image without noise.

The next step of image preprocessing is image enhancement. We adjust the brightness and darkness of images using gamma correction algorithm. Fig. 6 shows the original images of malignant and benign cases comparing to the improved results after applying the gamma correction technique. The gamma correction helps contrasting the tumor area from the fatty area. In Fig. 6(b), we can see that the tumor area has lighter intensity and density than the original image. In Fig. 6(d), after gamma correction process, the area of benign tumor is lighter than original image. If we compare between Fig. 6(b) and Fig. 6(d), they are rather different because Fig. 6(b) is the malignant case and it has more light and dense intensity pixels than those in Fig. 6(d) which is a benign case.

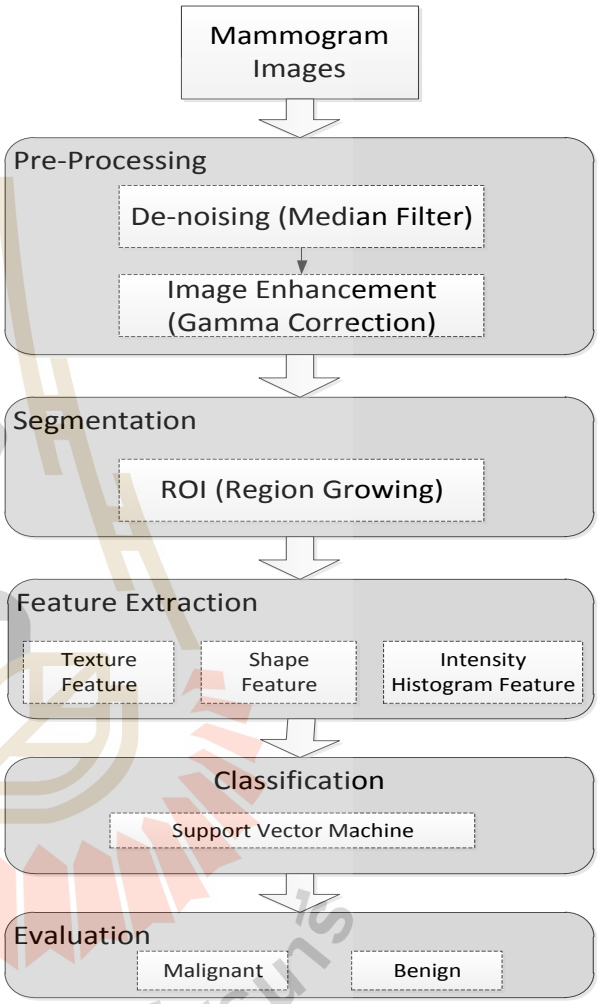


Fig. 5. The framework of the proposed system.

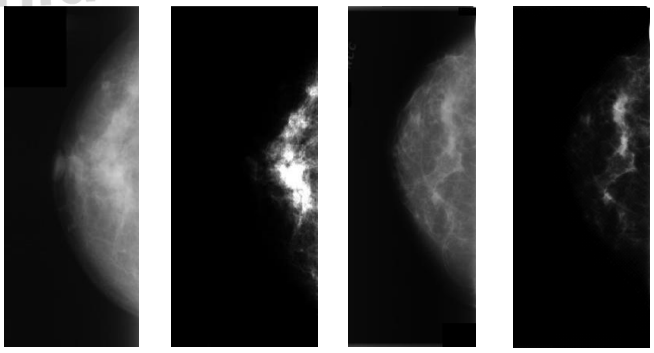


Fig. 6. Breast tumor images: (a) original malignant tumor, (b) malignant tumor after gamma correction, (c) original benign tumor, (d) benign tumor after gamma correction.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

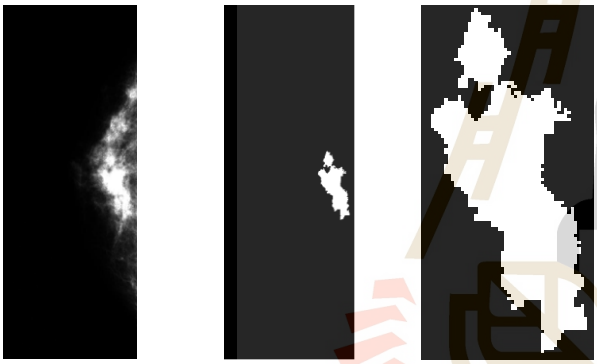
B. Segmentation

The segmentation process separates the tumor areas from the background tissue in mammogram images. In this step, we apply the region growing segmentation method. Region growing is a region-based method starting with seed points in the image, and then propagating seeds until the specified stopping criteria are satisfied. Appropriate seed point selection is important. Therefore, in our proposed work, we select seed point using the centroid of object computed from area and position of object (centroid), as shown in (7) and (8).

$$Area = \sum_{i=1}^{m}\sum_{j=1}^{n} W[i,j] \tag{7}$$

$$Centroid \quad \bar{x} = \frac{\sum_i \sum_j jW[i,j]}{Area} \; \bar{y} = \frac{\sum_i \sum_j iW[i,j]}{Area} \tag{8}$$

where W is the white pixel in the image and i, j are the position of white pixel. After the region growing process, we will get the region of interest (ROI, white pixels) and then we crop only the ROI (Fig. 7) to removing background that may affect the classification and clustering process.



(a) (b) (c)

Fig. 7. The result of region growing and the cropped image: (a) gamma corrected image, (b) the image after applying region growing technique, (c) cropped image.

C. Feature Extraction

The objective of feature extraction step is to represents the image in its reduced and compact form in order to facilitate and speed up the decision making process such as classification and clustering. In this paper, we extract three types of features: texture, shape, and intensity histogram features.

1) Texture Features

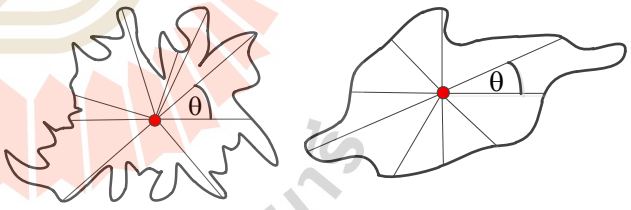
Texture is one of the important features used in identifying objects in an image. Texture features are based on gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The GLCM function characterizes the texture of an image by calculating how often pairs of pixels with specific values and in a specified spatial relationship occur in an image. We create a GLCM, and then extract statistical measures from this matrix such as contrast, correlation, and homogeneity in four directions (0^o, 45^o, 90^o, 135^o). We use these properties of texture to input into the classification process.

2) Intensity Histogram Features

The shape of the intensity histogram features provides several information to describe the properties of the image. Six statistic features obtained from the histogram are mean, variance, skewness, kurtosis, energy, and entropy. The mean is the average intensity level, whereas the variance is the variation of intensities around the mean. The skewness shows whether the histogram is symmetric. The histogram is symmetrical if the skewness is zero. For asymmetric cases, it is skewed above the mean if the skewness is positive, and skewed below the mean if the skewness is negative. The kurotosis is a measure of whether the data are peaked or flat relative to a normal distribution. Data with high kurtosis tend to have a distinct near the mean, and having heavy tails. Data with low kurtosis tend to have a flat top near the mean. Entropy is a metric to measure magnitude of disorder in a system.

3) Shape Features

In this process, we extract shape feature using the percentage of curvature. First we drag lines from centroid to every edge pixel and measure distance and angle from centroid to every edge pixel. After that, we plot the graph with angle along the x-axis and distance on the y-axis. From the graph, we can notice difference of curvature due to the distinct shape of malignant and benign tumor. We also do the normalization to find the percentage of curvature. As a result, we get the different percentage of curvature between malignant and benign tumor. We observe that malignant tumor shows many serrate along its contour and we can get the higher percentage of peak in this graph. In contrast, in the case of benign tumor, it has fewer serrate than the malignant contour. Fig. 8 illustrates example of curvature measurement. Fig. 9 shows the different graph of curvature between malignant and benign contour.



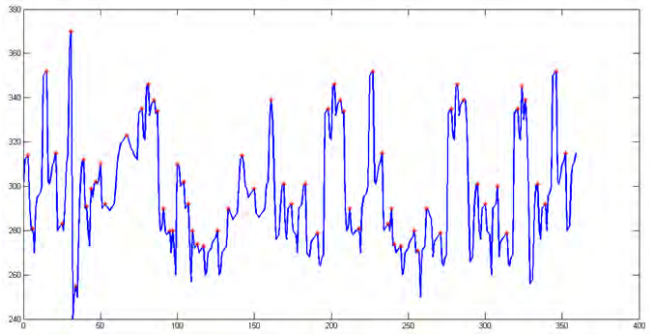
(a) (b)

Fig. 8. Measuring the curvature: (a) malignant contour (b) benign contour.

D. Classification

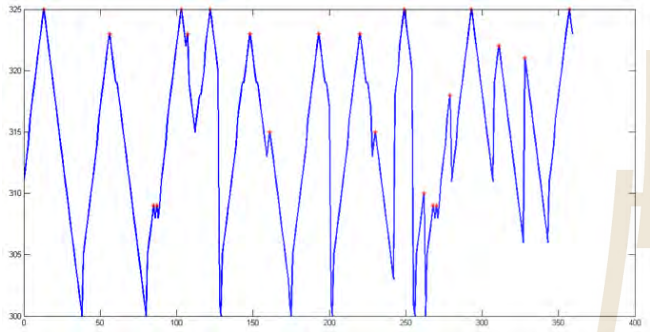
In this research work, we use Support Vector Machine with RBF kernel function to classify the mammogram images using three features including texture, shape (percentage of curvature), and intensity histogram. In the SVM training process, we train SVM with the 133 images (70% of 190 images selected from the DDSM database). In the classification evaluation process, 57 images are used for testing. Training and testing images have been preprocessed through the same steps.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)



……..1001001100101010010001101110001110110001111……..

(a)



……..1000000000001000000000001000001000000000100……..

(b)

Fig. 9. Graph of curvature: (a) malignant contour (b) benign contour.

IV. EXPERIMENTAL RESULTS

In this proposed work, we use data set from DDSM (Digital Database for Screening Mammography). We have selected from DDSM 190 images that include both cases of tumor, that is, malignant and benign (malignant case consists of 110 images and benign case consists of 90 images). This work has been implemented using MATLAB. We run our experiments on a core i5/2.4 GHZ computer with 4 GB RAM.

TABLE I
Classification results between features.

Features	Accuracy (%)	AUC
Texture, Shape, Histogram (TSH)	**89.47**	**0.89**
Histogram, Shape (HS)	87.37	0.87
Texture, Shape (TS)	85.26	0.84
Histogram, Texture (HT)	81.58	0.79

In the classification process, we compare between features using SVM classifier. The results are illustrated in Table I.

It can be noticed from the classification results summarized in Table I that the classification accuracy recognizing the benign and malignant images of the SVM (with RBF – radial basis kernel function) using combination between three features (texture, shape and intensity histrogram) represents the highest rate at 89.47%. In other three combining features as shown in Table I, the accuracy are 87.37%, 85.26% and 81.58%, respectively. We can conclude from this result that our proposed work using three types of feature and SVM classification has a higher accuracy than using only texture feature and intensity histogram feature.

We also show in Fig. 10, the area under curve (AUC) of the four features combination: TSH, HS, TS and HT have the AUC value, 089, 0.87, 0.84 and 0.79, respectively. The higher the AUC value indicates the more precise detection of true positive cases with less inclusion of unwanted false positive.

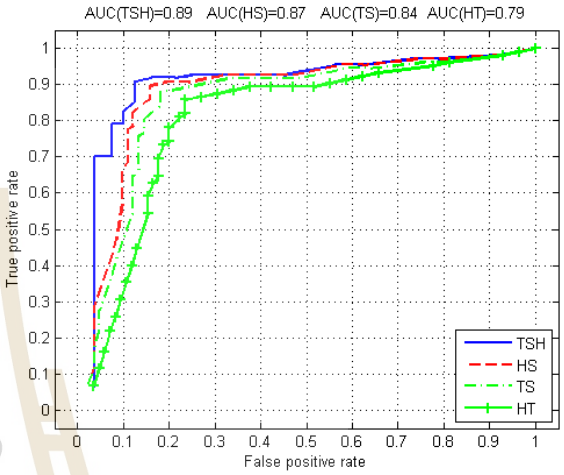


Fig. 10. Area under curve of four combination features.

V. CONCLUSIONS

Mammography classification using support vector machine with image enhancement and three types of extracted features that we proposed in our framework is the main contribution of this paper. Image enhancement using gamma correction can improve contrast of mammogram images to be seen clearly. After the image enhancement process, we extract the region of interest (ROI) using a well-known algorithm called region growing that can help the cropping of only the tumor object and at the same time eliminating the unnecessary background. After the ROI extraction, the three types of image features including texture, shape, and intensity histogram can be constructed. The processed images are then sent as input to the classification process using SVM with RBF kernel. The classification accuracy of SVM using all three features, especially when add the shape feature (89.47%) is higher than the other (87.37%, 85.26% and 81.58%).

Therefore, it is expected that undulated and ill-defined tumors tend to produce higher percentage of curvature than round and regular shapes, as illustrated in Table I. Among combination of features, percentage of curvature showed the most significant feature to distinguishing malignant and benign tumors.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

REFERENCES

[1] S. Huber, J. Danes, I. Zuna, J. Teubner, M. Medl, and S. Delmore, "Relevance of sonographic B-mode criteria and computer-aided ultrasonic tissue characterization in differential diagnosis of solid breast masses," *Ultrasound in Medicine and Biology*, vol. 26, no. 8, pp. 1243-1252, Aug. 2000.

[2] G. Rahbar, A.C. Sie, G.C. Hansen, J.S. Prince, M.L. Melany, H.E. Reynolds, V.P. Jackson, J.W. Sayre, and L.W. Bassett, "Benign versus malignant solid breast masses: US differentiation," *Radiology*, vol. 213, no. 12, pp.889-894, Dec. 1999.

[3] P. Skaane, "Ultrasonography as adjunct to mammography in the evaluation of breast tumors," *Acta Radiologica Supplementum* , vol. 40, no. 420, pp. 1-47, Dec. 1999

[4] M.A. Dennis, S.H. Parker, A.J. Klaus, A.T. Stavros, T.I. Kaske, and S.B. Clark, "Breast biopsy avoidance: the value of normal mammograms and normal sonograms in the setting of a palpable lump," *Radiology*, vol. 219, no. 1, pp.168-191, 2001.

[5] W.A. Berg, L. Gutierrez, M.S. NessAiver, W.B. Carter, M. Bhargavan, R.S. Lewis, and O.B. Ioffe, "Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer," *Radiolo-gy* , vol. 233, no. 3, pp. 830-849, 2004.

[6] M.J. Collins, J. Hoffmeister, and S.W. Worrell, "Computer-aided detection and diagnosis of breast cancer," *Seminars in Ultrasound, CT and MRI*, vol. 27, no. 4, pp.351-355, 2006.

[7] M.L. Giger, "Computerized analysis of images in the detection and diagnosis of breast cancer," *Seminars in Ultrasound, CT and MRI*, vol. 25, no. 5, pp.411-418, 2004

[8] F. Maes, D. Vandermeulen, and P. Suetens, "Medical image registration using mutual information," *Proceedings of the IEEE*, vol. 91, no. 10, pp. 1699-1722, 2003.

[9] A. Oliver, X. Llado, E. Perez, J. Pont, E. Denton, J. Freixener and J. Marti, "A statistical approach for breast density segmentation," *Journal of Digital Imaging*, vol.23, no.5, pp.527-537, 2010.

[10] D. Brzakovic, N. Vujovic, M. Neskovic, P. Brzakovic and K. Fogarty, "An approach to automated detection of tumors in mammograms," *IEEE Transaction in Medical Image*, vol.9, no.3, pp.233-241, 1990.

[11] Y.H. Chou, C.M. Tiu, G.S. Hung, S.C. Wu, T.Y. Chang, and H.K. Chiang, "Stepwise logistic regression analysis of tumor features for breast ultrasound diagnosis," *Ultrasound in Medicine and Biology*, vol. 27, no. 11, pp.1493-1498, Nov. 2001.

[12] A.V. Alvarenga, W.C.A. Pereira, A.F.C. Infantosi, and C.M. Azevedo, "Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images," *Medical Physics*, vol. 34, no. 2, pp. 379-387, 2007

[13] W.C. Pereira, A.V. Alvarenga, A.F. Infantosi, L. Macrini, and C. E. Pedreira, "A non-linear morphometric feature selection approach for breast tumor contour from ultrasonic images," Computer in Biology and Medicine, vol. 40, 2010.

Kedkarn Chaiyakhan is currently a Ph.D. student in the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in Computer Engineering from Rajamangala University of Technology Thanyaburi in 1998 , master degree in Computer Engineering from King Mongkut's University of Technology Thonbuti in 2007. Her current research includes image classification and image clustering.

Nittaya Kertprasop is and associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in Radiation Techniques from Mahidol University, Thailand, in 1985, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes Knowledge Discovery in Databases, Artificial Intelligence, Logic Programming, and Intelligent Databases.

Kittisak Kerdprasop is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A., in 1999. His current research includes Data mining, Artificial Intelligence, Functional and Logic Programming Languages, Computational Statistics.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

MAMMOGRAPHY IMAGES CATEGORIZATION WITH K-MEANS CLUTERING

Kedkarn Chaiyakhan*, Nittaya Kerdprasop, and Kittisak Kerdprasop
School of Computer Engineering, Suranaree University of Technology, Thailand

ABSTRACT

Mammography is an extraordinary type of low-powered x-ray process that provides detailed images of the internal structure of the breast. Many researches show that the dense masses in the breast density are one of the strongest indicators of developing breast cancer. In this paper, we proposed an approach to automatically appraise the density of breast using gamma correction to increase the intensity dense pixels which has light intensity vice versa decrease the intensity sparse pixels which has dark intensity. In clustering process we use k-means clustering to cluster image into 3 categories: benign, malignant and normal. The result shows that our approach be able to cluster three type of mammography after gamma correction process in the correct class which has rather high accuracy.

1. INTRODUCTION

Breast cancer is a type of cancer origination from breast tissue, and it accounts for 23% of all cancers in women. The most effective way to detect breast cancer is through the breast mammogram screening. However, the major limitation for mammography diagnosis is sensitivity. Mammography is the most common imaging technique to detect breast cancer. Many methodologies have been proposed to solve the problem providing assistance on the advanced cancer detection and diagnosis tools.

During the last year, different algorithms have been proposed for breast density segmentation. For instance, Oliver. et al.(2010), proposed a statistical approach for breast density segmentation They provide connected density clusters taking the spatial information of the breast into account. Quantitative and qualitative results show that their approach is able to correctly detect dense breasts, segmentation the tissue type accordingly. Brzakovic. et al. (2009), was presented a methodology that based on modeling a set of patched of either fatty or dense parenchyma using statistical analysis. They analyzed two different strategies to perform this modeling process such as principal component analysis and linear-discriminant-based model. Once the tissue models have been learned, each pixel of a new mammogram is classified as being fatty or dense tissue, taking its corresponding neighborhood into account. Ferrari, et al. (2004) and Aylward, et al. (1998), used mixtures of Gaussian for modeling and segmentation the breast into four and five regions, respectively. However, these related approaches do not take spatial information into account providing segmentations with too many disconnected regions. Moreover, an initial pre-processing step is needed to remove noisy pixels. Aiming to include this spatial information into account, Saha, et al. (2001), included a fuzzy affinity function in their proposed work, while Zwiggelarar (2004), employed textural features to take the spatial distribution of the pixel and its neighborhood into account. Shi, et al. (2010), presented fuzzy support vector machine to automatically detect and classify mass using ultrasound images. They also provided the feature extraction and feature selection using image preprocessing and membership value, respectively.

In this paper, we proposed the clustering method using k-means clustering, we also used image preprocessing technique algorithm namely gamma correction. After preprocessing process, we input the data into k-means clustering which set k=3, since we know that each image belong to one of three classes from the well-known DDSM database which annotated from the experts. In the experimental result shows that our purposed work has capability to categorized the images correctly, with pretty high accuracy that illustrated by the confusion matrix, the cluster plot and the silhouette plot.

2. METERIALS AND METHODS

2.1 Gamma Correction

Gamma correction is the name of nonlinear operation used to code and decode luminance on image systems. Each pixel in an image has brightness level, called luminance. This value is between 0 to 1, where 0 means absolute darkness (black), and 1 is brightest (white). Different camera devices do not correctly capture luminance and do not display luminance precisely. So, we need to correct them using gamma correction function. Gamma correction function is used to correct image's luminance. It controls the whole brightness of an image. Images which are not corrected can look either light region darker or dark region lighter. Suppose a computer

monitor has 2.2 powers function as intensity to voltage response curve. This just means that if we send a message to the monitor that a certain pixel should have intensity equal to x, it will actually display a pixel which has intensity equal to $x^{2.2}$. Because the range of voltages sent to monitor are between 0 and 1, it means that the intensity value displayed will be less than what we wanted it to be. Hence, the gamma corrected formula is written as

$$Corrected = 255 * (\frac{Image}{255})^{(\frac{1}{\gamma})} \quad (1)$$

where γ is the encoding or decoding value. If value $\gamma < 1$ is called and encoding gamma or gamma compression, conversely if $\gamma > 1$ is called a decoding gamma or gamma expansion. The effect of gamma correction on an image if $\gamma > 1$ shadow in image will be darker, whereas, if $\gamma < 1$ dark region will be lighter.

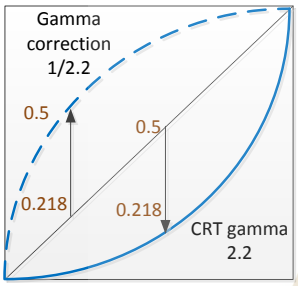


Fig. 1 Gamma correction model.

2.2 K-means Clustering

K-means clustering is one of the easiest unsupervised learning algorithms that solve the clustering problem. The procedure follows a simple and uncomplicated way to cluster a given data set through a certain number of clusters (suppose k clusters). The main concept is to determine k centers, one of each cluster. These centers should be locating them in the brilliant way because of different location causes different result. Therefore, the optimal choice is to locate them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. Clustering data are represented as $D = \{x_1, \dots, x_N\}$. Since the data is p-dimensional, then represent it as $X_n = \{x_{n,1}, \dots, x_{n,p}\}$. The distance function is $d(X_n, X_m)$ between two data points. The k groups has distinguish the data into $\{z_1, \dots, z_N\}$ where $x \in \{1, \dots, K\}$.

3. PROPOSED WORK

In our proposed medical image clustering system, we get benefit from the gamma correction and k-means clustering algorithm. As shown in Fig. 2, the proposed medical image clustering system consists of 6 stages: image acquisition, image resize, gamma correction, image to vector, vector to CSV and clustering images. In the first process we acquired images from DDSM database. Because of each image has very larch size about 3000x5000 pixels which effect long computation time. Thus, in the second process, we resized image to 300x500 pixels.

The main idea of doing the 3 classes (malignant, benign and normal) of image to the different properties is image preprocessing using gamma correction. Because the images from DDSM are gray scale image which 3 classes look rather similar intensity and low contrast. Therefore, we used gamma correction to increase bright pixel and decrease dark pixel. So we will get the different properties of 3 classes image because malignant case has the lighter intensity and dense pixel more than benign case. Likewise, benign image has the lighter intensity and dense pixel more than normal case. In the fourth and fifth process we converted every pre-pressing images to vector and save data in to CSV file, this process make less computation time because no need to read every images in the clustering process. In the last process, we input the CSV file into the clustering process using k-means that set k = 3.

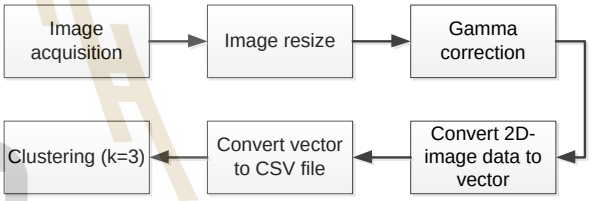


Fig. 2 The framework of the proposed work.

3.1 Image Preprocessing using Gamma Correction

In the image pre-processing process, we adjusted the brightness and darkness of image using gamma correction algorithm. In Fig. 3 shows the result of gamma correction with malignant case, benign case and normal case. In Fig. 3a illustrates the malignant tumor before gamma correction, it seem not clear between the tumor area and fatty area.

In Fig. 3(a), this image is malignant case, after we used gamma correction and get the result in Fig. 3(d), we will see that the tumor area has lighter intensity and density more than original image, that we can input the image after pre-processing in to clustering process using k-means. In Fig. 3(b), and Fig. 3(c) we use the gamma correction process same as Fig 3(a). In Fig 3(b) and Fig. 3(c) are benign tumor and normal tumor respectively. Consequently, we will see the result in Fig 3(e) that it is the benign tumor and after gamma correction process, the area of benign tumor is lighter more than original. If we compare between Fig. 3(d) and Fig. 3(e), they are rather different because Fig. 3(d) is the malignant tumor and it has light intensity pixel and dense intensity pixel more

than Fig. 3(e) that it is benign tumor. Accordingly, in Fig. 3(c), we also apply gamma correction in the image, it is normal case and it has no tumor in this image. Thus the result in Fig. 3(f) has poor light intensity pixel and low dense pixel.

3.2 K-means Clustering on Mammogram Images

After image preprocessing using gamma correction process. We obtained images that corrected brightness and darkness which illustrate in Fig. 3. Subsequently, we input images in clustering process using k-means which set k=3, because after image preprocessing step, the intensity and density of pixels in each image (malignant, benign and normal) rather different. Therefore, k-means can cluster images in the correct class accurately.

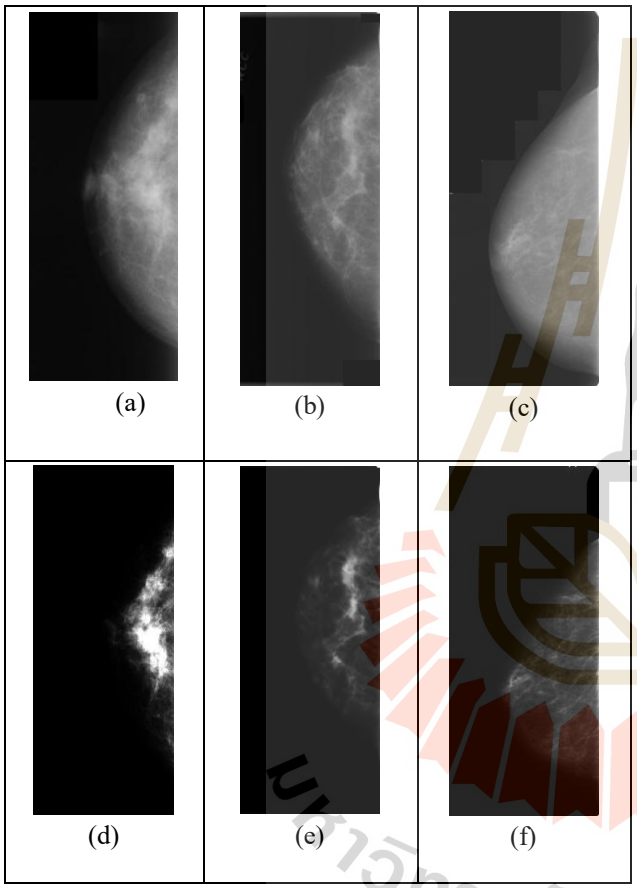


Fig. 3 (a) Original malignant, (b) original benign, (c) original normal, (d) corrected malignant, (e) corrected benign, (f) corrected normal.

4. EXPERIMENTAL RESULTS

In this research we used data set from DDSM (Digital Database for Screening Mammography). We selected 60 images from DDSM that include 3 cases such as malignant, benign and normal (each 20 images). This work was implemented using R language. We run our experiments on a core i5/2.4 GHZ computer with 4 GB RAM. Table 1 shows the result of clustering that pretty good clustering. The clustering process using k-means can clustered the images in a correct class such as benign case can clustered in class 1 which has 18 out of 20, malignant case can clustered in class 2 which has 19 out of 20 and normal case can clustered in class 3 which has 18 out of 20. Consequently, the accuracy rate of our proposed work is 91.67%.

In Fig. 4 demonstrates the two components of 3 clusters plot which are malignant case, benign case and normal case. The two-dimensional clustering plot of the three clusters and lines show the distance between clusters.

The result of clustering seems rather good, because it identifies three clusters, corresponding to three classes. Moreover, in Fig. 5 show their silhouettes plot. From the silhouette plot, the averages S_i are 0.22, 0.85 and 0.74, respectively. According to the silhouette, the first cluster is not well clustered, but the second and third clusters are well clustered. As a result, the average silhouette width is 0.62.

Table 1 Confusion matrix of 3 clusters.

	Benign	Malignant	Normal
1	**18**	0	1
2	2	**19**	1
3	0	1	**18**

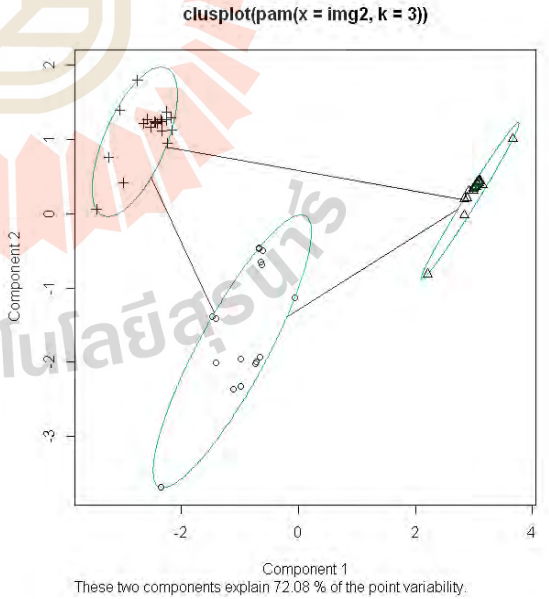


Fig. 4 Two components of clustering plot.

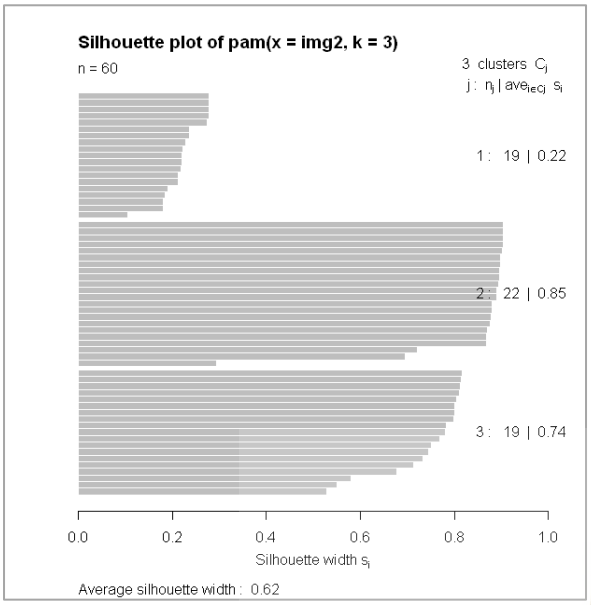


Fig. 5 The silhouette plot when k=3.

5. CONCLUSION

The k-means clustering with gamma correction method that we proposed in this paper can cluster the mammography images from the well-known DDSM database correctly. It clustered images into malignant case, benign case and normal case which has the accuracy 91.67%. Since gamma correction be able to improve the clearness of brightness intensity and it can decrease the poor dark intensity which mean that the area of malignant and benign tumor will appear explicitly and in the normal case which has no tumor area, it appear only fatty which dark intensity pixels. When we input the image after gamma correction process into k-means clustering with k=3, then the k-means be able to cluster the images into correct class, because of an intensity brightness level in images was different.

In our future work we will extract the region of interest (ROI) of tumor using other image preprocessing techniques and we will also use other classification techniques such as support vector machine or artificial neural network to improve the performance of classify and increase the accuracy rate.

REFERENCES

Oliver, A., Llado, X., Perez, E., Pont, J., Denton, E., Freixener, J., and Marti, J., Journal of digital imaging, vol. 23, no. 5, pp. 527-537, 2010.

Brzakovic, D., Vujovic, N., Neskovic, M., Brzakovic, P., and Fogarty, K., IEEE transaction in medical image, vol. 9, no. 3, pp. 233-241, 1990.

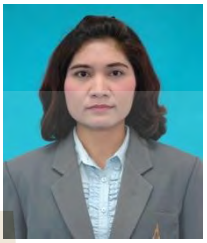
Ferrari, R., Rangayyan, R., Borges, R., and Frere, A., Medical biology engineering computation, vol. 42, pp. 378-387, 2004.

Aylward, S., Hemminger, B., and Pisano, E., International workshop in digital mammography, pp. 305-312, 1998.

Saha, P., Udupa, J., Conant, E., Chakraborty, P., and Sullivan, D., IEEE transaction in medical image, vol.20, no. 8, pp. 792-803, 2001.

Zwiggelaar, R., and Denton, E., International workshop in digital mammography, pp. 751-757, 2004.

Shi, X., Cheng, H., Liming, H., Wen, J., and Jiawei, T., Digital signal processing, vol. 20, pp. 824-836, 2010.



Kedkarn Chaiyakhan is currently a Ph.D. student in the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in Computer Engineering from Rajamangala University of Technology Thanyaburi in 1998 , master degree in Computer Engineering from King Mongkut's University of Technology Thonbuti in 2007. Her current research includes image classification and image clustering.

Nittaya Kertprasop is and associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in Radiation Techniques from Mahidol University, Thailand, in 1985, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes Knowledge Discovery in Databases, Artificial Intelligence, Logic Programming, and Intelligent Databases.

Kittisak Kerdprasop is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A., in 1999. His current research includes Data mining, Artificial Intelligence, Functional and Logic Programming Languages, Computational Statistics.

ภาคผนวก ข

ลิขสิทธิ์โปรแกรม

โปรแกรมจำแนกภาพรังสีเพื่อการวินิจฉัยมะเร็งเต้านม

(Mammography image classification for breast cancer diagnosis program)

ชื่อภาษาไทย	โปรแกรมจำแนกภาพรังสีเพื่อการวินิจฉัยมะเร็งเต้านม
ชื่อภาษาอังกฤษ	Mammography image classification for breast cancer diagnosis program
ทะเบียนข้อมูลเลขที่	ว1. 6552
ให้ไว้ ณ วันที่	12 มิถุนายน พ.ศ. 2560
คำอธิบายโปรแกรมโดยย่อ	ภาพแมมโมแกรมคือภาพถ่ายทางรังสีวิทยาเพื่อใช้ตรวจหามะเร็งเต้านม โปรแกรมจำแนกภาพแมมโมแกรมเพื่อตรวจจับมะเร็ง พัฒนาด้วยโปรแกรม MATLAB (ในส่วนการปรับปรุงภาพ) ร่วมกับโปรแกรมภาษา R (ในส่วนการจำแนกประเภทของภาพ) เพื่อใช้ช่วยในการวินิจฉัยภาพก้อนเนื้อ (Tumor) ว่าเป็นก้อนเนื้องอกชนิดไม่ร้ายแรง (Benign) หรือเป็นก้อนเนื้องอกชนิดร้ายแรง (Malignant) ที่พัฒนาเป็นมะเร็ง (Cancer) ที่เป็นอันตรายถึงชีวิตได้ เทคนิคการจำแนกภาพให้ได้ความแม่นยำสูงสุดของโปรแกรมที่พัฒนาขึ้นนี้ใช้วิธีการปรับปรุงภาพก่อนนำไปจำแนก (Pre-processing) ร่วมกับการปรับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ให้ได้ค่าความแม่นยำของการจำแนกสูงที่สุด การปรับปรุงภาพใช้วิธีการกำจัดสัญญาณรบกวนภายในภาพออกไปแล้วจึงทำการปรับปรุงภาพโดยทำให้ความเข้มสีบริเวณก้อนเนื้อในภาพชัดเจนขึ้น จากนั้นจึงใช้เทคนิคการประมวลผลภาพด้วยวิธีการหาขอบเขตที่น่าสนใจ โดยใช้ขั้นตอนวิธีในการตัดเฉพาะบริเวณก้อนเนื้อในภาพแมมโมแกรมเพื่อนำมาประมวลผล หลังจากได้บริเวณขอบเขตที่สนใจแล้ว ขั้นตอนก่อนการจำแนกอีกขั้นตอนหนึ่งคือการหาลักษณะสำคัญภายในบริเวณขอบเขตที่สนใจ โดยงานวิจัยนี้จะพิจารณาลักษณะสำคัญ 3 ลักษณะ คือ ลักษณะสำคัญของลวดลาย ลักษณะสำคัญของฮิสโตแกรม และ ลักษณะสำคัญของรูปร่าง โดยเฉพาะลักษณะสำคัญของรูปร่างได้มีการเพิ่มชุดข้อมูลต่อท้ายชุดข้อมูลเดิมโดยพิจารณาจากความถี่ของกราฟฮิสโตแกรมของรอยหยักบริเวณเส้นขอบของก้อนเนื้อ และในขั้นตอนสุดท้ายลักษณะสำคัญทั้ง 3 แบบจะถูกนำไปใช้ในการจำแนก ด้วยเทคนิควิธีในการจำแนกข้อมูลแบบมีผู้สอนที่ชื่อว่าซัพพอร์ตเวกเตอร์แมชชีน โดยซัพพอร์ตเวกเตอร์แมชชีนสามารถใช้ร่วมกับเคอร์เนลฟังก์ชันหลายแบบ จากผลการทดสอบพบว่าเคอร์เนล Radial basis function ให้ค่าความแม่นยำสูงที่สุดที่ 93%

สำเนา

รลข.01 ทะเบียนข้อมูลเลขที่ ว1. 6552

หนังสือรับรองการแจ้งข้อมูล

ลิขสิทธิ์

ออกให้เพื่อแสดงว่า

มหาวิทยาลัยเทคโนโลยีสุรนารี

ได้แจ้งข้อมูลลิขสิทธิ์ ประเภทงาน วรรณกรรม

ลักษณะงาน โปรแกรมคอมพิวเตอร์

ชื่อผลงาน โปรแกรมจำแนกภาพรังสีเพื่อการวินิจฉัยมะเร็งเต้านม

ไว้ต่อกรมทรัพย์สินทางปัญญา ตามคำขอแจ้งข้อมูลลิขสิทธิ์ เลขที่ 354899

เมื่อวันที่ 7 เดือน มิถุนายน พ.ศ. 2560

ให้ไว้ ณ วันที่ 12 เดือน มิถุนายน พ.ศ. 2560

ลงชื่อ.....

นางสาวอำพันธ์ เดชสกุลชัย

นักวิชาการพาณิชย์ชำนาญการ

ปฏิบัติราชการแทนผู้อำนวยการสำนักลิขสิทธิ์

หมายเหตุ 1. เอกสารนี้มิได้รับรองความเป็นเจ้าของลิขสิทธิ์

2. การเปลี่ยนแปลงรายการข้างต้น ให้ดูด้านหลัง

ประวัติผู้วิจัย

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ สำเร็จการศึกษาในระดับปริญญาเอกสาขา Computer Science จาก Nova Southeastern University เมือง Fort Lauderdale รัฐฟลอริดา สหรัฐอเมริกา เมื่อปีพุทธศักราช 2542 (ค.ศ. 1999) ด้วยทุนการศึกษาของกระทรวงวิทยาศาสตร์และเทคโนโลยี (หรือชื่อใหม่ในปัจจุบันคือ กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม) หลังสำเร็จการศึกษาได้ปฏิบัติราชการในตำแหน่งอาจารย์ ประจำสาขาคอมพิวเตอร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ต่อมาในปีพุทธศักราช 2543 ได้มาปฏิบัติงานในตำแหน่งอาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี จนถึงปัจจุบัน งานวิจัยที่ทำในขณะนี้คือการประยุกต์เทคโนโลยีเหมืองข้อมูลกับงานด้านการแพทย์ การสาธารณสุขและสิ่งแวดล้อม รวมถึงการพัฒนาเทคนิคเพื่อเพิ่มความสามารถในการจัดการความรู้ของระบบเหมืองข้อมูล

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ สำเร็จการศึกษาในระดับปริญญาเอกสาขา Computer Science จาก Nova Southeastern University เมือง Fort Lauderdale รัฐฟลอริดา สหรัฐอเมริกา เมื่อปีพุทธศักราช 2542 (ค.ศ. 1999) ด้วยทุนการศึกษาของทบวงมหาวิทยาลัย (หรือชื่อในปัจจุบันคือ กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม) หลังสำเร็จการศึกษาได้ปฏิบัติงานในตำแหน่งอาจารย์ ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ปัจจุบันดำรงตำแหน่งหัวหน้าหน่วยวิจัยวิศวกรรมความรู้ เน้นการวิจัยเกี่ยวกับการพัฒนาระบบเหมืองข้อมูลประสิทธิภาพสูง การประยุกต์เหมืองข้อมูลกับงานวิศวกรรมและการวิเคราะห์ข้อมูลเชิงสถิติ รวมถึงการวิจัยพื้นฐานเกี่ยวกับเทคนิคการวิเคราะห์ข้อมูลโดยวิธีอัตโนมัติ โดยมีผลงานวิจัยในด้านฐานข้อมูล การวิเคราะห์ข้อมูล การทำเหมืองข้อมูล และการค้นหาความรู้ ตีพิมพ์ในวารสารวิชาการและเอกสารการประชุมวิชาการทั้งระดับชาติและนานาชาติจำนวนมากกว่า 300 เรื่อง

อาจารย์ ดร.เกตุกาญจน์ ไชยขันธุ์ สำเร็จการศึกษาในระดับปริญญาเอกสาขาวิชาวิศวกรรมคอมพิวเตอร์ จากมหาวิทยาลัยเทคโนโลยีสุรนารี เมื่อปีพุทธศักราช 2559 (ค.ศ. 2017) ด้วยทุนการศึกษาของมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน หลังสำเร็จการศึกษาได้ปฏิบัติงานในตำแหน่งอาจารย์ ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์และสถาปัตยกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ปัจจุบันดำรงตำแหน่งอาจารย์และรองผู้อำนวยการฝ่ายทะเบียนและประเมินผล สำนักส่งเสริมวิชาการและงานทะเบียน งานวิจัยที่สนใจเกี่ยวข้องกับเทคนิคเหมืองข้อมูล และการประมวลผลภาพ โดยมีผลงานวิจัยตีพิมพ์ในวารสารวิชาการและเอกสารการประชุมวิชาการทั้งระดับชาติและนานาชาติ
