การจำแนกภาพแมมโมแกรมโดยใช้การประมวลผลภาพร่วมกับซัพพอร์ตเวกเตอร์แมชชีน

นางเกตุกาญจน์ ไชยขันธุ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2558

MAMMOGRAPHY IMAGE CLASSIFICATION USING IMAGE PROCESSING AND SUPPORT VECTOR MACHINE

Kedkarn Chaiyakhan

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy in Computer Engineering

Suranaree University of Technology

Academic Year 2015

การจำแนกภาพแมมโมแกรมโดยใช้การประมวลผลภาพร่วมกับ
ซัพพอร์ตเวกเตอร์แมชชีน

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

_____
(รศ. ดร.กิตติศักดิ์ เกิดประสพ)
ประธานกรรมการ

_____
(รศ. ดร.นิตยา เกิดประสพ)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

_____
(ผศ. ดร.ศุภกฤษฎิ์ นิวัฒนากูล)
กรรมการ

_____
(ผศ. ดร.ปรเมศวร์ ห่อแก้ว)
กรรมการ

_____
(ผศ. ดร.สายสุนีย์ จับโจร)
กรรมการ

_____
(ศ. ดร.ชูกิจ ลิมปิจำนงค์)
รองอธิการบดีฝ่ายวิชาการและนวัตกรรม

_____
(รศ. ร.อ. ดร.กนต์ธร ชำนิประศาสน์)
คณบดีสำนักวิชาวิศวกรรมศาสตร์

เกตุกาญจน์ ไชยขันธุ์ : การจำแนกภาพแมมโมแกรมโดยใช้การประมวลผลภาพร่วมกับซัพพอร์ตเวกเตอร์แมชชีน (MAMMOGRAPHY IMAGE CLASSIFICATION USING IMAGE PROCESSING AND SUPPORT VECTOR MACHINE)
อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.นิตยา เกิดประสพ, 110 หน้า.

การจำแนกมะเร็งเต้านมจากภาพแมมโมแกรม มีจุดประสงค์เพื่อทำการจำแนกก้อนเนื้อ (Tumor) ภายในภาพแมมโมแกรมว่าเป็นก้อนเนื้อไม่อันตราย (Benign) หรือก้อนเนื้ออันตราย (Malignant) เพื่อประโยชน์ในการช่วยนักรังสีวิทยาวินิจฉัยโรคมะเร็งเต้านม และยังช่วยทำให้ผู้ป่วยได้รู้ผลการวินิจฉัยเบื้องต้นจากภาพแมมโมแกรมโดยไม่จำเป็นต้องมีความเสี่ยงจากการผ่าตัดเพื่อนำชิ้นเนื้อในเต้านมไปตรวจสอบ ในปัจจุบันมีนักวิจัยจำนวนมากพัฒนาประสิทธิภาพของการจำแนกภาพแมมโมแกรมโดยใช้เทคนิควิธีต่าง ๆ ของการประมวลผลภาพร่วมกับเทคนิควิธีการเรียนรู้ของเครื่อง เพื่อเพิ่มความแม่นในการจำแนก การปรับปรุงภาพก่อนการนำไปจำแนก (Preprecessing) เป็นขั้นตอนที่สำคัญเนื่องจากภาพแมมโมแกรมอาจมีความไม่ชัดเจนหรือมีสัญญาณรบกวนในภาพทำให้การจำแนกได้ผลที่ไม่ดีนัก

ดังนั้นงานวิจัยของวิทยานิพนธ์ฉบับนี้เสนอวิธีการปรับปรุงภาพคือการกำจัดสัญญาณรบกวนภายในภาพออกไป แล้วจึงทำการปรับปรุงภาพโดยทำให้ความเข้มสีบริเวณก้อนเนื้อในภาพชัดเจนขึ้น จากนั้นจึงใช้เทคนิคการประมวลผลภาพด้วยวิธีการหาขอบเขตที่น่าสนใจ โดยใช้ขั้นตอนวิธีในการตัดเฉพาะบริเวณก้อนเนื้อในภาพแมมโมแกรมเพื่อนำมาประมวลผล หลังจากได้บริเวณขอบเขตที่สนใจแล้ว ขั้นตอนก่อนการจำแนกอีกขั้นตอนหนึ่งคือการหาลักษณะสำคัญภายในบริเวณขอบเขตที่สนใจ โดยงานวิจัยนี้จะพิจารณาลักษณะสำคัญ 3 ลักษณะ คือ ลักษณะสำคัญของลวดลาย ลักษณะสำคัญของฮิสโตแกรม และ ลักษณะสำคัญของรูปร่าง โดยเฉพาะลักษณะสำคัญของรูปร่างได้มีการเพิ่มชุดข้อมูลต่อท้ายชุดข้อมูลเดิมโดยพิจารณาจากความถี่ของกราฟฮิสโตแกรมของรอยหยักบริเวณเส้นขอบของก้อนเนื้อ และในขั้นตอนสุดท้าย ลักษณะสำคัญทั้ง 3 แบบจะถูกนำไปใช้ในการจำแนก ด้วยเทคนิควิธีในการจำแนกข้อมูลแบบมีผู้สอนที่ชื่อว่าซัพพอร์ตเวกเตอร์แมชชีน โดยซัพพอร์ตเวกเตอร์แมชชีนสามารถใช้ร่วมกับเคอร์เนลฟังก์ชันหลายแบบ งานวิจัยนี้จะเปรียบเทียบประสิทธิภาพการจำแนกระหว่างเทคนิควิธีซัพพอร์ตเวกเตอร์แมชชีนกับเทคนิคการจำแนกอื่น ๆ เช่น โครงข่ายประสาทเทียม และ นาอีฟเบย์

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา _____
ปีการศึกษา 2558 ลายมือชื่ออาจารย์ที่ปรึกษา _____

KEDKARN CHAIYAKHAN : MAMMOGRAPHY IMAGE CLASSIFICATION USING IMAGE PROCESSING AND SUPPORT VECTOR MACHINE. THESIS ADVISOR : ASSOC. PROF. NITTAYA KERDPRASOP, Ph.D., 110 PP.

MAMMOGRAPHY CLASSIFICATION/SUPPORT VECTOR MACHINE/ FEATURE SELECTION/IMAGE SEGMENTATION

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 Bays. The results of classification show that SVM gives good classification accuracy more than Artificial Neural Network and Naïve Bays.

School of Computer Engineering Student's Signature _____

Academic Year 2015 Advisor's Signature_____

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงด้วยดี ผู้วิจัยขอกราบขอบพระคุณ บุคคล และกลุ่มบุคคลที่ได้กรุณาให้คำปรึกษา แนะนำ ช่วยเหลืออย่างดียิ่ง ทั้งในด้านวิชาการ และด้านการดำเนินงานวิจัย ดังต่อไปนี้

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ อาจารย์ที่ปรึกษาวิทยานิพนธ์ และรองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ ที่ให้คำปรึกษาในการทำงานวิจัย การจัดรูปแบบวิทยานิพนธ์ และช่วยตรวจทานความถูกต้องของวิทยานิพนธ์

คุณสายฝน สิบพลกรัง เลขานุการสาขาวิศวกรรมคอมพิวเตอร์ ที่ให้ความช่วยเหลือในการประสานงานด้านเอกสารระหว่างศึกษา

ขอขอบคุณนักศึกษาร่วมชั้นเรียนทั้งปริญญาโทและปริญญาเอก ที่ให้คำแนะนำและปรึกษาด้านวิชาการและช่วยเหลือสนับสนุนด้วยดีมาตลอด

ขอบคุณมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ที่ให้การสนับสนุนทุนการศึกษา ทุนวิจัย และค่าใช้จ่ายต่าง ๆ

นอกจากนี้ขอขอบคุณ ครู อาจารย์ ทั้งในอดีตและปัจจุบันที่ให้ความรู้แก่ผู้วิจัยจนประสบความสำเร็จในชีวิต

ท้ายที่สุดขอกราบขอบพระคุณ บิดา มารดา ที่ให้กำเนิด อบรม เลี้ยงดูและส่งเสริมการศึกษาเป็นอย่างดีทำให้ผู้วิจัยมีความรู้ ความสามารถ มีจิตใจที่เข้มแข็ง รวมทั้งเป็นกำลังใจแก่ผู้วิจัยจนทำให้ผู้วิจัยประสบความสำเร็จในชีวิต

เกตุกาญจน์ ไชยขันธุ์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

มะเร็งเต้านม (Breast Cancer) เป็นสาเหตุสำคัญของการเสียชีวิตมากที่สุดของประชากรเพศหญิงทั่วโลก เมื่อเปรียบเทียบอัตราการเสียชีวิตกับมะเร็งชนิดอื่น ๆ ที่เกิดขึ้นในเพศหญิง (ผลสำรวจโดยหน่วยงาน World Cancer Research Fund International ข้อมูลจาก http://www.wcrf.org) ทั้งนี้เป็นที่ทราบกันดีว่าวิธีการที่จะลดอัตราการเสียชีวิตจากมะเร็งเต้านมที่ดีที่สุดคือการตรวจวินิจฉัย (Diagnosis) ก้อนเนื้อ (Tumor) ที่มีความผิดปกติในระยะแรก โดยทำการวินิจฉัยว่าก้อนเนื้อที่เกิดขึ้นในเต้านมนั้นเป็นก้อนเนื้อที่ไม่อันตราย (Benign) หรือเป็นก้อนเนื้อร้าย (Malignant)

ก้อนเนื้อที่ไม่อันตราย หมายถึง ก้อนเนื้อที่ไม่แพร่กระจายไปยังอวัยวะอื่น ส่วนก้อนเนื้อร้ายเป็นก้อนเนื้อที่สามารถแพร่กระจายไปยังอวัยวะอื่น และก่อให้เกิดการแบ่งเซลล์ที่ไม่สามารถควบคุมได้ทำให้เกิดเป็นมะเร็ง ก้อนเนื้อที่มีความน่าจะเป็นที่จะเป็นมะเร็งเต้านมนั้นเกิดจากความผิดปกติของเซลล์ในร่างกายมนุษย์ ซึ่งเซลล์ที่ผิดปกตินั้นจะมีการเปลี่ยนแปลงและมีขนาดใหญ่ขึ้นเรื่อย ๆ หากปล่อยไว้โดยไม่มีการตรวจวิเคราะห์และรักษาอาจทำให้ผู้ป่วยเสียชีวิตได้ (Huo et al., 2000; Wu et al., 1995) ส่วนใหญ่นั้นเซลล์มะเร็งจะมีลักษณะที่เป็นก้อนเนื้อที่มีความหนาแน่นสูง ตรวจสอบได้ด้วยการผ่าตัดเพื่อนำชิ้นเนื้อไปตรวจสอบ (Biopsy) หรือการวิเคราะห์จากภาพแมมโมแกรม (Mammography) (Oliver et al., 2010; Xie et al., 2015; Pak et al., in press) หรือ ภาพอัลตราซาวด์ (Ultrasound) (Shi et al., 2010; Kumar et al., 2015; Zhou et al., 2013) มะเร็งเต้านมเป็นมะเร็งที่ไม่แสดงอาการในเบื้องต้น ผู้หญิงส่วนใหญ่จึงรู้สึกถึงอาการของมะเร็งเต้านมในขณะที่มะเร็งได้ลุกลามไปในขั้นที่สูงแล้ว ดังนั้นหากมีการวิเคราะห์และวินิจฉัยก้อนเนื้อว่าเป็นเนื้อร้าย หรือเนื้อไม่อันตรายในเบื้องต้นแล้ว ก็จะทำให้ผู้ป่วยได้มีโอกาสในการรักษาและหายจากอาการป่วยและรอดชีวิตค่อนข้างสูง ในการตรวจวิเคราะห์ก้อนเนื้อในเต้านมนั้นมีหลายวิธี เช่นการผ่าตัดนำเอาก้อนเนื้อออกมาพิสูจน์ ซึ่งวิธีนี้จะค่อนข้างยุ่งยากและค่อนข้างอันตรายต่อคนไข้ ดังนั้นอีกวิธีหนึ่งที่นิยมใช้กันมากในปัจจุบันคือการนำภาพแมมโมแกรม และภาพอัลตราซาวด์ของเต้านมมาวิเคราะห์หาความผิดปกติของก้อนเนื้อ (Chen et al., 2015; Dhahbi et al., 2015)

มะเร็งเต้านมเป็นมะเร็งที่พบบ่อยมากกว่ามะเร็งปากมดลูก มะเร็งเต้านมนั้นสามารถพบได้ในผู้ชายเช่นกัน แต่พบในอัตราส่วนที่น้อยมาก มะเร็งเต้านมก่อกำเนิดขึ้นจากเนื้อเยื่อ (Tissue) ใน

ทรวงอก ซึ่งโดยทั่วไปแล้วจะเกิดขึ้นภายในต่อมน้ำนม (Mammary Gland) การวิเคราะห์เนื้อเยื่อภายในทรวงอกที่มีประสิทธิภาพมากในปัจจุบันคือวิธีการตรวจภาพรังสีเต้านมหรือแมมโมแกรม ซึ่งวิธีนี้เป็นการนำภาพแมมโมแกรมมาวิเคราะห์โดยนักรังสีวิทยา (Radiologist) ซึ่งโดยทั่วไปแล้วเนื้อเยื่อในทรวงอกที่มีความผิดปกตินั้น จากการวิเคราะห์ของนักรังสีวิทยาจะเป็นก้อนเนื้อที่มีความหนาแน่นสูง (Dense) และมีก้อนหินปูน (Calcified) เกาะอยู่ด้วย ซึ่งเนื้อเยื่อหรือก้อนเนื้อที่มีความผิดปกตินี้สามารถเป็นได้ทั้งก้อนเนื้อที่ไม่อันตรายและก้อนเนื้อร้ายที่สามารถก่อตัวเป็นมะเร็งเต้านม การวินิจฉัยขึ้นอยู่กับรูปร่างของก้อนเนื้อที่ตรวจพบ (Aguiar et al., 2015; Chen et al., 2015) จากความรู้ของนักรังสีวิทยา สรุปได้ว่าก้อนเนื้อที่ไม่อันตรายส่วนใหญ่มักจะมีรูปร่างที่เป็นทรงกลมหรือทรงรีและมีขอบที่ค่อนข้างเรียบมีความหยักน้อย ในขณะที่ก้อนเนื้อร้ายที่ก่อตัวเป็นมะเร็งเต้านมนั้นจะมีขอบบางส่วนที่เป็นทรงกลม และขอบบางส่วนที่มีลักษณะผิดรูปหรือมีรอยหยักค่อนข้างมาก (Lee et al., 2015) และโดยทั่วไปแล้วก้อนเนื้อที่มีความผิดปกตินั้นเมื่อดูจากภาพแมมโมแกรมมักจะมีความสว่างของสี และความหนาแน่นในบริเวณก้อนเนื้อมากกว่าบริเวณข้างเคียงหรือบริเวณที่เป็นไขมัน

อย่างไรก็ตามภาพแมมโมแกรมที่ได้มานั้นอาจยังมีสิ่งรบกวนภายในภาพ (Noise) ซึ่งจะทำให้ภาพไม่ชัดเจน ส่งผลให้การวิเคราะห์ของนักรังสีวิทยามีความผิดพลาดเกิดขึ้นได้ เพราะต้องทำการวิเคราะห์โดยใช้สายตาและความเชี่ยวชาญเฉพาะตน เพื่อลดปัญหาความผิดพลาด ในปัจจุบันได้มีผู้พัฒนาระบบคอมพิวเตอร์ในการตรวจหาและวิเคราะห์โรค ซึ่งเรียกว่า Computer Aided Detection (CAD) (Helena et al., 2015; Jalalian et al., 2013) เป็นเทคนิคที่ช่วยให้นักรังสีวิทยาวิเคราะห์ผลของมะเร็งต่าง ๆ ได้แม่นยำยิ่งขึ้น เทคนิคนี้เป็นการปรับปรุงให้ภาพถ่ายทางการแพทย์มีความชัดเจนและดึงลักษณะสำคัญของภาพ (Zyout et al., in press; Wajid et al., 2015) เพื่อสามารถนำมาวิเคราะห์ได้ละเอียดมากยิ่งขึ้น แต่เทคนิค CAD ยังไม่มีความสามารถจำแนกภาพอัตโนมัติ

ปัจจุบันจึงมีผู้วิจัยนำเสนองานวิจัยที่เกี่ยวกับการจำแนกมะเร็งเต้านมจากภาพแมมโมแกรมโดยใช้อัลกอริทึมที่เกี่ยวกับการเรียนรู้ของเครื่อง (Machine Learning) (Zaki, 2014; Xie et al., in press; Dietzel et al., 2012) ร่วมกับเทคนิคต่าง ๆ ของการประมวลผลภาพ (Image Processing) เช่น การปรับปรุงภาพ (Image Enhancement) (Beranek et al., 1998) การแบ่งขอบเขตในภาพ (Image Segmentation) (Szeliski, 2010) การดึงลักษณะสำคัญของภาพ (Image Feature Extraction) เพื่อมุ่งเน้นให้การจำแนกภาพมีความแม่นยำและประสิทธิภาพที่สูงขึ้น

ดังนั้นงานวิจัยนี้จึงได้เสนอการพัฒนาขั้นตอนวิธีเพื่อจำแนกมะเร็งเต้านมจากภาพแมมโมแกรมด้วยวิธีการประมวลผลภาพเป็นขั้นตอนก่อนการจำแนก (Pre-classification) และใช้วิธีการ

เรียนรู้ของเครื่องเพื่อจำแนกภาพมะเร็งเต้านมได้โดยอัตโนมัติ โดยได้เลือกใช้เทคนิคและอัลกอริทึมดังนี้

- การปรับปรุงภาพแมมโมแกรมจะใช้เทคนิค Grey-level Thresholding เพื่อปรับความเข้มสีของบริเวณวัตถุที่สนใจภายในภาพให้มีความชัดเจนขึ้น
- การแบ่งขอบเขตในภาพจะใช้วิธีการขยายพื้นที่ของส่วนภาพ (Region Growing) เพื่อตัดเฉพาะขอบเขตของก้อนเนื้อที่ผิดปกติ (Region of Interest)
- การดึงลักษณะสำคัญของภาพจะพิจารณาจากลวดลาย (Texture) รูปร่าง (Shape) และความเข้มสี (Intensity) เพื่อใช้เป็นลักษณะสำคัญในการจำแนก (Feature for Classification)
- การจำแนกข้อมูลภาพจะใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

ข้อแตกต่างของงานวิจัยในวิทยานิพนธ์นี้กับงานวิจัยอื่นที่เกี่ยวข้องกับการจำแนกมะเร็งเต้านมจากภาพแมมโมแกรม คือ เทคนิคการดึงลักษณะสำคัญของภาพ ในวิทยานิพนธ์นี้จะใช้พื้นความรู้ (Background Knowledge) ที่เป็นข้อสรุปจากประสบการณ์ของนักรังสีวิทยา เกี่ยวกับข้อแตกต่างระหว่างภาพก้อนเนื้อที่ไม่อันตรายและก้อนเนื้อร้าย เป็นพื้นฐานสำคัญในการพัฒนาเทคนิคการดึงลักษณะสำคัญที่จะช่วยให้การจำแนกภาพมะเร็งเต้านมมีความแม่นมากขึ้น

นอกจากวิธีที่กล่าวข้างต้นแล้ว งานวิจัยนี้ยังได้มีการทดสอบประสิทธิภาพของอัลกอริทึมที่พัฒนาขึ้นด้วยการเปรียบเทียบประสิทธิภาพในการจำแนกมะเร็งเต้านมในภาพแมมโมแกรม กับอัลกอริทึมการเรียนรู้ของเครื่องแบบอื่น ๆ อีก 2 อัลกอริทึม คือ โครงข่ายประสาทเทียม (Artificial Neural Networks) และ นาอีฟเบย์ (Naïve Bays) โดยเกณฑ์ที่ใช้ในการทดสอบประสิทธิภาพคือความแม่น (Accuracy) ในการจำแนกข้อมูล

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาและพัฒนาขั้นตอนวิธีการจำแนกมะเร็งเต้านม จากภาพแมมโมแกรมโดยเน้นในส่วนของเทคนิคการปรับปรุงคุณภาพของภาพ การพิจารณาหาขอบเขตของภาพ และการพิจารณาหาลักษณะสำคัญของภาพ
2. เพื่อศึกษาและพัฒนาขั้นตอนวิธีการปรับปรุงภาพก่อนการจำแนก
3. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอกับอัลกอริทึมการเรียนรู้ของเครื่องแบบอื่น ๆ

1.3 ขอบเขตของการวิจัย

1. การทดสอบใช้ข้อมูลภาพแมมโมแกรมจาก Digital Database for Screening Mammography หรือ DDSM จากเว็บไซต์ของ University of South Florida ประเทศสหรัฐอเมริกา (http://marathon.csee.usf.edu/Mammography/Database.html)
2. การเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการจำแนกจะทำการเปรียบเทียบกับอัลกอริทึมการเรียนรู้ของเครื่อง 2 แบบ คือ นาอีฟเบย์ และ โครงข่ายประสาทเทียม
3. การเปรียบเทียบประสิทธิภาพจะใช้เกณฑ์ร้อยละของความแม่นในการจำแนก ค่า Sensitivity ค่า Specificity ค่า F-measure และ พื้นที่ใต้กราฟ ROC

1.4 ประโยชน์ที่จะได้รับ

จากการศึกษาและพัฒนางานวิจัยนี้ ผู้วิจัยคาดว่าอัลกอริทึมที่พัฒนาขึ้นจะเกิดประโยชน์ต่อผู้ใช้ในการนำไปจำแนกมะเร็งเต้านมในภาพแมมโมแกรมในประเด็นต่าง ๆ ดังนี้

1. การปรับปรุงภาพและการดึงลักษณะสำคัญจากภาพทำให้ได้ ลักษณะสำคัญที่มีประโยชน์ช่วยให้การจำแนกมีประสิทธิภาพมากยิ่งขึ้น
2. ค่าความแม่นในการจำแนกข้อมูลภาพสูงขึ้นเมื่อเปรียบเทียบกับอัลกอริทึมอื่น ๆ
3. ทำให้การวิเคราะห์มะเร็งเต้านมจากภาพแมมโมแกรมมีความแม่นมากยิ่งขึ้นส่งผลให้การรอดชีวิตของผู้ป่วยสูงขึ้น

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

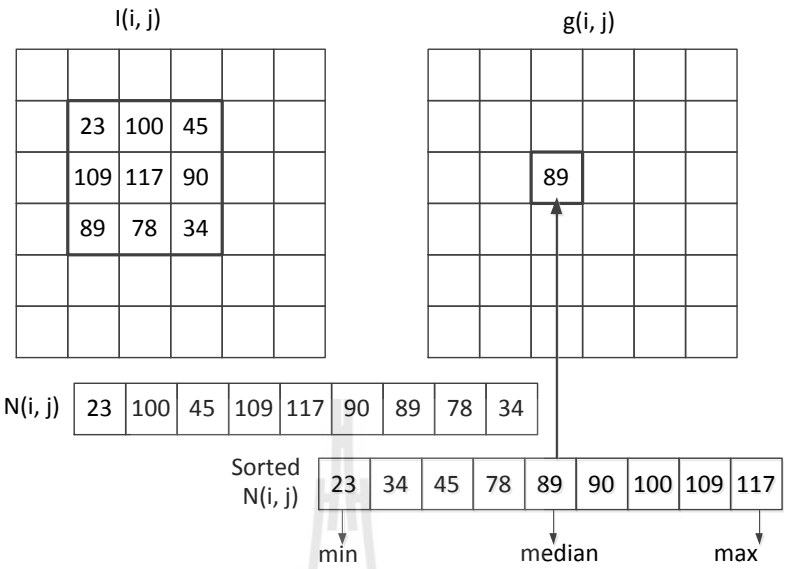
เนื้อหาในบทนี้ประกอบด้วยการทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้อง ซึ่งประกอบด้วยรายละเอียดทฤษฎีการปรับปรุงภาพ (Image Enhancement) การกำหนดขอบเขตภาพ (Region of Interest) ด้วยวิธีการขยายพื้นที่ การหาลักษณะสำคัญจากภาพ (Image Feature Extraction) การจำแนกประเภทข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน (Data Classification with Support Vector Machine) และงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีการปรับปรุงภาพ

2.1.1 ตัวกรองมัธยฐาน (Median Filter)

ตัวกรองมัธยฐาน หรือมีเดียนฟิลเตอร์ เป็นเทคนิคการกรองสัญญาณภาพแบบไม่เป็นเชิงเส้น หรือนอนลิเนียร์ ซึ่งนิยมใช้กันมากในการกำจัดสัญญาณรบกวนภายในภาพ เช่น ใช้ในกระบวนการก่อนประมวลผลภาพ (Preprocessing) เพื่อปรับภาพในเรียบ และมีความชัดเจนมากยิ่งขึ้นก่อนนำภาพไปเข้าสู่กระบวนการประมวลผล ซึ่งตัวกรองมัธยฐานนั้นสามารถกำจัดสัญญาณรบกวนที่มีค่าความเข้มสีที่แตกต่างจากพิกเซลข้างเคียงภายในภาพได้ โดยยังคงรักษาความคมชัดของภาพและขอบของวัตถุ (Edge Contour) ในภาพไว้ได้

หลักการสำคัญของตัวกรองมัธยฐาน คือ ใช้หน้าต่างขนาดเล็ก (Window) ทำการประมวลผลในภาพซึ่งขนาดหน้าต่างที่นิยมใช้กันนั้นคือขนาด 3×3 พิกเซล โดยหน้าต่างขนาดเล็กนี้จะเลื่อนเพื่อทำการประมวลผลไปในทุก ๆ พิกเซลในภาพจากซ้ายไปขวาโดยทำการเลือกตำแหน่งพิกเซลปัจจุบัน (i, j) จากนั้นจึงดูค่าความเข้มสีของพิกเซลรอบข้างอีก 8 พิกเซล แล้วนำค่าความเข้มสีของทั้ง 9 พิกเซล (รวมพิกเซลปัจจุบัน) มาทำการเรียงลำดับจากน้อยไปมาก แล้วคัดเลือกค่ากลางหรือค่ามัธยฐาน (Median) เพื่อนำไปแทนที่ในตำแหน่งพิกเซลปัจจุบัน (i, j) รูปที่ 2.1 แสดงตัวอย่างการปรับปรุงภาพด้วยการใช้ตัวกรองมัธยฐาน ค่าในพิกเซลปัจจุบันคือ 117 เมื่อได้รับการปรับปรุงแล้วจะถูกเปลี่ยนเป็นค่า 89 ซึ่งเป็นค่ามัธยฐานของข้อมูล 9 พิกเซล



รูปที่ 2.1 ตัวอย่างการทำงานของตัวกรองมัธยฐาน

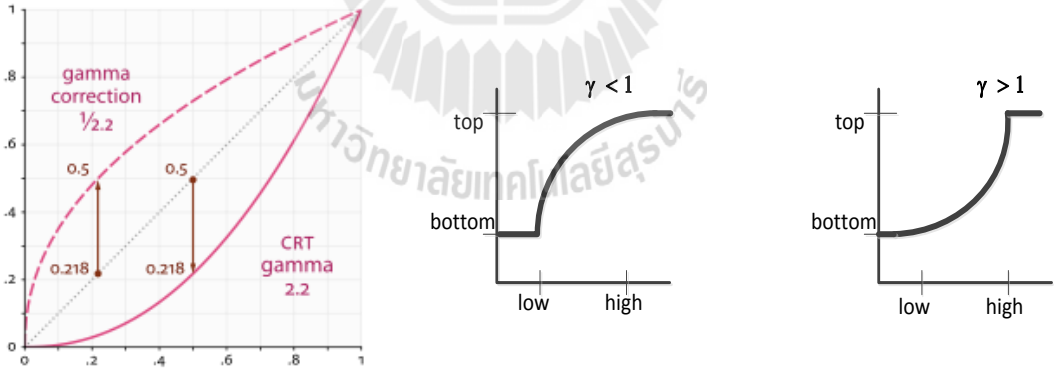
2.1.2 การแก้ไขแกมมา (Gamma Correction)

การปรับปรุงภาพแมมโมแกรม หรือ Image Enhancement ก่อนนำมาจำแนกว่าเป็นก้อนเนื้อไม่อันตรายหรือก้อนเนื้อร้ายนั้นมีความจำเป็นอย่างมาก ที่จะช่วยให้ผลของการจำแนกมีความแม่นและถูกต้องมากยิ่งขึ้น วิธีปรับปรุงภาพแมมโมแกรมที่นิยมนำมาใช้ในปัจจุบันและเป็นวิธีที่ไม่ซับซ้อน เช่น การใช้เทคนิค Grey-level Thresholding ซึ่งเป็นการปรับความเข้มของระดับสีเทาของบริเวณวัตถุที่สนใจภายในภาพให้มีความชัดเจนมากยิ่งขึ้น ในทำนองเดียวกันก็ปรับลดความเข้มสีในส่วนของไขมันที่เป็นพื้นหลังให้มีความเข้มสีที่ต่ำลง เทคนิค Grey-level Thresholding ที่นำมาใช้ในงานวิจัยครั้งนี้คือ วิธีปรับปรุงภาพด้วยเทคนิคการแก้ไขแกมมาหรือแกมมาคอเร็คชัน ซึ่งเป็นวิธีแบบนอนลิเนียร์ (Nonlinear) ทำหน้าที่ในการแปลงความสว่างของภาพให้มีความเหมาะสม โดยปรับเปลี่ยนพารามิเตอร์ที่ชื่อว่า แกมมา (Gamma) โดยทำการพิจารณาความเข้มของสีในทุก ๆ จุดพิกเซลในภาพ โดยภาพแมมโมแกรมที่ใช้ในงานวิจัยนี้จะเป็นภาพระดับสีเทาซึ่งมีค่าอยู่ระหว่าง 0 ถึง 255 โดยที่ 0 หมายถึงจุดพิกเซลที่มีความมืดสูงสุด และ 255 หมายถึงจุดพิกเซลที่มีความสว่างสูงสุด จึงทำให้บริเวณภาพที่เป็นก้อนเนื้อนั้นจะมีความสว่างมากกว่าบริเวณภาพที่เป็นไขมันในเต้านม ซึ่งอุปกรณ์ที่รับภาพแมมโมแกรมมานั้นอาจจะแสดงความแตกต่างของสีระหว่างบริเวณก้อนเนื้อและไขมันไม่ชัดเจนนัก ดังนั้นการปรับปรุงภาพแมมโมแกรมด้วยวิธีการแก้ไขแกมมา จะปรับความเข้มสีให้บริเวณที่สว่างซึ่งสันนิษฐานว่าเป็นบริเวณก้อนเนื้อให้มีความสว่างมากยิ่งขึ้น

และในทางกลับกันการแก้ไขแกมมา ก็จะทำการปรับความเข้มสีบริเวณที่เป็นไขมันซึ่งมีความสว่างน้อยอยู่แล้วให้ยิ่งมืดลง ทำให้เราได้ภาพความแตกต่างระหว่างบริเวณก้อนเนื้อและบริเวณไขมันชัดเจนมากยิ่งขึ้น การคำนวณค่าแก้ไขแกมมา ทำได้ดังสมการที่ (2-1)

$$\text{Corrected} = 255 * (\frac{\text{Image}}{255})^{(\frac{1}{\gamma})} \quad (2\text{-}1)$$

เมื่อ γ คือค่าที่แปลงความเข้มของสีในภาพ โดยถ้า γ<1 จะเรียกว่า Encoding Gamma หรือ Gamma Compression ในทางกลับกันหากค่า γ>1 จะเรียกว่า Decoding Gamma หรือ Gamma Expansion โดยผลของการแก้ไขแกมมา เป็นได้ 2 ลักษณะ คือ ถ้า γ>1 จะทำให้ส่วนที่เป็นเงาหรือส่วนมืดภายในภาพยิ่งมีความเข้มสีที่น้อยลงหรือมืดมากขึ้นไปอีก และหาก γ<1 จะทำให้ส่วนที่เป็นเงาหรือส่วนที่มืดมีความความเข้มสีที่มากขึ้นหรือทำให้บริเวณที่มืดสว่างขึ้นนั่นเอง รูปที่ 2.2 (ก) แสดงความสัมพันธ์ระหว่างการแก้ไขแกมมา จากอุปกรณ์แสดงภาพหรือหน้าจอ CRT (Cathode Ray Tube) และรูปที่ 2.2 (ข) แสดงกราฟความสัมพันธ์ระหว่างภาพก่อนการแก้ไขแกมมากับภาพหลังการแก้ไขแกมมา โดยแกน X แสดงความเข้มสีของภาพก่อนการแก้ไขแกมมา และ แกน Y แสดงความเข้มสีของภาพหลังการปรับแกมมา



(ก) (ข)

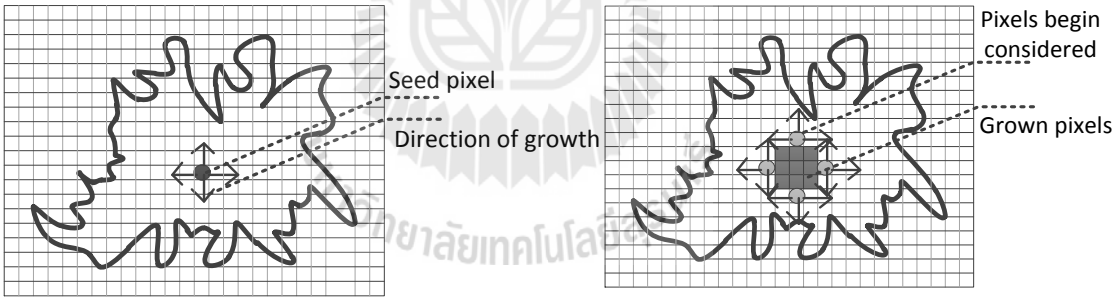
รูปที่ 2.2 ภาพ (ก) จำลองวิธีแก้ไขแกมมา และ (ข) แสดงผลของความเข้มสีเมื่อปรับค่า γ

(https://en.wikipedia.org/wiki/Gamma_correction)

2.2 การกำหนดขอบเขตของภาพด้วยวิธีการขยายพื้นที่

การขยายพื้นที่ของส่วนภาพ (Region Growing) เป็นการแบ่งส่วนภาพบนพื้นฐานของพื้นที่ (Region Based) ที่มีการกำหนดจุดกึ่งกลาง (Seed Point) ในภาพ แล้วทำการขยายพื้นที่ (Growing) ออกไปยังจุดพิกเซลใกล้เคียง โดยพิจารณาจากค่าระดับสีเทา จนกระทั่งขอบเขตนั้นสิ้นสุดเมื่อมีความเข้มสีของพิกเซลปัจจุบันและพิกเซลใกล้เคียงที่ต่างกันมาก ภายหลังเสร็จสิ้นขั้นตอนการขยายพื้นที่ ก็จะได้จำนวนขอบเขตในภาพหลาย ๆ ขอบเขต ซึ่งแบ่งแยกจากกันอย่างชัดเจน ช่วยให้ขั้นตอนต่อไปสามารถรวมส่วนภาพที่อยู่ติดกันและมีค่าระดับสีเทาใกล้เคียงกันเข้าไว้ด้วยกัน (Merging) รูปที่ 2.3 แสดงตัวอย่างของการขยายพื้นที่ของส่วนภาพโดยการพิจารณาพิกเซลใกล้เคียง

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



รูปที่ 2.3 การขยายส่วนพื้นที่ของภาพ

การคำนวณเพื่อการขยายพื้นที่ของส่วนภาพนั้น แสดงดังสมการที่ (2-2) ถึง (2-4) โดยที่ R แทนพื้นที่ที่สนใจ S แทนพิกเซลทั้งหมดที่พิจารณา และ P แทน Logical Predicate หมายถึงการพิจารณาพิกเซลใกล้เคียง ยกตัวอย่างเช่น หากค่าพิกเซลปัจจุบันมีค่าความเข้มสีคล้ายหรืออยู่ในช่วงขีดแบ่ง (Threshold) ที่กำหนด นั่นคือ $P(R_i) = TRUE$ แต่ทั้งนี้ค่า Logical Predicate ก็อาจขึ้นอยู่กับการพิจารณาด้วยวิธีการหรือค่าอื่น ๆ ได้ด้วย

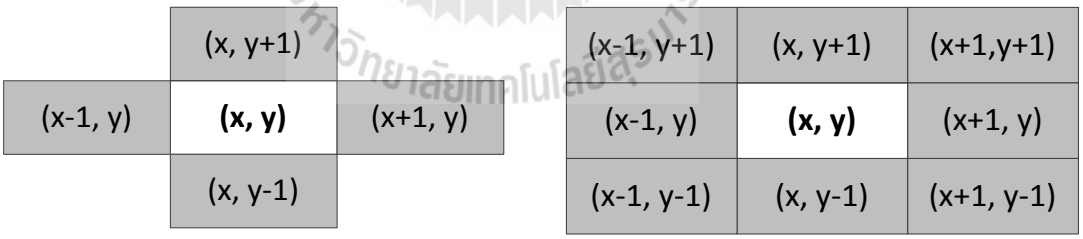
$$R = \bigcup_{i=1}^{S} R_i \ , \qquad R_i \cap R_j = 0 \ , \qquad i \neq j \tag{2-2}$$

$$P(R_i) = \text{TRUE} \ , \qquad i = 1, 2, \dots, S \tag{2-3}$$

$$P(R_i \cup R_j) = \text{FALSE} \ , \qquad i \neq j, \quad R_i \ \text{adjacent to} \ R_j \tag{2-4}$$

จากสมการที่ (2-2) หมายถึงการแบ่งขอบเขตต้องสมบูรณ์ (Completeness) โดยที่ทุก ๆ พิกเซลจะต้องอยู่ภายในขอบเขตใด ๆ และแต่ละขอบเขตจะต้องไม่ซ้อนทับกัน (Disjointness) สมการที่ (2-3) หมายถึง คุณสมบัติหรือความเข้มสีในระดับสีเทาของพิกเซลใด ๆ ที่อยู่ในขอบเขตเดียวกันจะต้องมีคุณสมบัติที่คล้ายกัน (Satisfiability) เช่น ความเข้มสีใกล้เคียงกัน และสมการที่ (2-4) หมายถึง ขอบเขตใด ๆ ที่ทำการแบ่งแยกแล้วจะแบ่งแยกจากกันอย่างชัดเจน หรือ แต่ละขอบเขตภายหลังการขยายพื้นที่ของส่วนภาพจะต้องแบ่งแยกกันอย่างสิ้นเชิง (Segmentability) ซึ่งในบางกรณี เช่น การแบ่งขอบเขตภาพก้อนเนื้อในเต้านม บางขอบเขตที่อยู่ติดกันหรือมีความห่างเพียงเล็กน้อยก็สามารถนำขอบเขตเหล่านี้มาผสานกันได้ เพื่อความสะดวกในการนำไปดึงลักษณะสำคัญต่อไป

การขยายส่วนพื้นที่ในภาพจะใช้การพิจารณาค่าของพิกเซลปัจจุบันร่วมกับค่าของพิกเซลใกล้เคียง (Neighborhood) โดยการพิจารณาพิกเซลใกล้เคียงสามารถพิจารณาได้สองแบบ คือ การพิจารณาพิกเซลใกล้เคียงจำนวน 4 พิกเซล (4-Neighberhoods) ดังแสดงในรูป2.4 (ก) และการพิจารณาพิกเซลใกล้เคียงจำนวน 8 พิกเซล (8-Neighberhoods) ดังแสดงในรูป 2.4 (ข)

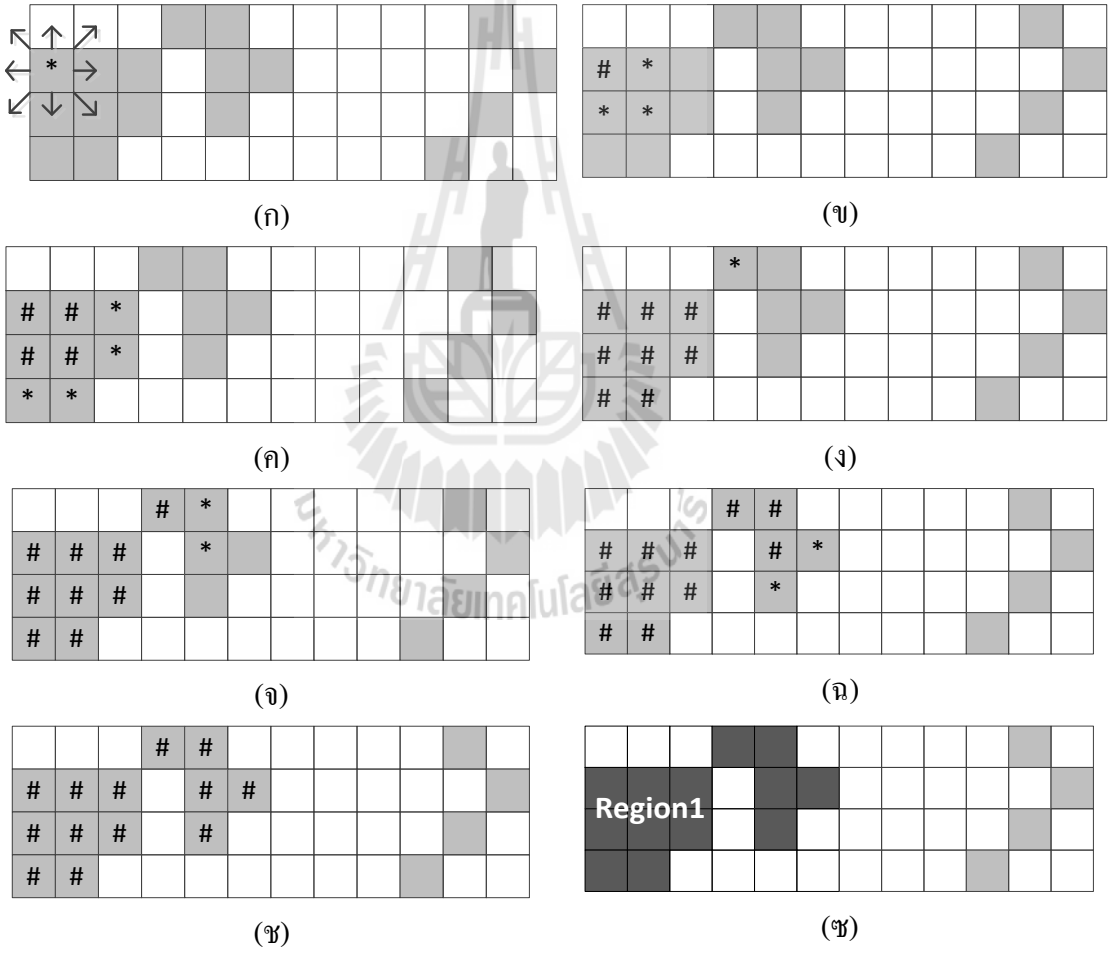


(ก) (ข)

รูปที่ 2.4 ตำแหน่งการพิจารณาพิกเซลใกล้เคียง (ก) 4 พิกเซล (ข) 8 พิกเซล

สำหรับตัวอย่างการขยายส่วนพื้นที่ แสดงในรูปที่ 2.5 โดยทำการพิจารณาพิกเซลใกล้เคียงจำนวน 8 พิกเซล เริ่มต้นที่รูป 2.5 (ก) กำหนดจุดเริ่มต้นในการขยายส่วนพื้นที่โดยแทนด้วยเครื่องหมายดอกจัน (Asterisk) จากนั้นทำการพิจารณาพิกเซลข้างเคียงจำนวน 8 พิกเซลโดยเปรียบเทียบค่าพิกเซลปัจจุบันกับพิกเซลใกล้เคียงหากมีค่าความเข้มสีใกล้เคียงกัน หรือมีค่าความ

เข้มสีอยู่ในช่วงขีดแบ่ง หรือ Threshold ที่กำหนด ก็ให้ทำเครื่องหมายชาร์ป (Sharp) ในพิกเซลใกล้เคียงเหล่านั้น ดังแสดงในรูปที่ 2.5 (ข) ถึง 2.5 (ช) แต่หากพิกเซลใกล้เคียงใดที่มีความเข้มสีต่างกัน หรือมีค่าความเข้มสีไม่อยู่ในช่วงขีดเบ่งก็ไม่ต้องทำเครื่องหมายและไม่ต้องนำมาพิจารณา จากนั้นจึงทำกระบวนการเดิมซ้ำไปเรื่อย ๆ จนกระทั่งไม่มีจุดพิกเซลใกล้เคียงที่มีความเข้มสีคล้ายกันจึงหยุดการขยายพื้นที่ สุดท้ายจะได้บริเวณขอบเขตที่สนใจโดยใช้วิธีการขยายพื้นที่ดังแสดงในรูป 2.5 (ซ) แต่หากยังมีขอบเขต อื่น ๆ ในภาพที่ยังไม่ได้ทำการขยายพื้นที่ก็ทำตามกระบวนการเดิมคือกำหนดจุดเริ่มต้นของขอบเขตอื่น ๆ และพิจารณาพิกเซลใกล้เคียงต่อไปจนครบทั้งภาพ



รูปที่ 2.5 ตัวอย่างการขยายส่วนพื้นที่โดยพิจารณาพิกเซลใกล้เคียง 8 พิกเซล

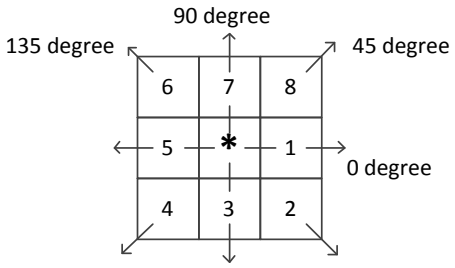
2.3 การหาลักษณะสำคัญของภาพ

2.3.1 ลักษณะสำคัญของลวดลาย (Texture Feature)

ลวดลายภายในภาพเป็นหนึ่งในลักษณะสำคัญที่ใช้ในการระบุวัตถุ หรือขอบเขตที่น่าสนใจในภาพ ไม่ว่าจะเป็นภาพถ่ายทางอากาศ (Aerial Photograph) ภาพถ่ายจากดาวเทียม (Satellite Images) หรือภาพถ่ายทางการแพทย์ (Medical Images) ซึ่งลวดลายในภาพเหล่านี้สามารถอธิบายถึงคุณสมบัติของภาพ และสามารถตีความหรือบอกความแตกต่างของภาพได้ คุณสมบัติของลวดลายนั้นยังสามารถนำไปใช้ในการจำแนกภาพทางการแพทย์เช่น ภาพแมมโมแกรม และ ภาพอัลตราซาวด์ ลักษณะสำคัญหรือคุณลักษณะของลวดลายนั้นจะเป็นข้อมูลที่เกี่ยวกับการกระจายของรูปแบบโทนสี (Tone) ภายในภาพ ดังนั้นลักษณะการเปลี่ยนแปลงของโทนสีภายในภาพจึงเป็นข้อมูลสำคัญซึ่งนำมาใช้ในการจำแนกภาพได้

ลักษณะสำคัญของลวดลายนั้นสามารถหาได้จากเมตริกซ์ของระดับสีเทาที่เกิดขึ้นร่วมกัน (Grey-level Co-occurrence Matrix) หรือ GLCM ฟังก์ชันใน GLCM นั้นจะทำการคำนวณและเปรียบเทียบการเกิดขึ้นของระดับสีเทาในภาพหรือรูปแบบ (Pattern) ของระดับสีเทาระหว่างพิกเซลในภาพ โดยใช้ความน่าจะเป็นในการแสดงผลของความชัดเจนของลวดลาย (Contrast) การเกิดขึ้นร่วมกันของลวดลาย (Correlation) และการเป็นเนื้อเดียวกันของลวดลาย (Homogeneity) ซึ่งลักษณะสำคัญของลวดลายเหล่านี้จะถูกนำไปใช้ในกระบวนการจำแนกภาพ

การสร้าง GLCM นั้นมีองค์ประกอบคือ P(i, j, d, θ) ซึ่งเป็นการหาความน่าจะเป็นที่เกิดขึ้นร่วมกันของระดับสีเทา (Joint Probability) ที่ตำแหน่ง i และ j โดยมีการกำหนดระยะห่างระหว่างแต่ละพิกเซลนั่นคือค่า d ร่วมด้วยการกำหนดทิศทางที่เป็นไปได้ในการดูการเกิดขึ้นร่วมกันระหว่างแต่ละพิกเซลนั่นคือค่า θ ซึ่งโดยทั่วไปนั้นทิศทางที่ใช้จะมีค่า $0^{o}, 90^{o}, 45^{o}$ และ 135^{o} (Horizontal, Vertical, Right Diagonal and Left Diagonal) ดังแสดงในรูปที่ 2.6 และสำหรับค่า d นั้นส่วนใหญ่จะใช้เป็นการวัดระยะห่างระหว่างจุดแบบแมนฮัตทัน (Manhattan Distance) สำหรับการหาค่า P(i, j, d, θ) ของ GLCM ในทิศทางทั้ง 4 ทิศทางนั้นสามารถแสดงดังสมการที่ (2-5) ถึง (2-8)



รูปที่ 2.6 ทิศทางการนับความสัมพันธ์ 4 ทิศทางในการสร้างเมตริกซ์ GLCM

$$P(i,j,d,0^o) = \left[\{((x_1,y_1),(x_2,y_2)) \text{ โดยที่ } |x_2 - x_1| = d, y_2 - y_1 = 0\}\right] \quad (2\text{-}5)$$

$$P(i,j,d,45^o) = [\{((x_1,y_1),(x_2,y_2)) \text{ โดยที่ } (x_2 - x_1 = d, y_2 - y_1 = d) \text{ หรือ } (x_2 - x_1 = -d, y_2 - y_1 = -d)\}] \quad (2\text{-}6)$$

$$P(i,j,d,90^o) = \left[\{((x_1,y_1),(x_2,y_2)) \text{ โดยที่ } x_2 - x_1 = 0, |y_2 - y_1| = d\}\right] \quad (2\text{-}7)$$

$$P(i,j,d,135^o) = [\{((x_1,y_1),(x_2,y_2)) \text{ โดยที่ } (x_2 - x_1 = d, y_2 - y_1 = -d) \text{ หรือ } (x_2 - x_1 = -d, y_2 - y_1 = -d)\}] \quad (2\text{-}8)$$

รูปที่ 2.7 แสดงตัวอย่างการหาลักษณะสำคัญของลวดลาย จากรูปที่ 2.7 (ก) เป็นรูปโทนสีเทาขนาด 4 × 4 พิกเซล ซึ่งมีค่าโทนสีคือ 0, 1, 2 และ 3 รูปที่ 2.7 (ข) แสดงรูปแบบของเมตริกซ์ของการนับการเกิดขึ้นร่วมกันของโทนสีเทาในภาพ ยกตัวอย่างเช่น จากรูปที่ 2.7 (ข) ตำแหน่งที่ (0, 0) หากทำการนับโทนสีเทา (เครื่องหมาย # คือการนับจำนวน) ที่เกิดขึ้นร่วมกันในแนวนอน (Horizontal) ของโทนสี (0,0) และกำหนดค่า d=1 ในรูปที่ 2.7 (ก) ผลลัพธ์ในการนับโทนสีเทาที่เกิดขึ้นร่วมกันในตำแหน่ง (0, 0) แสดงดังรูปที่ 2.7 (ค) ซึ่งค่าที่นับได้ คือ 4 (นับทิศทางแนวนอนทั้งซ้ายและขวา) ในขณะเดียวกัน รูปที่ 2.7 (ง) 2.7 (จ) และ 2.7 (ฉ) ก็คือการนับในทิศทางแนวตั้ง แนวทแยงมุมด้านซ้าย และแนวทแยงมุมด้านขวา ตามลำดับ

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

(ก)

Grey Tone	0	1	2	3
0	#(0, 0)	#(0, 1)	#(0, 2)	#(0, 3)
1	#(1, 0)	#(1, 1)	#(1, 2)	#(1, 3)
2	#(2, 0)	#(2, 1)	#(2, 2)	#(2, 3)
3	#(3, 0)	#(3, 1)	#(3, 2)	#(3, 3)

(ข)

$$0^{o}\quad P_H = \begin{pmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \qquad 90^{o}\quad P_V = \begin{pmatrix} 6 & 0 & 2 & 0 \\ 0 & 4 & 2 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 2 & 2 \end{pmatrix}$$

(ค) (ง)

$$135^{o}\quad P_{LD} = \begin{pmatrix} 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix} \qquad 45^{o}\quad P_{RD} = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 4 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

(จ) (ฉ)

รูปที่ 2.7 ตัวอย่างการสร้างเมตริกซ์ GLCM

หลังจากได้เมตริกซ์ที่เก็บจำนวนนับของการเกิดขึ้นร่วมกันของโทนสีเทาแล้ว ให้ทำการหาความน่าจะเป็นที่เกิดขึ้นร่วมกันของระดับสีเทาในทุก ๆ ทิศทาง โดยนำผลรวมของค่าทั้งหมดในเมตริกซ์มาหารกับค่าทุกตำแหน่ง (i, j) ในเมตริกซ์ที่เก็บจำนวนนับ รูปที่ 2.8 แสดงตัวอย่างการหาความน่าจะเป็นที่เกิดขึ้นร่วมกันของระดับสีเทาในทิศทาง 0º หลังจากนั้นก็จะสามารถคำนวณหาลักษณะสำคัญของลวดลายได้ด้วยการหา ความเป็นเนื้อเดียวกันของลวดลาย ความชัดเจนของลวดลาย และ การเกิดขึ้นร่วมกันของลวดลายจากสมการที่ (2-9) (2-10) และ (2-11) ตามลำดับ

$$P_H^{0^{o}} = 1/24 * \begin{pmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 0.167 & 0.083 & 0.041 & 0 \\ 0.083 & 0.167 & 0 & 0 \\ 0.041 & 0 & 0.250 & 0.041 \\ 0 & 0 & 0.041 & 0.083 \end{pmatrix}$$

รูปที่ 2.8 ตัวอย่างการหาความน่าจะเป็นที่เกิดขึ้นร่วมกันของระดับสีเทาในทิศทาง 0º

$$\text{Homogeneity} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{P(i,j)}{1+|i-j|}\right) \tag{2-9}$$

$$\text{Contrast} = \sum_{i=1}^{m}\sum_{j=1}^{n}(i-j)^2 P(i,j) \tag{2-10}$$

$$\text{Correlation} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\{i \times j\} \times P(i,j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \tag{2-11}$$

โดยที่ m และ n คือ ความกว้างและความสูงของเมตริกซ์ GLCM

P(i, j) คือ ความน่าจะเป็นที่ตำแหน่ง (i, j) ในเมตริกซ์ GLCM

μ_x และ μ_y คือค่าเฉลี่ยของเมตริกซ์ GLCM ตามแถวและคอลัมน์

σ_x และ σ_y คือค่าเบี่ยงเบนมาตรฐานของเมตริกซ์ GLCM ตามแถวและคอลัมน์

2.3.2 ลักษณะสำคัญของฮิสโตแกรม (Histogram Based Feature)

รูปร่างลักษณะและคุณสมบัติของฮิสโตแกรม เป็นลักษะสำคัญอีกประเภทหนึ่งที่นิยมนำมาใช้เป็นลักษณะสำคัญในการจำแนกภาพ ซึ่งกราฟฮิสโตแกรมจะให้ข้อมูลสถิติของความเข้มสีที่เกิดขึ้นในภาพ และสามารถหาความน่าจะเป็นของความเข้มสีระดับสีเทาที่เกิดขึ้นในภาพ ดังสมการที่ (2-12)

$$P(i) = \frac{h(i)}{NM} \tag{2-12}$$

โดยที่ i คือ 0, 1, 2, ……., G-1

h คือ จำนวนพิกเซลที่เกิดขึ้นในแต่ละความเข้มสีระดับสีเทาเมื่อ i คือความเข้มสีระดับสีเทามีค่าตั้งแต่ 0 ถึง 255

G คือ ระดับโทนสีเทาในภาพ

N คือ จำนวนพิกเซลของภาพในแนวนอน (Width)

M คือ จำนวนพิกเซลของภาพในแนวตั้ง (Height)

กราฟฮิสโตแกรมนั้นมีคุณลักษณะสำคัญทางสถิติ (Statistic Feature) ทั้งหมด 4 ค่า ได้แก่

1. ค่าเฉลี่ย (Mean) คือ ค่าเฉลี่ยความเข้มสี ซึ่งคำนวณได้จากสมการที่ (2-13)

$$\mu = \sum_{i=1}^{G-1} iP(i) \tag{2-13}$$

2. ค่าความแปรปรวน (Variance) คือ การเปลี่ยนแปลงความเข้มสีรอบค่าเฉลี่ย μ ซึ่งคำนวณได้จากสมการที่ (2-14)

$$\sigma^2 = \sum_{i=1}^{G-1} (i - \mu)^2 P(i) \tag{2-14}$$

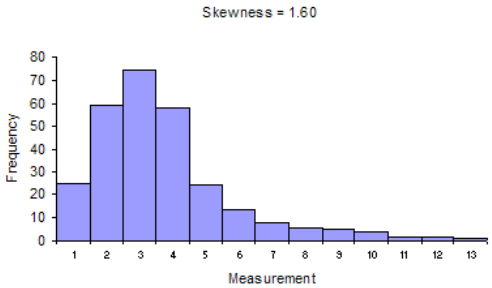
3. ความเบ้ (Skewness) คือ ค่าที่แสดงถึงความสมมาตรของกราฟฮิสโตแกรม หากฮิสโตแกรมมีความสมมาตรแล้ว ความเบ้จะมีค่าเป็น 0 ซึ่งสามารถคำนวณได้จากสมการที่ (2-15)

$$s = \sigma^{-3} \sum_{i=1}^{G-1} (i - \mu)^3 P(i) \tag{2-15}$$

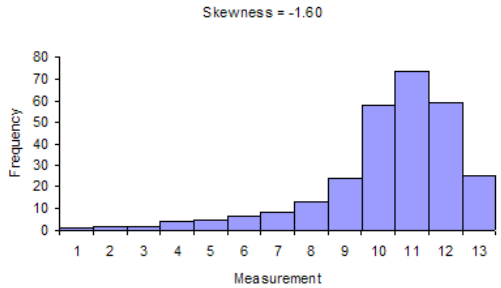
4. ความโด่ง (Kurtosis) คือ ค่าที่วัดจุดสูงสุดและต่ำสุดภายในกราฟในฮิสโตแกรม ซึ่งมีความสัมพันธ์กับการกระจายข้อมูลแบบปกติ (Normal Distribution) คำนวณได้จากสมการที่ (2-16)

$$k = \sigma^{-4} \sum_{i=1}^{G-1} (i - \mu)^4 P(i) \tag{2-16}$$

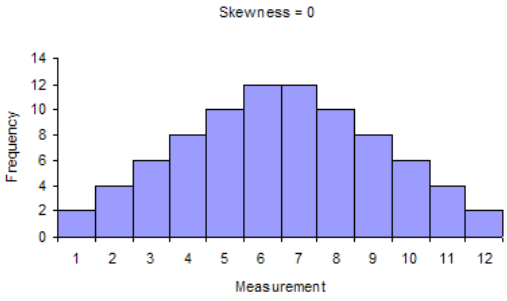
จากรูปที่ 2.9 แสดงลักษณะความเบ้ของกราฟฮิสโตแกรมในลักษณะที่แตกต่างกันไป โดยรูปที่ 2.9(ก) แสดงกราฟฮิสโตแกรมเบ้ทางขวา โดยมีค่าความเบ้เป็นค่าบวก รูปที่ 2.9(ข) แสดงกราฟฮิสโตแกรมเบ้ทางซ้าย โดยมีค่าความเบ้เป็นค่าลบ และรูปที่ 2.9(ค) แสดงกราฟฮิสโตแกรมที่สมมาตร โดยมีค่าความเบ้เข้าใกล้ศูนย์หรือเท่ากับศูนย์



(ก)



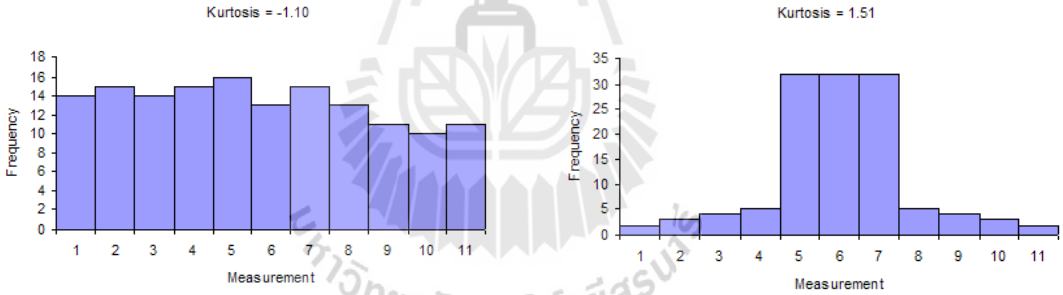
(ข)



(ค)

รูปที่ 2.9 ลักษณะความเบ้ของกราฟฮิสโตแกรม (ก) เบ้ทางขวา (ข) เบ้ทางซ้าย (ค) สมมาตร (https://www.spcforexcel.com/knowledge/basic-statistics)

จากรูปที่ 2.10 แสดงลักษณะความโด่งของกราฟฮิสโตแกรมในลักษณะที่แตกต่างกันไป โดยรูปที่ 2.10(ก) แสดงกราฟฮิสโตแกรมที่มีค่าความโด่งน้อย โดยมีค่าความโด่งเป็นค่าลบ ส่วนรูปที่ 2.10(ข) แสดงกราฟฮิสโตแกรมที่มีค่าความโด่งมาก โดยมีค่าความโด่งเป็นค่าบวก



(ก) (ข)

รูปที่ 2.10 ลักษณะความโด่งของกราฟฮิสโตแกรม (ก) ความโด่งน้อย (ข) ความโด่งมาก (https://www.spcforexcel.com/knowledge/basic-statistics)

2.3.3 ลักษณะสำคัญของรูปร่าง (Shape Feature)

การหาลักษณะสำคัญของรูปร่างนั้นเป็นองค์ประกอบที่สำคัญอย่างหนึ่งในการจำแนกภาพ สมมติว่าหากเราต้องการที่จะจำแนกวัตถุสองชนิดซึ่งมีรูปร่างต่างกันจากภาพ ลักษณะสำคัญของรูปร่างของวัตถุจะเป็นตัวระบุความแตกต่างของวัตถุแต่ละชนิด ลักษณะสำคัญของรูปร่างที่สำคัญได้แก่ พื้นที่ (Area) เส้นผ่าศูนย์กลาง (Diameter) ส่วนนูนของรูปร่าง (Convex Area) โครง

ร่าง (Skeleton) เส้นรอบรูป (Perimeter) ระยะทางจากจุดศูนย์กลางไปยังเส้นขอบ (Centroid to Distance)

ในการจำแนกมะเร็งเต้านมจากภาพแมมโมแกรมนั้น โดยทั่วไปแล้วลักษณะรูปร่างของก้อนเนื้อที่ไม่เป็นอันตรายและก้อนเนื้อร้ายจะมีรูปร่างที่ต่างกัน คือ ก้อนเนื้อที่ไม่เป็นอันตรายจะมีลักษณะเป็นรูปร่างค่อนข้างกลมและขอบของก้อนเนื้อจะมีรอยหยักน้อย ซึ่งในทางกลับกันก้อนเนื้อร้ายที่มีแนวโน้มจะเป็นมะเร็งนั้นลักษณะรูปร่างจะไม่เป็นรูปทรง มีความบิดเบี้ยวและโค้งงอมาก และขอบของก้อนเนื้อจะมีรอยหยักมาก ดังนั้นในงานวิจัยนี้จึงเลือกใช้ลักษณะสำคัญของรูปร่าง โดยทำการวัดระยะทางจากจุดศูนย์กลางของก้อนเนื้อไปยังเส้นขอบ ซึ่งวิธีการนี้จะเป็นการดูความเปลี่ยนแปลงของความโค้งจากจุดศูนย์กลางไปยังเส้นขอบ ซึ่งถ้ามีการเปลี่ยนแปลงความโค้งจากจุดศูนย์กลางไปยังเส้นขอบน้อยก็สันนิษฐานได้ว่าเป็นก้อนเนื้อไม่อันตราย และหากมีการเปลี่ยนแปลงความโค้งจากจุดศูนย์กลางไปยังเส้นขอบค่อนข้างมากก็สันนิษฐานได้ว่าเป็นก้อนเนื้อที่มีโอกาสเป็นมะเร็ง ซึ่งแสดงการหาระยะทางจากจุดศูนย์กลางไปยังเส้นขอบดังสมการที่ (2-17)

$$r(n) = [(x(n) - g_x)^2 + (y(n) - g_y)^2]^{1/2} \tag{2-17}$$

โดยที่ r(n) คือ ระยะทางจากจุดศูนย์กลางไปเส้นขอบใด ๆ ของวัตถุ

x(n) คือ พิกัด x ของจุดพิกเซลที่เป็นเส้นขอบ

y(n) คือ พิกัด y ของจุดพิกเซลที่เป็นเส้นขอบ

g_x คือ พิกัด x ของจุดกึ่งกลางวัตถุ

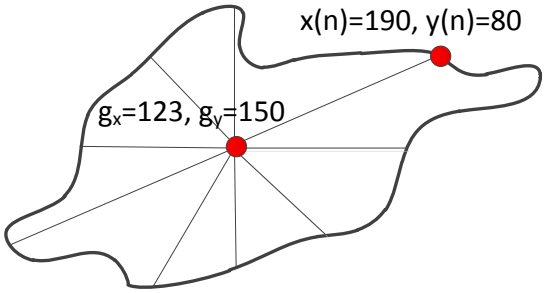
g_y คือ พิกัด y ของจุดกึ่งกลางวัตถุ

นอกจากการหาระยะทางจากจุดศูนย์กลางไปยังเส้นขอบแล้ว ค่ามุม (Angle) ระหว่างจุดสองจุดเมื่อวัดจากแกน X ก็เป็นอีกค่าหนึ่งที่สำคัญ โดยต้องมีการคำนวณหาเพื่อใช้ในการบ่งบอกลักษณะสำคัญของรูปร่าง สมการที่ (2-18) แสดงการหามุมระหว่างจุดสองจุด

$$\theta = \arctan\left(\frac{y(n) - g_y}{x(n) - g_x}\right) \times \frac{180}{\pi} \tag{2-18}$$

รูปที่ 2.11 แสดงตัวอย่างการคำนวณระยะทาง จากจุดศูนย์กลาง g_x = 123 และ g_y =150 ไปยังพิกัดเส้นขอบ x(n) = 190 และ y(n) = 80 สามารถคำนวณได้ดังนี้

$$r(n) = [(190 - 123)^2 + (80 - 150)^2]^{\frac{1}{2}} = 96.90$$



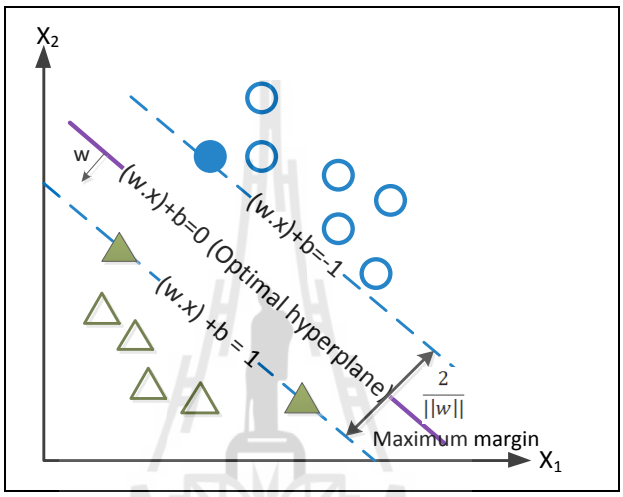
รูปที่ 2.11 การหาระยะทางจากจุดศูนย์กลางไปยังเส้นขอบ

2.4 การจำแนกประเภทข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน

ในปัจจุบันการจำแนกข้อมูล (Data Classification) เป็นงานวิเคราะห์ข้อมูลพื้นฐานที่สำคัญของการเรียนรู้ของเครื่อง (Machine learning) และสถิติ (Statistics) การจำแนกสามารถนำไปประยุกต์ใช้กับข้อมูลได้หลายประเภท อาทิเช่น การจำแนกข้อมูลจากฐานข้อมูล (Database) การจำแนกข้อมูลภาพ (Image Classification) เป็นต้น

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 2.12 เปรียบเสมือนการตัดถนนผ่านบ้านเรือนประชาชน ซึ่งจะต้องสร้างถนนให้กว้างมากที่สุดเท่าที่จะเป็นไปได้ และจะต้องทำลายบ้านเรือนประชาชนให้น้อยที่สุดด้วย ตัวแปรต่างๆ (Parameter) ของวิธีการหาไฮเปอร์เพลนนั้น ได้มาจากวิธีการหาคำตอบที่ดีที่สุดแบบควอดราติกส์ (Quadratic Programming Optimization Problem)

กำหนดให้ $D = \{(x_i, y_i)\}_{i=1}^{n}$ เป็นเซตของข้อมูลที่จะทำการจำแนก โดยมีจำนวนจุดซึ่งแทนข้อมูลแต่ละตัวเท่ากับ n จุด และ มีมิติข้อมูลเท่ากับ d มิติ (d-dimension) และสมมติว่าเราจะทำการจำแนกคลาส (Class) เพียงแค่สองคลาส คือ $y_i \in \{+1, -1\}$, โดยกำหนดเป็นคลาส +1 และคลาส -1 ดังนั้นขั้นตอนการทำงานหลักของซัพพอร์ตเวกเตอร์แมชชีน คือ การหาไฮเปอร์เพลน การคำนวณระยะทางจากจุดข้อมูลไปยังไฮเปอร์เพลน การกำหนดมาร์จินและซัพพอร์ตเวกเตอร์ของไฮเปอร์เพลน



รูปที่ 2.12 การจำแนกข้อมูล 2 คลาส โดยวิธีการซัพพอร์ตเวกเตอร์แมชชีน

2.4.1 ไฮเปอร์เพลน (Hyperplane)

การจำแนกเชิงเส้นด้วยไฮเปอร์เพลน h กับข้อมูล x ขนาด d มิติ นั้นสามารถแสดงได้ดังสมการที่ (2-19)

$$\begin{aligned} h(x) &= w^T x + b \\ &= w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b \end{aligned} \tag{2-19}$$

โดย w แทนค่า Weight Vector หรือเวกเตอร์ถ่วงน้ำหนัก ใน d-dimension และ b เป็นสเกลาร์ (Scalar) เรียกว่าค่าไบแอส (Bias) โดยในแต่ละจุดที่อยู่บนไฮเปอร์เพลนนั้น แสดงดังสมการที่ (2-20)

$$h(x) = w^T x + b = 0 \tag{2-20}$$

ไฮเปอร์เพลน $h(x)$ จะทำการแบ่งแยก d-dimension space ออกอย่างละครึ่ง ซึ่งเรียกว่าการแบ่งแยกเชิงเส้น (Linear Separable) โดย w และ b คือ ตัวแปรที่ใช้เบี่ยงเบนเส้นที่ใช้แยกคลาส ส่วน x นั้นเป็นอินพุทเวกเตอร์ ส่วน y จะเป็นตัวระบุว่าข้อมูลนั้นเป็นคลาส $+1$ หรือ -1 ซึ่งถ้าข้อมูลที่นำเข้ามาสามารถจำแนกแบบลิเนียร์ได้ เราก็จะสามารถหาไฮเปอร์เพลนที่เป็นลักษณะของเส้นตรงที่สามารถแบ่งแยกข้อมูลได้ สำหรับจุดใด ๆ ที่มีค่า $y_i = -1$ จะได้ค่า $h(x_i) < 0$ และสำหรับจุดใด ๆ ที่มีค่า $y_i = +1$ จะได้ค่า $h(x_i) > 0$ ซึ่งสมการที่ (2-21) แสดงให้เห็นถึงวิธีการทำนายว่าข้อมูลนั้นสังกัดอยู่คลาสใด

$$y = \begin{cases} +1 & \text{if } h(x) > 0 \\ -1 & \text{if } h(x) < 0 \end{cases} \tag{2-21}$$

กำหนดให้ a_1 และ a_2 เป็นจุดสองจุดใด ๆ ที่อยู่บนไฮเปอร์เพลน ซึ่งจากสมการที่ (2-20) จะได้

$$h(a_1) = w^T a_1 + b = 0 \tag{2-22}$$

$$h(a_2) = w^T a_2 + b = 0 \tag{2-23}$$

เมื่อนำสมการที่ (2-22) และ (2-23) มาลบกัน จะได้

$$w^T(a_1 - a_2) = 0 \tag{2-24}$$

ในที่นี้หมายถึงเวกเตอร์ถ่วงน้ำหนัก w คือเส้นที่ลากไปตั้งฉากกับไฮเปอร์เพลน เนื่องจากเป็นเส้นตั้งฉากที่ลากไปยังเวกเตอร์ใด ๆ $(a_1 - a_2)$ บนไฮเปอร์เพลน และเวกเตอร์ถ่วงน้ำหนัก w ยังเป็นตัวกำหนดทิศทางและปรับความเอียงหรือทิศทางของไฮเปอร์เพลนด้วย ส่วนค่าไบแอส b จะเป็นตัวกำหนดออฟเซ็ทหรือระยะห่างของไฮเปอร์เพลน

2.4.2 ระยะทางจากจุดไปยังไฮเปอร์เพลน (Distance from Point to Hyperplane)

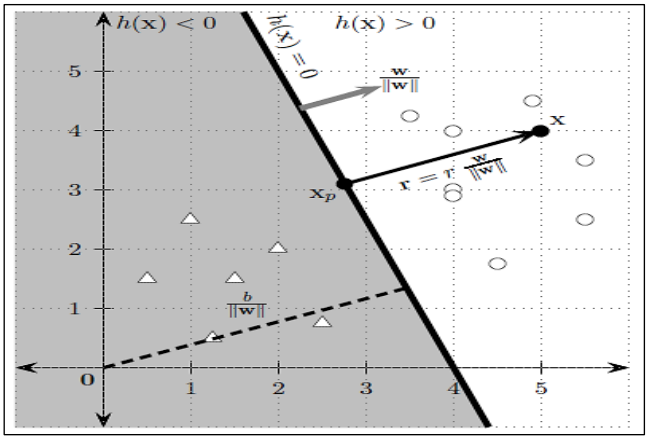
สำหรับทิศทางของจุดใด ๆ ไปยังไฮเปอร์เพลนนั้น กำหนดจุด $x \in R^d$ โดยจุด x แต่ละจุดนั้นไม่ได้อยู่บนไฮเปอร์เพลน กำหนดให้ x_p เป็นจุดบนไฮเปอร์เพลนที่ลากเส้นไปตั้งฉากกับจุด x ดังนั้น $r = x - x_p$ ดังแสดงในรูปที่ 2.13 และสามารถหาค่า x ได้ดังสมการที่ (2-25)

$$x = x_p + r \tag{2-25}$$
$$x = x_p + r\frac{w}{\|w\|}$$

เมื่อ r เป็นระยะทางจากจุด x ไปยัง x_p นั่นคือ r จะให้ค่าออฟเซ็ต (Offset) ของ จุด x ไปยัง x_p ซึ่งจะอยู่ในเทอมของยูนิตเวทเวกเตอร์ $\frac{w}{\|w\|}$ ซึ่งค่าออฟเซ็ต r มีเป็นค่าบวก ถ้า r อยู่ในทิศทางเดียวกันกับ w และ r จะมีค่าลบเมื่อ r อยู่ในทิศทางตรงกันข้ามกับ w หากแทนสมการที่ (2-25) ในสมการไฮเปอร์เพลนที่ (2-22) จะได้

$$\begin{aligned} h &= h\left(x_p + r\frac{w}{\|w\|}\right) + b \\ &= w^T\left(x_p + r\frac{w}{\|w\|} + b\right. \\ &= w^T x_p + b + r\frac{w^T w}{\|w\|} \\ &= h(x_p) + r\|w\| \\ &= r\|w\| \end{aligned} \tag{2-26}$$

จากรูปที่ 2.13 สำหรับจุดวงกลมที่อยู่ในคลาส +1 และจุดสามเหลี่ยมอยู่ในคลาส -1 ไฮเปอร์เพลน $h(x) = 0$ จะเป็นเส้นที่แบ่งสเปซของแต่ละคลาสออกจากกัน สำหรับจุด x พื้นที่ที่ระบายสีจะอยู่ในส่วนของ $h(x) < 0$ ในทางตรงกันข้าม จุด x ที่อยู่ในพื้นที่ที่ไม่ระบายสีจะอยู่ในส่วนของ $h(x) > 0$ สำหรับยูนิตเวทเวกเตอร์ $\frac{w}{\|w\|}$ (เส้นสีเทา) คือเส้นที่ลากไปตั้งฉากกับไฮเปอร์เพลน ส่วนทิศทางจากจุดออริจิน (Origin) ไปยังไฮเปอร์เพลน คือ $\frac{b}{\|w\|}$



รูปที่ 2.13 ไฮเปอร์เพลน 2 มิติ (Zaki, 2014)

ในขั้นตอนสุดท้ายนั้นจาก $h(x_p) = 0$ เนื่องจาก x_p เป็นจุดที่อยู่บนไฮเปอร์เพลน หากใช้ผลของสมการ (2-26) เราจะได้ระยะทางของจุดแต่ละจุดไปยังไฮเปอร์เพลนดังสมการที่ (2-27)

$$r = \frac{h(x)}{\|w\|} \tag{2-27}$$

ระยะทางที่จะได้นั้นจะต้องไม่เป็นค่าติดลบ ดังนั้นจึงนำเอาค่า r ไปคูณกับ y ของแต่ละจุด เนื่องจาก $h(x) < 0$ เป็นคลาส -1 และ $h(x) > 0$ เป็นคลาส +1 ดังนั้นระยะทางของจุด x จากไฮเปอร์เพลน (แทนระยะทางด้วย δ) จะได้ดังนี้

$$\delta = yr = \frac{y\, h(x)}{\|w\|} \tag{2-28}$$

และสำหรับระยะทางจากจุดออริจิน $x = 0$ จะได้ระยะทาง ดังสมการที่ (2-29)

$$r = \frac{h(0)}{\|w\|} = \frac{w^T 0 + b}{\|w\|} = \frac{b}{\|w\|} \tag{2-29}$$

พิจารณาตัวอย่างข้อมูล 2 มิติ จากรูปที่ 2.13 ไฮเปอร์เพลนเป็นเส้นตรง และกำหนดให้ทุกจุด $X = (x_1, x_2)^T$ จากรูป ไฮเปอร์เพลนที่แสดงจะอยู่ในรูปของสมการ

$$h(x) = w^T x + b = w_1 x_1 + w_2 x_2 + b = 0 \tag{2-30}$$

หากจัดสมการใหม่ จะได้เป็น

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2} \tag{2-31}$$

เมื่อ $-\frac{w_1}{w_2}$ เป็นความชันของเส้น และ $-\frac{b}{w_2}$ เป็นตัวแบ่งตามมิติที่ 2 และเมื่อพิจารณา 2 จุดใด ๆ บนไฮเปอร์เพลน ดังนี้ $p = (p_1, p_2) = (4,0)$ และ $q = (q_1, q_2) = (2,5)$ ดังนั้นความชันจะมีค่าเป็น

$$-\frac{w_1}{w_2} = \frac{q_2 - p_2}{q_1 - p_1} = -\frac{5 - 0}{2 - 4} = -\frac{5}{2} \tag{2-32}$$

ซึ่งหมายถึง $w_1 = 5$ และ $w_2 = 2$ สำหรับจุด (4,0) บนไฮเปอร์เพลนเราสามารถคำนวณค่า b ได้ดังนี้

$$b = -5x_1 - 2x_2 - 5 \,.\, 4 - 2 \,.\, 0 = -20 \tag{2-33}$$

ดังนั้น $w = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$ คือเวกเตอร์ถ่วงน้ำหนัก และ $b = -20$ คือค่าไบแอส และสมการของไฮเปอร์เพลน เป็นดังนี้

$$w^T x + b = (5 \;\; 2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 20 = 0 \tag{2-34}$$

และเราสามารถหาระยะทางจากจุดออริจินไปยังไฮเปอร์เพลนได้ ดังนี้

$$\delta = y\, r = -1\, r = \frac{-b}{\|w\|} = \frac{-(-20)}{\sqrt{29}} = 3.71 \tag{2-35}$$

2.4.3 มาร์จินและซัพพอร์ตเวกเตอร์ของไฮเปอร์เพลน (Margin and Support Vector of Hyperplane)

เมื่อกำหนดข้อมูลฝึกสอน (Training Dataset) ของแต่ละจุด คือ $D = \{x_i, y_i\}_{i=1}^{n}$ โดยที่คลาส $y_i \in \{+1, -1\}$ และไฮเปอร์เพลน $h(x) = 0$ เราสามารถหาระยะ δ_i จากแต่ละจุด x_i ไปยังไฮเปอร์เพลน ได้ ดังนี้

$$\delta_i = \frac{y_i h(x_i)}{\|w\|} = \frac{y_i(w^T x_i + b)}{\|w\|} \tag{2-36}$$

พิจารณาทุก n จุด โดยสามารถหามาร์จินของการจำแนกแบบลิเนียร์ ซึ่งเป็นระยะทางที่สั้นที่สุดจากจุดไปยังไฮเปอร์เพลน ดังนี้

$$\delta_i = \min_{x_i} \left\{ \frac{y_i(w^T x_i + b)}{\|w\|} \right\} \tag{2-37}$$

ถ้ากำหนดให้ δ^* คือระยะทางที่สั้นที่สุดจากจุด x ไปยังไฮเปอร์เพลน และ $\delta^* \neq 0$ เนื่องจาก $h(x)$ ถูกกำหนดให้เป็นเส้นไฮเปอร์เพลนที่ทำการจำแนก เราจะเรียกทุก ๆ จุด หรือเวกเตอร์ที่มีค่าระยะทางสั้นที่สุดจากมาร์จินไปยังไฮเปอร์เพลนว่า ซัพพอร์ตเวกเตอร์ ซึ่งแสดงสมการคำนวณระยะทางสั้นที่สุดได้ดังนี้

$$\delta^* = \frac{y^*(w^T x^* + b)}{\|w\|} \tag{2-38}$$

จากสมการที่ (2-38) ตัวแปร y^* เป็น Class Label ของ x^* โดยที่ x^* คือจุดที่เป็นซัพพอร์ตเวกเตอร์ และจากสมการ $y^*(w^T x^* + b)$ จะให้ค่าของระยะทางที่แท้จริงของซัพพอร์ตเวกเตอร์ไปยังไฮเปอร์เพลน ในขณะที่ตัวหาร $\|w\|$ จะเป็น Relative Distance ในเทอมของ w เมื่อพิจารณา Canonical Hyperplane ซึ่งมีหน่วยเป็น Scalar จากสมการที่ (2-39)

$$sh(x) = sw^T x + s\,b = (sw)^T x + (sb) = 0 \tag{2-39}$$

เพื่อที่จะให้ได้ Canonical Hyperplane ที่เหมาะสมนั้น เราจะทำการเลือกค่า scalar **s** ที่มีระยะทางจากไฮเปอร์เพลนไปซัพพอร์ตเวกเตอร์ที่มีค่าเท่ากับ 1 นั่นคือ

$$sy^*(w^T x^* + b) = 1 \tag{2-40}$$

ซึ่งหมายถึง

$$s = \frac{1}{y^*(w^T x^* + b)} = \frac{1}{y^* h(x^*)} \tag{2-41}$$

และตั้งแต่นี้เป็นต้นไป เราจะสมมติว่า Separating Hyperplane ใด ๆ เป็น Canonical นั่นคือเป็นไฮเปอร์เพลนที่เหมาะสมที่ใช้ในการจำแนก และมาร์จินสามารถคำนวณหาได้ ดังนี้

$$\delta^* = \frac{y^* h(x^*)}{\|w\|} = \frac{1}{\|w\|} \tag{2-42}$$

สำหรับ Canonical Hyperplane และสำหรับ ซัพพอร์ตเวกเตอร์ x_i^* (กับ label y_i^*) แต่ละตัว จะได้ $y_i^* h(x_i^*) = 1$ และสำหรับจุดใด ๆ ที่ไม่ใช่ซัพพอร์ตเวกเตอร์ นั่นคือ $y_i h(x_i) > 1$ เนื่องจากจุดใด ๆ ที่ไม่ใช่ซัพพอร์ตเวกเตอร์ จุดเหล่านั้นจะมีระยะห่างจากไฮเปอร์เพลนมากกว่าจุดที่เป็นซัพพอร์ตเวกเตอร์ดังนั้นสมการด้านล่างนี้จะเป็นการแสดงถึงจุดใด ๆ ที่อยู่ในชุดข้อมูล D

$$y_i(w^T x_i + b) \geq 1, \quad \text{สำหรับ } x_i \in D \tag{2-43}$$

จากรูปที่ 2.14 แสดงซัพพอร์ตเวกเตอร์และ มาร์จินของไฮเปอร์เพลน ซึ่งกำหนดสมการของไฮเปอร์เพลนดังนี้

$$h(x) = \begin{pmatrix} 5 \\ 2 \end{pmatrix}^T x - 20 = 0$$

พิจารณาซัพพอร์ตเวกเตอร์ $x^* = (2,2)$ และอยู่ในคลาส $y^* = -1$ จากสมการในการหา Canonical Hyperplane นั้นเราจะต้องทำการแปลงเวกเตอร์ถ่วงน้ำหนัก และไบแอส ให้เป็น Scalar s โดยใช้สมการที่ (2-41)

$$s = \frac{1}{y^* h(x^*)} = \frac{1}{-1(\binom{5}{2}^T \binom{2}{2} - 20)} = \frac{1}{6}$$

ดังนั้น จะได้เวกเตอร์ถ่วงน้ำหนัก ที่แปลงแล้ว เป็น

$$w = \frac{1}{6}\binom{5}{2} = \binom{5/6}{2/6}$$

และจะได้ไบแอส ที่แปลงแล้ว เป็น

$$b = \frac{-20}{6}$$

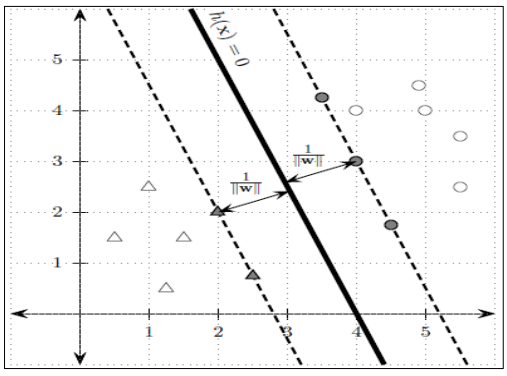
Canonical ของ ไฮเปอร์เพลน แสดงได้ดังนี้

$$h(x) = \binom{5/6}{2/6} x - \frac{20}{6} = \binom{0.833}{0.333}^T x - 3.33$$

และมาร์จินของ Canonical Hyperplane คือ

$$\delta^* = \frac{y^* h(x^*)}{\|w\|} = \frac{1}{\sqrt{(\frac{5}{6})^2 + (\frac{2}{6})^2}} = \frac{6}{\sqrt{29}} = 1.114$$

ดังนั้นในตัวอย่างการคำนวณนี้ จากรูปที่ 2.14 จะเห็นว่ามีซัพพอร์ตเวกเตอร์จำนวน 5 จุด (แสดงจุดที่ระบายสีทึบ) คือ (2,2) และ (2.5, 0.75) สังกัดอยู่ในคลาส $y = -1$ แสดงเป็นรูปสามเหลี่ยม และ (3.5, 4.25), (4, 3) และ (4.5, 1.75) สังกัดอยู่ในคลาส $y = +1$ แสดงเป็นรูปวงกลม



รูปที่ 2.14 มาร์จินของไฮเปอร์เพลนที่เหมาะสม (Zaki, 2014)

2.4.4 การแบ่งแยกเชิงเส้น (Linear Separation)

การแบ่งแยกเชิงเส้นเพื่อจำแนกข้อมูลจะใช้ไฮเปอร์เพลนที่มีความกว้างของมาร์จินมากที่สุด (Maximum Margin Hyperplane) ด้วยหลักการของซัพพอร์ตเวกเตอร์แมชชีนที่จะทำการเลือก Canonical Hyperplane ซึ่งถูกกำหนดโดยค่าเวกเตอร์ถ่วงน้ำหนัก w และ ไบแอส b และผลลัพธ์ก็คือจะได้ค่ามาร์จินที่กว้างมากที่สุดที่สามารถจะจำแนกได้โดยไฮเปอร์เพลนที่เหมาะสมที่สุดดังสมการ $h(x) \equiv w^T x + b = 0$ ถ้า δ_h^* คือค่ามาร์จินสำหรับไฮเปอร์เพลน $h(x) = 0$ ดังนั้นจุดมุ่งหมายของการหาไฮเปอร์เพลนที่ดีและเหมาะสมที่สุดคือการหาค่า h^*

$$h^* = \arg\max_h\{\delta_h^*\} = \arg\max_{w,b}\left\{\frac{1}{\|w\|}\right\} \tag{2-44}$$

สำหรับเป้าหมายของซัพพอร์ตเวกเตอร์แมชชีนในการหาไฮเปอร์เพลนที่มีความกว้างของมาร์จินมากที่สุดนั่นคือ $\frac{1}{\|w\|}$ ซึ่งสอดคล้องกับสมการที่ (2-43) นั่นคือ $y_i w^T x_i + b) \geq 1$ สำหรับทุกจุด $X_i \in D$ ข้อสังเกตคือ แทนที่เราจะทำการหาค่ามาร์จินที่กว้างที่สุดด้วยค่า $\frac{1}{\|w\|}$ เราก็จะทำการหาค่า $\|w\|$ ที่น้อยสุดซึ่งจะได้สูตรดังสมการด้านล่างนี้

Objective Function: $\min_{w,b}\left\{\frac{\|w\|^2}{2}\right\}$ (2-45)

Linear Constraint: $y_i(w^T x_i + b) \geq 1, \quad \forall x_i \in D$ (2-46)

แนวคิดที่สำคัญอีกอย่างหนึ่งเพื่อที่จะทำการหาไฮเปอร์เพลนที่เหมาะสมที่สุดนั้น คือ การนำค่า Lagrange Multiplier α_i เข้าไปคูณกับสมการที่ (2-45) โดยอยู่ในเงื่อนไขของ Karush-Kuhn-Tucker (KKT) ดังนั้นจากผลการคูณจะได้สมการที่ (2-47)

$$\alpha_i(y_i(w^T x_i + b) - 1) = 0 \quad \text{and} \quad \alpha_i \geq 0 \tag{2-47}$$

สำหรับการหาค่าเวกเตอร์ถ่วงน้ำหนัก และ ค่าไบแอส เราจะต้องได้ค่า α_i สำหรับ i = 1,…,n มาก่อน และหลังจากนั้นก็จะสามารถหาค่าเวกเตอร์ถ่วงน้ำหนัก w และ ค่าไบแอส b ได้ ซึ่งสอดคล้องกับเงื่อนไขของ KKT ดังสมการที่ (2-47) และจะได้ผลลัพธ์ออกมาเป็น 2 กรณี คือ

i) $\alpha_i = 0$, หรือ

ii) $y_i(w^T x_i + b) - 1 = 0$, ซึ่งหมายถึง $y_i(w^T x_i + b) = 1$

ผลลัพธ์ที่ได้มานี้มีความสำคัญมาก เนื่องจาก ถ้า $\alpha_i > 0$ จะทำให้กรณี $y_i(w^T x_i + b) = 1$ เป็นจริง โดยจุด x_i คือซัพพอร์ตเวกเตอร์ ในทางกลับกัน ถ้า $y_i(w^T x_i + b) > 1$ จะหมายถึงกรณี $\alpha_i = 0$ เป็นจริง นั่นคือถ้าหากจุดใด ๆ ที่ไม่ใช่ซัพพอร์ตเวกเตอร์ จะได้ $\alpha_i = 0$

เนื่องจากค่า α_i จะต้องถูกนำไปคูณกับทุกจุดข้อมูล ดังนั้นเราสามารถคำนวณค่าเวกเตอร์ถ่วงน้ำหนัก w จากสมการที่ (2-48)

$$w = \sum_{i,\alpha_i>0} \alpha_i y_i x_i \tag{2-48}$$

กล่าวอีกนัยหนึ่งนั้นค่า w ที่ได้จะเป็น Linear Combination ของซัพพอร์ตเวกเตอร์ทั้งหมด ซึ่ง α_i แต่ละตัวจะบ่งบอกถึงน้ำหนักของซัพพอร์ตเวกเตอร์ ดังนั้นหากค่า $\alpha_i = 0$ นั่นหมายถึงจุดนั้น ๆ ไม่ใช่ซัพพอร์ตเวกเตอร์ จึงไม่จำเป็นจะต้องคำนวณหาค่า w

สำหรับการคำนวณหาค่าไบแอส b นั้น เราจะต้องคำนวณหาค่า b ของซัพพอร์ตเวกเตอร์ แต่ละตัวก่อน ดังนี้

$$\alpha_i(y_i(w^T x_i + b) - 1) = 0 \tag{2-49}$$

$$y_i(w^T x_i + b) = 1 \tag{2-50}$$

$$b_i = \frac{1}{y_i} - w^T x_i = y_i - w^T x_i \tag{2-51}$$

และสามารถหาค่าเฉลี่ยของไบแอสจากซัพพอร์ตเวกเตอร์ทั้งหมดได้ ดังนี้

$$b = \text{avg}_{\alpha_i > 0}\{b_i\} \tag{2-52}$$

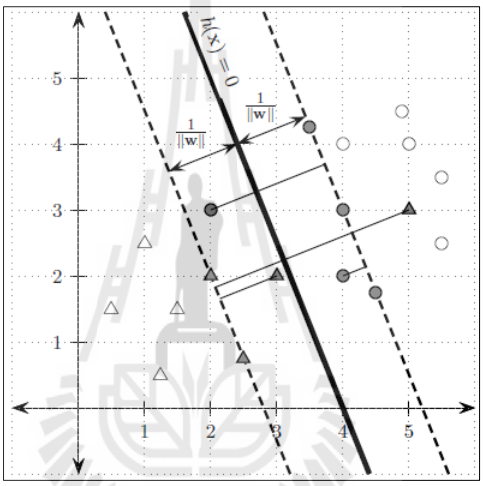
การจำแนกด้วยซัพพอร์ตเวกเตอร์ จุดประสงค์สำคัญคือการหาค่าไฮเปอร์เพลนที่เหมาะสมที่สุด นั่นคือ $h(x) = w^T x + b$ ซึ่งสำหรับจุดใหม่ที่เราจะนำเข้ามาทำการจำแนกนั้น เราสามารถทำนายได้ดังสูตรนี้

$$\hat{y} = \text{sign}\big(h(z)\big) = \text{sign}(w^T z + b) \tag{2-53}$$

เมื่อค่า sign(.) จะได้ผล +1 ถ้าผลลัพธ์เป็นบวก และให้ค่า -1 เมื่อผลลัพธ์เป็นลบ

2.4.5 ซัพพอร์ตเวกเตอร์แมชชีนแบบซอฟต์มาร์จิน (Soft Margin SVM)

จากหัวข้อที่ผ่านมาเราได้สมมติว่า การจำแนกแบบลิเนียร์นั้นสามารถทำได้อย่างสมบูรณ์แบบ โดยซัพพอร์ตเวกเตอร์แต่ละตัวจะอยู่บนเส้นของมาร์จินเท่านั้น แต่ในความเป็นจริงแล้ว การจำแนกอาจไม่จำเป็นต้องสมบูรณ์แบบ ซึ่งอาจจะอนุญาตให้จุดที่เป็นซัพพอร์ตเวกเตอร์ ไม่จำเป็นต้องอยู่บนเส้นมาร์จินเสมอไป เราสามารถอนุญาตให้ซัพพอร์ตเวกเตอร์บางจุด อยู่ในขอบเขตระหว่างมาร์จินกับไฮเปอร์เพลนได้ ซึ่งทำให้การจำแนกนั้นมีความยืดหยุ่นขึ้น ดังแสดงในรูปที่ 2.15 วิธีการนี้จะทำการหาค่า Slack Variable เพิ่มเติมในการตัดสินความยืดหยุ่นของมาร์จิน



รูปที่ 2.15 ซอฟต์มาร์จินไฮเปอร์เพลน จุดที่ระบายสีทึบหมายถึงซัพพอร์ตเวกเตอร์ ส่วนความกว้างของมาร์จินเท่ากับ $\frac{1}{\|w\|}$ (Zaki, 2014)

จากที่ได้กล่าวมาแล้วเราสามารถทำการจำแนกข้อมูลแบบลิเนียร์และได้ไฮเปอร์เพลนที่เหมาะสมจากสมการ $y_i(w^T x_i + b) \geq 1$ ซัพพอร์ตเวกเตอร์แมชชีนสามารถตัดสินจุดใด ๆ ที่ไม่สามารถจำแนกได้ด้วยการพิจารณาค่า Slack Variables ξ_i ดังสมการที่ (2-54)

$$y_i(w^T x_i + b) \geq 1 - \xi_i \tag{2-54}$$

เมื่อ $\xi_i \geq 0$ คือค่า Slack Variable สำหรับจุด x_i ซึ่งค่า Slack Variable นี้จะแสดงถึงความล่วงล้ำเขตของจุดซัพพอร์ตเวกเตอร์ ภายในบริเวณของมาร์จิน ค่าของ Slack Variable จะแบ่งเป็น 3 ประเภทดังนี้ ถ้า $\xi_i = 0$ จุดข้อมูลจะอยู่ห่างจากไฮเปอร์เพลนด้วยระยะทางอย่างน้อย $\frac{1}{\|w\|}$ ถ้า $0 < \xi_i < 1$ แสดงว่าจุดนั้นอยู่ภายในมาร์จินและถูกจำแนกได้ถูกคลาสและอยู่ถูกฝั่งของ

ไฮเปอร์เพลน อย่างไรก็ตามหาก $\xi_i \geq 1$ แสดงว่าจุดนั้นถูกจำแนกผิดคลาสและอยู่ผิดฝั่งของไฮเปอร์เพลน

ในกรณีของ Non-separable หรือจุดข้อมูลที่ไม่สามารถจำแนกคลาสได้อย่างสมบูรณ์นั้น จะใช้เทคนิค ซอฟต์มาร์จิน ซึ่งเป้าหมายหลักของซัพพอร์ตเวกเตอร์ก็คือยังต้องหาไฮเปอร์เพลนที่เหมาะสมและมีมาร์จินกว้างที่สุด โดยอนุญาตให้มี Slack Variable ได้แต่ต้องเป็นค่าที่เหมาะสมไม่มากหรือน้อยเกินไป โดยแสดงดังสมการ (2-55) และ (2-56)

$$\textbf{Objective Function:} \quad \min_{w,b,\xi_i} \left\{ \frac{\|w\|^2}{2} + C \sum_{i=1}^{n} (\xi_i)^k \right\} \tag{2-55}$$

$$\textbf{Linear Constraints:} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \ \ \forall x_i \in D \tag{2-56}$$
$$\xi_i \geq 0 \, \forall x_i \in D$$

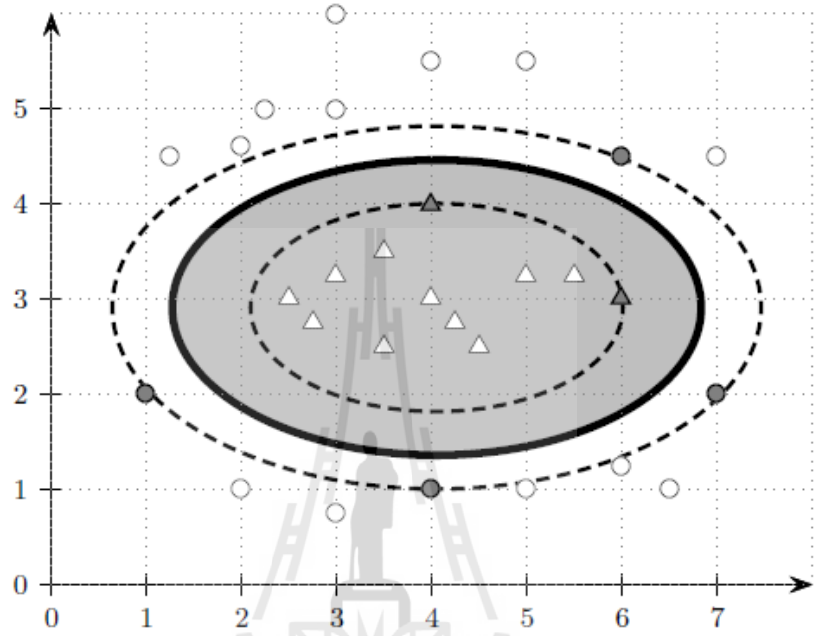
เมื่อ C และ k เป็นค่าคงที่ซึ่งเป็นตัวบ่งบอกถึงค่าใช้จ่าย (Cost) ของการจำแนกที่ผิดพลาด (Misclassification) ตัวอย่างคือ หากค่า C มีค่าเป็น 0 หรือเข้าใกล้ 0 นั่นหมายถึงไม่อนุญาตให้จุดใด ๆ ถูกจำแนกผิด ดังนั้นสมการ (2-55) จึงทำการหามาร์จินที่กว้างที่สุดตามปกติ ในทางกลับกันหาก C มีค่ามากขึ้น เช่น C เข้าใกล้ค่า ∞ (Infinity) ก็จะมีผลกระทบกับการหาค่ามาร์จินที่เหมาะสมอย่างมาก สำหรับค่าคงที่ k นั้นจะมีการตั้งค่าเป็น 2 ค่าด้วยกัน คือ 1 และ 2 ซึ่งถ้าหาก k=1 จะเรียกว่า Hinge Loss จุดประสงค์คือต้องการหาจำนวนผลรวมของค่า Slack Variables ที่น้อยที่สุด แต่หาก k=2 เรียกว่า Quadratic Loss จุดประสงค์คือต้องการหาค่าผลรวมยกกำลังสองที่น้อยที่สุดของ Slack Variables

2.4.6 เคอร์เนลฟังก์ชันกับซัพพอร์ตเวกเตอร์แมชชีน (Kernel Function with SVM)

ในบางกรณีการจำแนกข้อมูลแบบลิเนียร์นั้นไม่สามารถที่จะจำแนกข้อมูลได้อย่างถูกต้อง จึงจำเป็นต้องนำเคอร์เนลฟังก์ชัน (Kernel Function) มาใช้ร่วมกับซัพพอร์ตเวกเตอร์แมชชีน หลักการคือจะต้องทำการแมพ (Map) หรือแปลง d-dimensional ของจุด x_i ในอินพุทสเปซไปเป็นจุด $\emptyset(x_i)$ ในมิติข้อมูลที่สูงขึ้น (High-dimensional Feature Space)

รูปที่ 2.16 แสดงตัวอย่างลักษณะข้อมูลที่การจำแนกข้อมูลแบบลิเนียร์ไม่สามารถที่จะจำแนกได้อย่างถูกต้อง อย่างไรก็ตามจากข้อมูลดังรูปที่ 2.16 นั้นสามารถใช้ Quadratic Classifier ในการจำแนกคลาสสองคลาสได้ โดยกำหนดให้ข้อมูลในอินพุทสเปซขนาด 2 มิติ $X = (x_1, x_2)^T$ ถูกแปลงไปเป็นฟีเจอร์สเปซ (Feature Space) ที่มีห้ามิติเป็น $(x_1, x_2, x_1^2 x_2^2, x_1, x_2)$ โดยใช้สูตรการแปลงดังนี้ $\emptyset(X) = (\sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2)^T$ ดังนั้นจึงเป็นไปได้ที่จะทำการจำแนก

ข้อมูลแบบลิเนียร์บนข้อมูลมิติที่สูงขึ้นในฟีเจอร์สเปซ ซึ่งจากรูปที่ 2.16 จะเห็นได้ว่ารูปวงรีที่เป็นเส้นทึบสีดำ จะสามารถจำแนกคลาส 2 คลาสออกจากกันได้ (คลาสแสดงรูปสามเหลี่ยมและวงกลม) ส่วนจุดต่าง ๆ ที่เป็นซัพพอร์ตเวกเตอร์นั้นจะแสดงจุดที่ระบายสีทึบ ซึ่งอยู่บนเส้นมาร์จิน (แสดงด้วยเส้นประ)



รูปที่ 2.16 ซัพพอร์ตเวกเตอร์แมชชีนแบบไม่เป็นเชิงเส้น จุดที่ระบายสีทึบคือซัพพอร์ตเวกเตอร์ (Zaki, 2014)

สำหรับเคอร์เนลฟังก์ชันมีหลายแบบให้เลือกใช้งาน ซึ่งหากลักษณะของข้อมูลไม่สามารถที่จะจำแนกแบบลิเนียร์แล้วสามารถเลือกใช้เคอร์เนลฟังก์ชันแต่ละแบบได้ เนื่องจากชุดข้อมูลแต่ละชุดนั้นมีการกระจายและมิติข้อมูลที่แตกต่างกัน ซึ่งส่วนใหญ่แล้วนักวิจัยจะเลือกใช้เคอร์เนลฟังก์ชันหลาย ๆ แบบ และนำผลความแม่นมาเปรียบเทียบกัน ซึ่งอาจบอกไม่ได้ว่าเคอร์เนลฟังก์ชันใดเหมาะกับข้อมูลประเภทใด ดังนั้นจึงต้องมีการทดลองกับเคอร์เนลฟังก์ชันมากกว่าหนึ่งแบบ สำหรับเคอร์เนลฟังก์ชันที่นิยมใช้กันในปัจจุบันสามารถแสดงได้ดังตารางที่ 2.1

ตารางที่ 2.1 เคอร์เนลฟังก์ชันที่นิยมใช้ร่วมกับซัพพอร์ตเวกเตอร์แมชชีน

Kernel Type	Function $K(x, x_i)$:
Radial Basis Function (RBF)	$\exp(-\gamma\|x - x_i\|^2), \quad \gamma > 0$
Inverse multiquadratic	$\frac{1}{\sqrt{\|x - x_i\| + \eta}}$

ตารางที่ 2.1 เคอร์เนลฟังก์ชันที่นิยมใช้ร่วมกับซัพพอร์ตเวกเตอร์แมชชีน (ต่อ)

Kernel Type	Function $K(x, x_i)$:
Polynomial of degree *d*	$((x^T . x_i) + \eta)^d$
Sigmoidal	$\tanh(\gamma(x^T . x_i) + \eta), \gamma > 0$
Linear	$x^T . x_i$

2.5 เกณฑ์ที่ใช้ในการวัดประสิทธิภาพของโมเดล

2.5.1 เกณฑ์ความแม่น (Accuracy)

เกณฑ์ความแม่นเป็นเกณฑ์ที่ใช้วัดระดับความถูกต้องหรือความแม่นในการจำแนกประเภทข้อมูลของโมเดลที่เรียนรู้จากชุดข้อมูลฝึก และจะแสดงผลลัพธ์ที่จำแนกได้ว่าเป็นคลาสบวก (Positive Class: P) หรือ คลาสลบ (Negative Class: N) ซึ่งผลลัพธ์ที่ได้สามารถมีได้ 4 แบบ คือ

1. True Positive (TP) หมายความว่า ผลลัพธ์ที่ได้จากการทำนายคือ P และค่าจริง ๆ ก็คือ P ด้วย เช่น หากโมเดลทำนายผู้ป่วยคนหนึ่งว่าเป็นมะเร็งเต้านมแล้วผลการตรวจสอบบอกว่าเป็นมะเร็งเต้านมจริง
2. False Positive (FP) หมายความว่า ผลลัพธ์ที่ได้จากการทำนายคือ P แต่ว่าค่าจริง ๆ แล้วคือ N เช่น หากโมเดลทำนายผู้ป่วยคนหนึ่งว่าเป็นมะเร็งเต้านม แต่ผลการตรวจสอบบอกว่าไม่เป็นมะเร็งเต้านม
3. True Negative (TN) หมายความว่า ผลลัพธ์ที่ได้จากการทำนายคือ N และค่าจริง ๆ ก็คือ N ด้วย เช่น หากโมเดลทำนายผู้ป่วยคนหนึ่งว่าไม่เป็นมะเร็งเต้านมแล้วผลการตรวจสอบบอกว่าไม่เป็นมะเร็งเต้านมจริง
4. False Negative (FN) หมายความว่า ผลลัพธ์ที่ได้จากการทำนายคือ N แต่ว่าค่าจริง ๆ แล้วคือ P เช่น หากโมเดลทำนายผู้ป่วยคนหนึ่งว่าไม่เป็นมะเร็งเต้านม แต่ผลการตรวจสอบบอกว่าเป็นมะเร็งเต้านม

ซึ่งสามารถคำนวณค่าความแม่นได้ดังสมการที่ (2-57)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2-57}$$

2.5.2 ค่า Sensitivity

ค่า Sensitivity หรือค่า Recall จะบ่งบอกถึงความสามารถของการทดสอบที่สามารถวินิจฉัยผู้ป่วยที่เป็นโรคมะเร็งได้อย่างถูกต้อง โดยพิจารณาข้อมูลทำนายที่อยู่ในคลาส Positive เทียบกับข้อมูลจริงทั้งหมดของคลาส Positive สามารถคำนวณหาค่า Sensitivity ได้ดังสมการที่ (2-58)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2-58}$$

2.5.3 ค่า Specificity

ค่า Specificity ใช้สำหรับบ่งบอกความสามารถของการทดสอบที่สามารถวินิจฉัยผู้ที่ไม่ป่วยเป็นมะเร็งได้อย่างถูกต้อง สามารถคำนวณหาค่า Specificity ได้ดังสมการที่ (2-59)

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2-59}$$

2.5.4 ค่า Precision

ค่า Precision เป็นการวัดความแม่นของโมเดล โดยพิจารณาข้อมูลทำนายที่อยู่ในคลาส Positive เทียบกับจำนวนข้อมูลที่ทำนายว่าเป็นคลาส Positive ทั้งหมด แสดงได้ดังสมการที่ (2-60)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2-60}$$

2.5.5 ค่า F-measure

ค่า F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดล หรือค่าเฉลี่ยที่ให้ความสำคัญกับ Precision และ Recall เท่า ๆ กัน แสดงได้ดังสมการที่ (2-61)

$$\text{F} - \text{measure} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{2-61}$$

2.5.6 Confusion Matrix

โดยทั่วไปแล้วการจะแสดงประสิทธิภาพของโมเดลจำแนกนั้นนิยมที่จะใช้ตาราง Confusion Matrix มาแสดง โดยการเก็บข้อมูลจำนวนแถวที่จำแนกจากกลุ่มข้อมูลจริงและกลุ่ม

ข้อมูลจากการทำนาย ตาราง Confusion Matrix นั้นจะมีขนาด m x m โดยที่ m คือจำนวนของคลาส ดังแสดงในรูปที่ 2.17 เป็นตัวอย่างในการจำแนกข้อมูลที่มีทั้งหมด 80 ข้อมูล

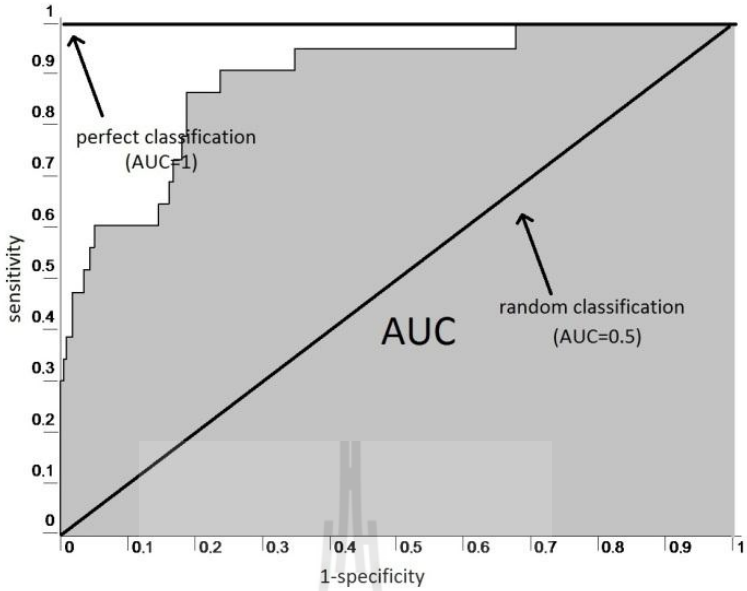
		ค่าความจริง (Actual)		
		Positive	Negative	
ค่าทำนาย (Predict)	Positive	True Positive (TP) = 68	False Positive (FP) = 4	**Accuracy** =(TP+TN)/ (TP+TN+FP+FN) = (68+5)/ (68+5+4+3) = **91.25%**
	Negative	False Negative (FN) = 3	True Negative (TN) = 5	
		Sensitivity = TP/(TP+FN) = 68/(68+3) = **95.77%**	**Specificity** = TN/(FP+TN) = 5/(4+5) = **55.56%**	

รูปที่ 2.17 ตัวอย่าง Confusion Matrix

2.5.7 กราฟ Receiver Operation Characteristic (ROC)

ROC เป็นมาตรฐานการวัดประสิทธิภาพอีกประเภทหนึ่งซึ่งมีวัตถุประสงค์เพื่อนำมาใช้ในการประเมินประสิทธิภาพในการจำแนกข้อมูล การสร้างกราฟ ROC จะเป็นการพล็อต (Plot) กราฟ 2 มิติ โดยใช้แกน X แทนค่า False Positive Rate (1- Specificity) และ แกน Y แทนค่า True Positive Rate (Sensitivity) โดยการแปรค่าจุดตัด (Cut-off point) และหาค่าพื้นที่ใต้กราฟ (Area under Curve)

เครื่องมือในการตรวจวินัจฉัยที่ดีควรมีค่า Sensitivity สูง และมี Specificity สูง ซึ่งประการหลังจะทำให้มีค่า False Positive Rate ต่ำ ส่งผลให้กราฟ ROC เข้าชิดมุมซ้ายบนมากที่สุดจึงทำให้มีพื้นที่ใต้กราฟมาก นอกจากนี้การสร้างกราฟ ROC ยังช่วยในการเปรียบเทียบประสิทธิภาพของการตรวจวินิจฉัยโดยการใช้เครื่องมือหรือวิธีการที่แตกต่างกันได้ด้วย โดยเปรียบเทียบพื้นที่ใต้เส้นโค้งของการตรวจวินัยฉัยแต่ละชนิด พื้นที่ใต้โค้งที่มากกว่าแสดงถึงประสิทธิภาพที่สูงกว่า ยกตัวอย่างเช่น การเปรียบเทียบประสิทธิภาพการจำแนกมะเร็งเต้านมจากภาพแมมโมแกรมโดยใช้อัลกอริทึมต่างชนิดกัน สามารถนำผลการจำแนกของอัลกอริทึมต่าง ๆ มาสร้างกราฟ ROC เพื่อเปรียบเทียบประสิทธิภาพของแต่ละโมเดลได้ รูปที่ 2.18 แสดงตัวอย่างกราฟ ROC



รูปที่ 2.18 กราฟ ROC และพื้นที่ใต้กราฟเพื่อใช้ในการวัดประสิทธิภาพของโมเดลการจำแนก (http://www.intechopen.com/books/data-mining-applications-in-engineering-and-medicine)

2.6 งานวิจัยที่เกี่ยวข้อง

การวิเคราะห์ภาพแมมโมแกรมเพื่อจำแนกมะเร็งเต้านมนั้นเป็นงานวิจัยที่น่าสนใจ เนื่องจากการวิเคราะห์ด้วยภาพนั้นช่วยให้การวิเคราะห์สะดวกขึ้น โดยคนไข้ที่มีก้อนเนื้อผิดปกติในบริเวณทรวงอกและต้องการทราบว่าตนป่วยเป็นมะเร็งเต้านมหรือไม่นั้นสามารถเข้ารับการตรวจวิเคราะห์ได้ โดยไม่ต้องเข้ารับการผ่าตัดชิ้นเนื้อซึ่งมีความเสี่ยงมากกว่าการถ่ายภาพรังสี ดังนั้นจึงมีงานวิจัยที่หลากหลายในการพยายามที่จะวิเคราะห์และจำแนกมะเร็งเต้านมให้มีความแม่นและมีประสิทธิภาพสูง จากความรู้ของนักรังสีวิทยาลักษณะรูปร่างของก้อนเนื้อในทรวงอกที่เป็นก้อนเนื้อไม่อันตรายและก้อนเนื้อร้ายนั้นมีรูปร่าง ลักษณะ และความหนาแน่นที่แตกต่างกัน แต่ในหลายปีที่ผ่านมาการวิเคราะห์ของนักรังสีวิทยาอาจมีความผิดพลาดได้ การวิเคราะห์นั้นจึงต้องใช้ประสบการณ์ การฝึกฝน และความรู้เฉพาะทาง แต่ถึงกระนั้นก็ยังมีการสำรวจพบว่าประมาณ 10% ของก้อนเนื้อร้ายทั้งหมดในภาพแมมโมแกรม ถูกอ่านและวิเคราะห์ผิดโดยนักรังสีวิทยา ซึ่งทำให้มีค่า False Positive ค่อนข้างสูง และก้อนเนื้อส่วนใหญ่ที่มีการวิเคราะห์ผิดพลาดนั้นจะอยู่ในทรวงอกที่มีความหนาแน่นสูง (Jackson et al., 1993)

Sivaramakrishna และ คณะ(2002) ได้ทำการสรุปความผิดพลาดในการวิเคราะห์ภาพแมมโมแกรมของนักรังสีวิทยาว่าเกิดจากสาเหตุสำคัญ 3 ประการ คือ 1.ลักษณะสำคัญของก้อนเนื้อจากภาพแมมโมแกรมที่จะทำการวินิจฉัยมีความไม่ชัดเจน 2.สัญญาณรบกวนภายในภาพแมมโมแกรม ซึ่งอาจเกิดจากคุณภาพของเครื่องเอ็กซเรย์ 3.ลักษณะสำคัญบางประการของภาพยังคลุมเครือและไม่สามารถวิเคราะห์และวินิจฉัยได้ด้วยตาเปล่า ดังนั้นจึงมีงานวิจัยที่นักวิจัยหลายคนได้นำเสนอเพื่อทำการจำแนกมะเร็งเต้านมจากภาพแมมโมแกรมให้มีความแม่นมากยิ่งขึ้น ดังนี้

Karmilasari และ คณะ(2014) เสนออัลกอริทึมการจัดกลุ่มภาพแมมโมแกรมที่ชื่อว่า Sample k-means ซึ่งทำการวิเคราะห์ภาพแมมโมแกรมของผู้ป่วยมะเร็งเต้านมขั้นที่ 1 ขั้นที่ 2 และขั้นที่ 3 โดยใช้ขนาดพื้นที่ของก้อนเนื้อเป็นข้อมูลตัดสินการลุกลามของก้อนเนื้อมะเร็งทั้ง 3 ขั้น ในขั้นตอนแรกใช้วิธี Otsu ในการปรับความเข้มสีภายในภาพให้บริเวณก้อนเนื้อมีความชัดเจนมากยิ่งขึ้น และใช้วิธีการขยายส่วนพื้นที่ของภาพในการหาขอบเขตที่สนใจของก้อนเนื้อในภาพแมมโมแกรม หลังจากนั้นจึงใช้วิธีการหาขอบภาพ (Edge Detection Method) จากขอบเขตที่สนใจโดยวิธีตรวจหาขอบแบบแคนนี (Canny Edge Detection) และเนื่องจากขอบภาพที่ได้จากก้อนเนื้อในภาพแมมโมแกรมนั้นอาจมีขอบที่บาง ดังนั้นจึงมีการใช้เทคนิคการขยายเส้นขอบ (Dilation) ในการปรับให้เส้นขอบมีความหนาและชัดเจนขึ้น หลังจากนั้นจึงนำภาพที่ได้จากขั้นตอนการหาขอบนำมาหาพื้นที่ และนำเข้าสู่กระบวนการจัดกลุ่มโดยใช้วิธีการจัดกลุ่มด้วย k-means ซึ่งผู้ป่วยขั้นที่ 1 จะมีขนาดของก้อนเนื้อระหว่าง 3,000 ถึง 35,000 พิกเซล ผู้ป่วยขั้นที่ 2 จะมีขนาดของก้อนเนื้อระหว่าง 35,000 ถึง 85,000 พิกเซล และ ผู้ป่วยขั้นที่ 3 จะมีขนาดของก้อนเนื้อระหว่าง 85,000 ถึง 250,000 พิกเซล ในงานวิจัยนี้ได้ใช้ภาพแมมโมแกรมของผู้ที่ตรวจพบก้อนเนื้อร้ายจำนวน 33 ภาพจากฐานข้อมูล MIAS (Mammographic Image Analysis Society)

Braz และคณะ (2009) เสนออัลกอริทึมในการจำแนกภาพแมมโมแกรมชื่อ Moran's Index และ Geary's Coefficient ซึ่งใช้เฉพาะภาพบางส่วนที่เป็นบริเวณก้อนเนื้อซึ่งทำการตัดมาจากภาพแมมโมแกรมขนาดใหญ่ โดยทำการจำแนกก้อนเนื้อในภาพเป็น 2 คลาส คือ ก้อนเนื้อไม่อันตรายและก้อนเนื้อร้าย โดยในขั้นตอนแรกจะใช้วิธีการปรับปรุงภาพให้มีความเข้มสีบริเวณก้อนเนื้อมีความชัดเจนขึ้นโดยใช้วิธี Histogram Equalization จากนั้นจึงนำภาพหลังผ่านการปรับปรุงแล้วไปทำการหาลักษณะเด่นของลวดลายโดยใช้เมตริกซ์ GLCM ร่วมกับการดึงลักษณะสำคัญโดยใช้วิธี Moran's Index และ Geary's Coefficient หลังจากนั้นจึงใช้ซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกร่วมกับเคอร์เนลฟังก์ชันเรเดียลเบสิส การวัดประสิทธิภาพของอัลกอริทึมนี้ใช้ค่าความแม่น และพื้นที่ใต้กราฟ ROC เป็นเครื่องมือในการวัดประสิทธิภาพ ในการวิจัยนี้ได้ใช้ภาพแมมโมแกรมจากฐานข้อมูล DDSM จำนวน 584 ภาพซึ่งเป็นการตัดเฉพาะภาพบริเวณก้อนเนื้อมาประมวลผล

Rouhi และ คณะ (2015) เสนออัลกอริทึมในการจำแนกก้อนเนื้อไม่อันตรายและก้อนเนื้อร้ายจากภาพแมมโมแกรม โดยใช้วิธีการขยายส่วนพื้นที่ของภาพร่วมกับ Cellular Neural Network (CNN) โดยในขั้นตอนแรกจะเป็นการกำจัดสัญญาณรบกวนในภาพโดยใช้ตัวกรองมัธยฐาน หลังจากนั้นจึงทำการหาขอบเขตที่สนใจ หรือ ROIโดยใช้วิธีการขยายส่วนพื้นที่ของภาพ จากนั้นจึงนำ ROI มาหาลักษณะสำคัญ 3 แบบคือ ลักษณะสำคัญของลวดลาย ลักษณะสำคัญของฮิสโตแกรม และลักษณะสำคัญของรูปร่าง และใช้ Genetic Algorithm ช่วยในการคัดเลือกลักษณะเด่นที่เหมาะสม ในขั้นตอนสุดท้ายจึงใช้ CNN ในการจำแนก การวัดประสิทธิภาพของอัลกอริทึมนี้ใช้ค่าความแม่น และพื้นที่ใต้กราฟ ROC เป็นเครื่องมือในการวัดประสิทธิภาพ ในการวิจัยนี้ได้ใช้ภาพแมมโมแกรมจากฐานข้อมูล 2 ฐานข้อมูล คือ ภาพจากฐานข้อมูล DDSM จำนวน 170 ภาพ และภาพจากฐานข้อมูล MIAS จำนวน 93 ภาพ

Oliver และคณะ (2010) เสนออัลกอริทึมการจำแนกความหนาแน่นของก้อนเนื้อในทรวงอกด้วยวิธีการทางสถิติ (Statistical for Breast Density Segmentation) โดยพิจารณาและประเมินความหนาแน่นของจุดพิกเซลภายในภาพ วิธีการนี้จะทำการพิจารณาลักษณะเฉพาะของเนื้อเยื่อ (Tissue) ในภาพแมมโมแกรม และทำการจำแนกภาพออกเป็นบริเวณที่เป็นไขมันและบริเวณที่เป็นเนื้อเยื่อที่มีความหนาแน่นสูง โดยใช้เทคนิค PCA และ LDA ในการจำแนกภาพ และนำผลมาเปรียบเทียบกัน ในการวิจัยนี้ได้ใช้ภาพแมมโมแกรมจากฐานข้อมูล MIAS จำนวน 125 ภาพ

Al-Najdawi และคณะ (2015) เสนออัลกอริทึมสำหรับปรับปรุงภาพหลายแบบร่วมกับการจำแนกภาพโดยพิจารณาความหนาแน่นของก้อนเนื้อ ขั้นตอนแรกทำการกำจัดสัญญาณรบกวนในภาพโดยใช้ตัวกรองมัธยฐาน ทำการปรับให้ภาพมีความเนียนขึ้นโดยใช้เกาเชียนฟิลเตอร์ (Gaussian Smoothing Filter) หลังจากนั้นขั้นตอนในการปรับปรุงภาพใช้วิธี Histogram Equalization เพื่อทำการเพิ่มความเข้มสีและเพิ่มความชัดเจนในภาพแมมโมแกรม ขั้นต่อไปเป็นการหาขอบเขตของภาพหรือ ROI โดยใช้วิธี Otsu หลังจากได้ ROI แล้วจะเป็นขั้นตอนในการเติมช่องว่าง (Hole-filling) ภายใน ROI เนื่องจากภาพแมมโมแกรมบางภาพนั้น เมื่อได้ ROI มาแล้วมักมีช่องว่างเกิดขึ้นภายในภาพ ซึ่งอัลกอริทึมนี้จะต้องทำการเติมช่องว่างโดยพิจารณาจากค่าเทรสโฮล (Threshold) ที่ตั้งไว้ หลังจากนั้นจึงทำการดึงลักษณะสำคัญของความหนาแน่นในพิกเซลของภาพ โดยพิจารณาค่าเฉลี่ยความหนาแน่นของพิกเซล ในการวิจัยนี้ผู้วิจัยได้ใช้ภาพแมมโมแกรมจากฐานข้อมูลจากโรงพยาบาล King Hussein Cancer จำนวน 1,300 ภาพ

Shi และ คณะ (2010) เสนออัลกอริทึมสำหรับตรวจจับก้อนเนื้อและจำแนกภาพก้อนเนื้อจากภาพอัลตราซาวด์ (Ultrasound Image) ในงานวิจัยผู้วิจัยได้พัฒนาอัลกอริทึมสำหรับวินิจฉัยก้อนเนื้อในเต้านม่ร่วมกับการใช้ฟัซซีซัพพอร์ตเวกเตอร์แมชชีน (Fuzzy Support Vector Machine:

FSVM) โดยในขั้นตอนแรกจะทำการปรับปรุงภาพโดยใช้วิธี Histogram Equalization หลังจากนั้นจึงทำการหาขอบเขตที่สนใจ หรือ ROI ในภาพโดยใช้วิธี Markov Random Field หลังจากนั้นจึงทำการหาลักษณะสำคัญในภาพ ซึ่งในงานวิจัยนี้ได้ใช้ลักษณะสำคัญ 2 แบบ คือ ลักษณะสำคัญของลวดลาย และ ลักษณะสำคัญของฮิสโตแกรม และได้ใช้วิธี Stepwise Regression ในการคัดเลือกลักษณะสำคัญที่เหมาะสม ในขั้นตอนสุดท้ายจะเป็นการจำแนกโดยใช้ FSVM และผู้วิจัยได้ใช้วิธีประเมินประสิทธิภาพของอัลกอริทึม คือ Accuracy, Sensitivity, Specificity, Positive Predictive Value และ Negative Predictive Value

Wu และ คณะ (2012) เสนออัลกอริทึม CSVM ในการจำแนกมะเร็งเต้านมโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนร่วมกันอัลกอริทึม Genetic โดยใช้ภาพอัลตราซาวด์จำนวน 210 ภาพ ในขั้นตอนแรกจะเป็นการหาขอบเขตที่น่าสนใจในภาพด้วยวิธี Level Set จากนั้นจะทำการหาลักษณะสำคัญในภาพ ซึ่งในการวิจัยนี้จะใช้การหาลักษณะสำคัญ 2 แบบ คือ การหาลักษณะสำคัญของลวดลาย และการหาลักษณะสำคัญของรูปร่าง ต่อไปจึงใช้อัลกอริทึม Genetic ในการคัดเลือกลักษณะสำคัญที่เหมาะสม หลังจากนั้นจึงนำเข้าสู่กระบวนการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีน สำหรับการประเมินประสิทธิภาพของงานวิจัยนี้จะใช้ค่า Accuracy, Sensitivity, Specificity, Positive Predictive Value และ Negative Predictive Value รวมทั้งใช้พื้นที่ใต้กราฟ ROC ในการประเมินประสิทธิภาพด้วย

Lo และ คณะ (2012) เสนออัลกอริทึมในการใช้ซัพพอร์ตเวกเตอร์แมชชีนจำแนกภาพ MRI (Magnetic Resonance Imaging) โดยใช้ลักษณะสำคัญของลวดลาย 4 แบบ ภายในภาพ MRI คือ Fatty, Glandular, Tumor และ Muscle เป็นลักษณะสำคัญที่ใช้ในการจำแนก โดยมีการเปรียบเทียบประสิทธิภาพในการจำแนกระหว่างซัพพอร์ตเวกเตอร์แมชชีนและซีมีนส์ (C-means) ซึ่งจากผลการทดลองซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพในการจำแนกที่สูงกว่าซีมีนส์

Hassanien และ คณะ (2012) เสนออัลกอริทึมในการวินัจฉัยมะเร็งเต้านมจากภาพ MRI โดยใช้ซัพพอร์ตเวกเตอร์แมชชีนร่วมกับ Pulse Coupled Neural Networks (PCNNs) โดยใช้วิธีการเว็ฟเล็ท (Wavelet-based) ในการดึงลักษณะสำคัญจากภาพ และขั้นตอนการปรับปรุงภาพก่อนจำแนกได้ใช้วิธีฟัซซีเซ็ต (Fuzzy Sets) ในการทำให้ภาพมีความคมชัดขึ้น ในงานวิจัยนี้ได้ทำการเปรียบเทียบความแม่นระหว่างอัลกอริทึมที่นำเสนอกับอัลกอรึทึมการจำแนกอื่น ๆ ด้วย เช่น ต้นไม้ตัดสินใจ (Decision Trees) และ นิวรอลเน็ตเวิร์ก

จากการทบทวนงานวิจัยที่เกี่ยวข้อง (สรุปดังตารางที่ 2.2) พบว่างานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ภาพแมมโมแกรมเพื่อจำแนกมะเร็งเต้านม กลุ่มนักวิจัยส่วนใหญ่จะเสนอแนวคิดไปในทางเดียวกันคือมีการทำการปรับปรุงภาพก่อนด้วยวิธีการประมวลผลภาพที่หลากหลาย เช่น การกำจัด

สัญญาณรบกวนภายในภาพและการปรับความเข้มสี หลังจากนั้นจึงนำภาพที่ปรับปรุงแล้วมาทำการแบ่งแยกขอบเขตเพื่อจะพิจารณเฉพาะบริเวณก้อนเนื้อซึ่งโดยปกติจะมีความเข้มสีและความหนาแน่นของพิกเซลมากกว่าบริเวณพื้นหลัง เช่นการหา ROI โดยวิธีการหาส่วนขยายของพื้นที่ต่อจากนั้นจึงนำ ROI ที่ได้ ไปเข้ากระบวนการหาลักษณะสำคัญภายในภาพ ซึ่งเป็นการนำเอาลักษณะเฉพาะของ ROI ออกจากภาพ เช่นลักษณะสำคัญของรูปร่าง ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของฮิสโตแกรม เป็นต้น ส่วนขั้นตอนสุดท้ายคือกระบวนการจำแนก ซึ่งจากการทบทวนวรรณกรรมที่เกี่ยวข้องการจำแนกมะเร็งเต้านมพบว่า นักวิจัยส่วนใหญ่จะใช้ซัพพอร์ตเวกเตอร์แมชชีนและนิวรอลเน็ตเวิร์กในการจำแนก

ในงานวิจัยของวิทยานิพนธ์นี้เสนอขั้นตอนหลักเป็นไปในแนวทางเดียวกับงานวิจัยอื่น ๆ คือ ใช้การประมวลผลภาพปรับปรุงคุณภาพ และลดขนาดของภาพ ก่อนที่จะทำการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีน แต่ในรายละเอียดของขั้นตอนการประมวลผลภาพ วิทยานิพนธ์นี้จะเน้นที่กระบวนการหาลักษณะสำคัญของภาพที่จะช่วยให้ซัพพอร์ตเวกเตอร์แมชชีน สร้างโมเดลเพื่อการจำแนกที่มีความแม่นสูง

ตารางที่ 2.2 สรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการจำแนกมะเร็งเต้านมในภาพแมมโมแกรม

กระบวนการทำงาน	**งานวิจัยที่เกี่ยวข้อง[1]**									
	ก	**ข**	**ค**	**ง**	**จ**	**ฉ**	**ช**	**ซ**	**ฌ**	**ญ[2]**
เทคนิคการปรับปรุงภาพ										
ปรับปรุงภาพด้วย Median Filter			✓	✓						✓
ปรับปรุงภาพด้วย Gamma Correction										✓
ปรับปรุงภาพด้วย Histogram Equalization		✓			✓	✓				
ปรับปรุงภาพด้วยเทคนิคอื่น	✓			✓					✓	
หาขอบเขตที่น่าสนใจด้วย Region Growing			✓							✓
หาขอบเขตที่น่าสนใจด้วย Otsu Method					✓					
หาขอบเขตที่น่าสนใจด้วยวิธีอื่น ๆ		✓				✓	✓	✓		
หาลักษณะสำคัญด้วย Texture Feature		✓	✓			✓	✓	✓		✓
หาลักษณะสำคัญด้วย Histogram Feature	✓		✓	✓	✓	✓				✓
หาลักษณะสำคัญด้วย Shape Feature			✓				✓			✓

ตารางที่ 2.2 สรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการจำแนกมะเร็งเต้านมในภาพแมมโมแกรม (ต่อ)

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง[1]									
	ก	ข	ค	ง	จ	ฉ	ช	ซ	ฌ	ญ[2]
เทคนิคที่ใช้ในการจำแนกภาพมะเร็งเต้านม										
จำแนกด้วย Support Vector Machine		✓				✓	✓	✓	✓	✓
จำแนกด้วย Neural Network			✓						✓	
จำแนกด้วย PCA และ LDA				✓						
จำแนกหรือจัดกลุ่มด้วยวิธีอื่นๆ	✓				✓					
ข้อมูลที่ใช้ในการทดสอบ										
MIAS Database (ภาพแมมโมแกรม)	✓			✓						
DDSM Database (ภาพแมมโมแกรม)		✓	✓							✓
อื่น ๆ (ภาพอัลตราซาวด์ หรือ ภาพ MRI)					✓	✓	✓	✓	✓	

[1]งานวิจัยที่เกี่ยวข้องประกอบด้วย

ก แทนงานวิจัยของ Karmilasari และ คณะ (2014)

ข แทนงานวิจัยของ Braz และ คณะ (2009)

ค แทนงานวิจัยของ Rouhi และ คณะ (2015)

ง แทนงานวิจัยของ Oliver และ คณะ (2010)

จ แทนงานวิจัยของ Al-Najdawi และ คณะ (2015)

ฉ แทนงานวิจัยของ Shi และ คณะ (2010)

ช แทนงานวิจัยของ Wu และ คณะ (2012)

ซ แทนงานวิจัยของ Lo และ คณะ (2012)

ฌ แทนงานวิจัยของ Hassanien และ คณะ (2012)

ญ[2] แทนงานวิจัยของวิทยานิพนธ์ฉบับนี้

บทที่ 3
วิธีดำเนินการวิจัย

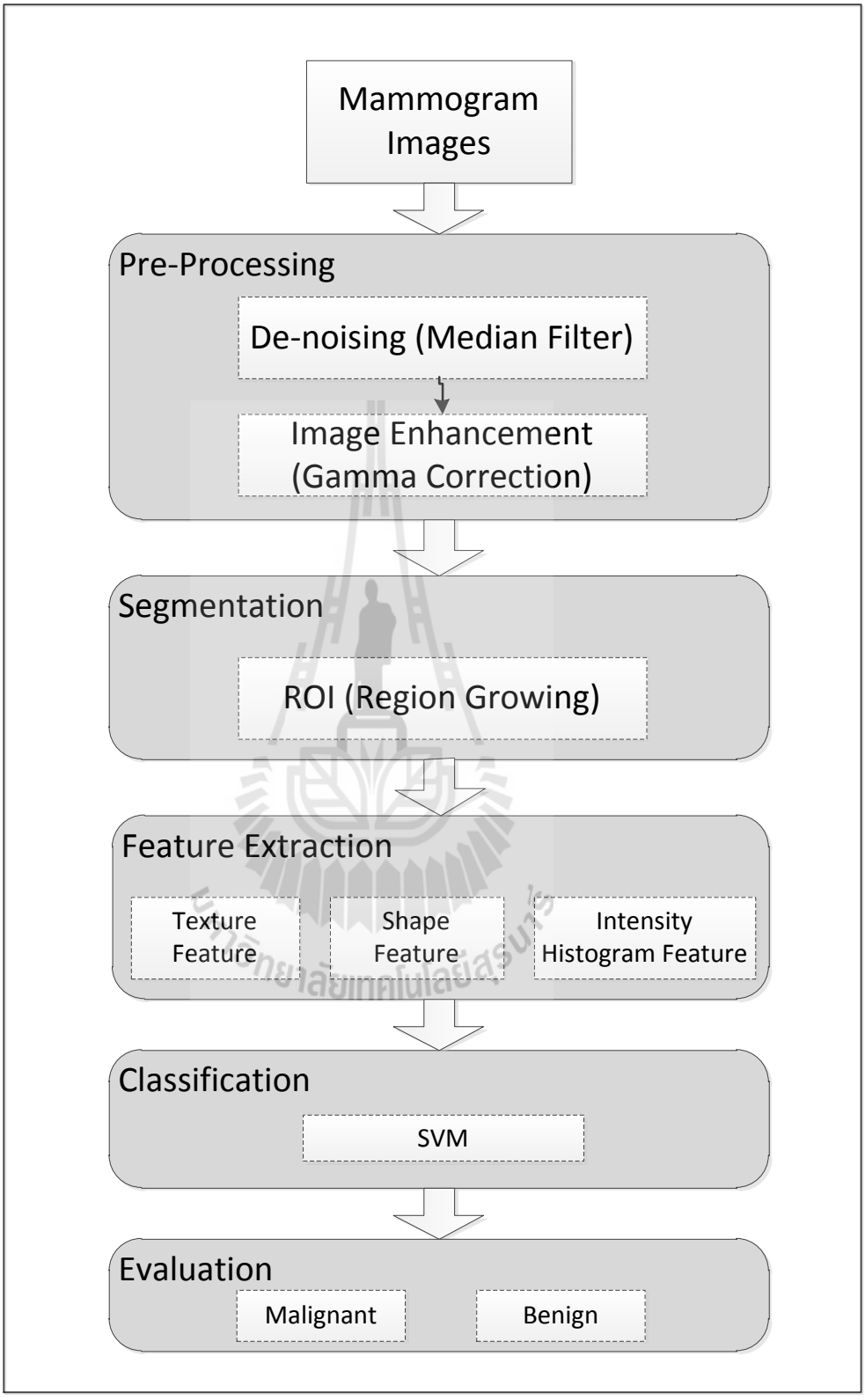
งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาอัลกอริทึมที่ใช้จำแนกมะเร็งเต้านมจากภาพก้อนเนื้อแมมโมแกรมด้วยซัพพอร์ตเวกเตอร์แมชชีนให้มีความถูกต้องในการจำแนกอยู่ในเกณฑ์ดี ในบทนี้จะกล่าวถึง วิธีการวิจัย เครื่องมือที่ใช้ในการวิจัย และกระบวนการต่าง ๆ ของการวิจัย โดยมีรายละเอียดดังนี้

3.1 ขั้นตอนวิธีของการวิจัย

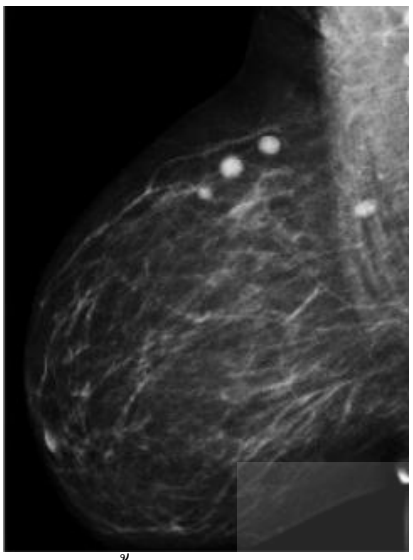
แนวคิดหลักของงานวิจัยนี้ คือ การจำแนกมะเร็งเต้านมจากภาพแมมโมแกรม โดยอาศัยวิธีการประมวลผลภาพในเบื้องต้น เพื่อให้ได้ลักษณะสำคัญของภาพ ซึ่งทำให้ประสิทธิภาพในการจำแนกข้อมูลด้วยภาพมีความแม่นมากขึ้น ขั้นตอนวิธีของการวิจัยนี้ประกอบด้วย 5 ส่วนคือ การปรับปรุงรูปภาพ การแบ่งขอบเขตภายในภาพ การดึงลักษณะสำคัญภายในภาพ การจำแนกภาพ และ การประเมินความถูกต้องของการจำแนก ดังแสดงในรูปที่ 3.1

3.1.1 ภาพแมมโมแกรม (Mammogram Images)

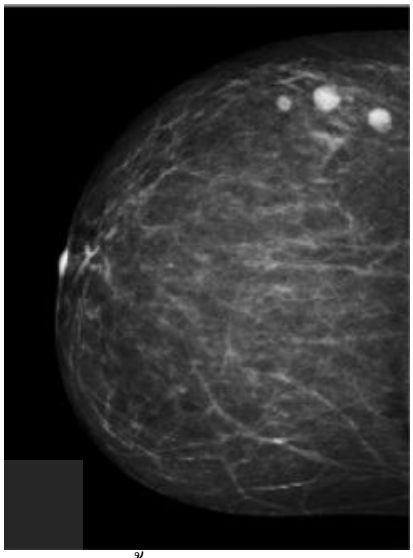
ข้อมูลที่ใช้ในการทดสอบอัลกอริทึม เป็นข้อมูลภาพจากชุดข้อมูลมาตรฐานของ University of South Florida Digital Mammography Home Page ชื่อ Digital Database for Screening Mammography (DDSM) (http://marathon.csee.usf.edu/Mammography/Database.html) ข้อมูลนี้ได้มาจากกลุ่มตัวอย่างคนไข้ 2,500 คน โดยในชุดข้อมูล DDSM มีการเก็บภาพแมมโมแกรมของเต้านมด้านซ้ายและด้านขวา ในมุมมอง 2 แบบ คือ มุมมองแบบ MLO (Mediolateral View) ซึ่งเป็นการถ่ายภาพรังสีในแนวขวางลำตัว เช่น จากด้านขวาไปด้านซ้าย และ มุมมองแบบ CC (Caudocranial View) ที่เป็นการถ่ายภาพรังสีในแนวดิ่ง จากส่วนบนลำตัวพุ่งลงด้านล่าง โดยมีทั้งภาพแมมโมแกรมของผู้ป่วยที่มีก้อนเนื้อที่ไม่เป็นอันตรายและก้อนเนื้อร้าย ดังแสดงในรูปที่ 3.2



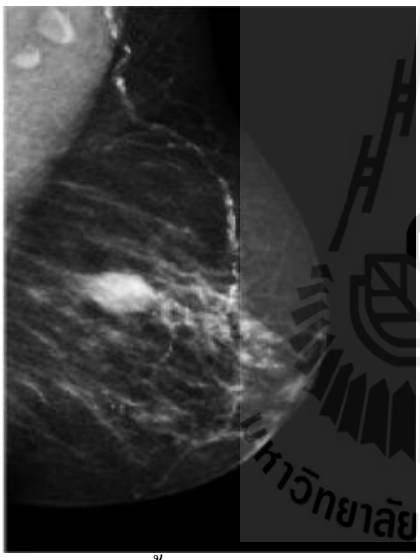
รูปที่ 3.1 ขั้นตอนวิธีในการจำแนกมะเร็งเต้านมด้วยซัพพอร์ตเวกเตอร์แมชชีน



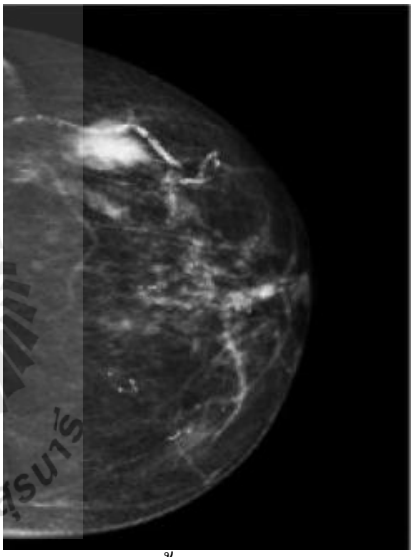
(ก) ก้อนเนื้อไม่อันตรายในมุมมอง MLO



(ข) ก้อนเนื้อไม่อันตรายในมุมมอง CC



(ค) ก้อนเนื้อมะเร็งในมุมมอง MLO



(ง) ก้อนเนื้อมะเร็งในมุมมอง CC

รูปที่ 3.2 ภาพแมมโมแกรมในมุมมอง MLO และ CC

ข้อมูลภาพที่ได้จากฐานข้อมูล DDSM นี้จะมีขนาดภาพประมาณ 3000 × 4000 พิกเซล และเป็นภาพระดับสีเทา โดยมีระดับความเข้มของสี 0 ถึง 255 โดยผู้เผยแพร่ข้อมูลได้ให้ข้อมูลของแต่ละภาพไว้ด้วย เช่น ประเภทมุมมองของภาพ คลาส ขนาดของภาพ และบริเวณขอบเขตที่เป็นก้อนเนื้อ เพื่อใช้ในการเปรียบเทียบประสิทธิภาพของอัลกอริทึม ในงานวิจัยนี้ผู้วิจัยได้คัดเลือกภาพแมมโมแกรมเพื่อใช้ในการทดสอบอัลกอริทึมจำนวน 190 ภาพ

3.1.2 การปรับปรุงภาพแมมโมแกรม (Image Enhancement)

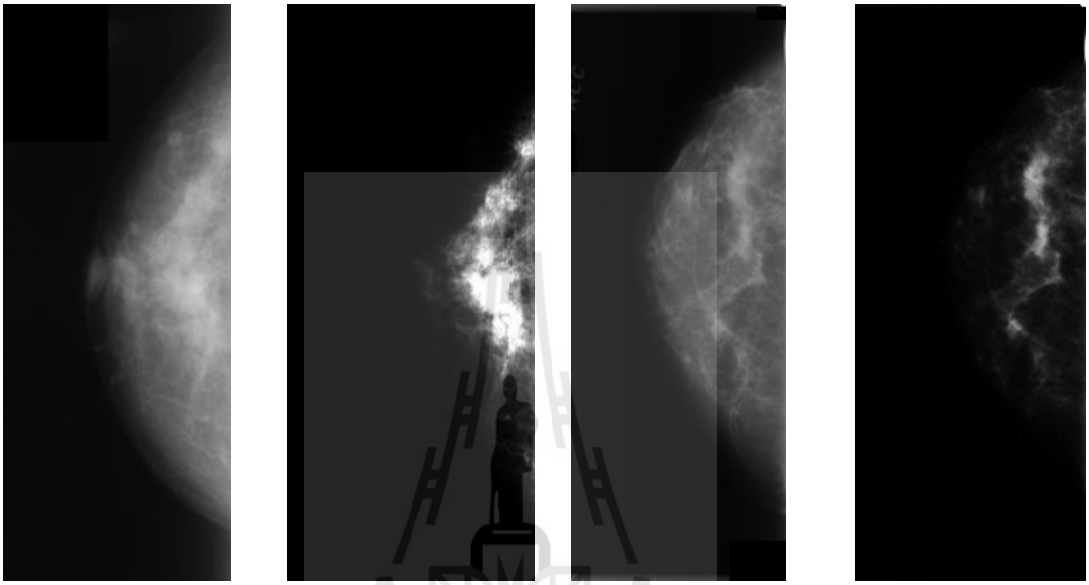
โดยปกติแล้วภาพแมมโมแกรมมักจะมีสัญญาณรบกวนภายในภาพ ทำให้ภาพไม่ชัดเจน สัญญาณรบกวนในภาพแมมโมแกรมนั้นมี 2 แบบ คือ สัญญาณรบกวนแบบเกาเซียน (Gaussian Noise) และสัญญาณรบกวนที่เป็นจุดดำหรือจุดขาวเล็ก ๆ หรือเรียกว่าสัญญาณรบกวนแบบเกลือและพริกไทย (Salt and Pepper Noise) ในขั้นตอนนี้จะเป็นการปรับปรุงภาพโดยใช้วิธีการกำจัดสัญญาณรบกวนภายในภาพด้วยวิธีมีเดียนฟิลเตอร์ (Median Filter) ซึ่งสัญญาณรบกวนทั้ง 2 แบบนี้เกิดจากคุณภาพของอุปกรณ์ในการรับภาพ หลักการของมีเดียนฟิลเตอร์คือการใช้หน้าต่างขนาดเล็กเช่น 3×3 พิกเซล หรือ 5×5 พิกเซลเลื่อนไปบนภาพที่ต้องการกำจัดสัญญาณรบกวน โดยในขณะที่เลื่อนนั้นในหน้าต่างขนาดเล็กก็ทำหน้าที่ให้การคำนวณและเปลี่ยนแปลงค่าพิกเซล ณ จุดใด ๆ โดยทำการเรียงค่าความเข้มสีของบริเวณหน้าต่างที่ครอบภาพ และทำการเรียงค่าจากน้อยไปมาก จากนั้นจึงทำการคัดเลือกค่าที่มัธยฐานแล้วนำค่ามัธยฐานแทนที่ลงในพิกเซลปัจจุบัน ดังนั้นเมื่อภาพผ่านกระบวนการของมีเดียนฟิลเตอร์แล้ว ภาพจะมีความชัดเจนขึ้นในระดับหนึ่ง โดยยังรักษาความคมชัด และขอบของภาพไว้ได้ ดังแสดงในรูปที่ 3.3

(ก) ภาพก่อนผ่านมีเดียนฟิลเตอร์ (ข) ภาพหลังผ่านการปรับปรุงโดยมีเดียนฟิลเตอร์

รูปที่ 3.3 ภาพก่อนและหลังการปรับปรุงด้วยมีเดียนฟิลเตอร์

ขั้นตอนต่อไปจะเป็นการปรับปรุงความชัดเจนของภาพ โดยเพิ่มความเข้มสีในบริเวณที่คาดว่าเป็นก้อนเนื้อเพื่อทำให้เห็นภาพบริเวณก้อนเนื้อได้ชัดเจนยิ่งขึ้น ในขณะเดียวกันก็ทำการปรับลดความเข้มสีในบริเวณที่เป็นพื้นหลังลงด้วย ในขั้นตอนนี้จะให้วิธีแก้ไขแกมมา (Gamma Correction) จากรูปที่ 3.4 แสดงภาพก้อนเนื้อในเต้านมทั้งในกรณีที่เป็นก้อนเนื้อร้ายและก้อนเนื้อไม่อันตราย โดยเมื่อเปรียบเทียบระหว่างรูป 3.4(ก) 3.4(ค) และ 3.4(ข) 3.4(ง) แล้วจะเห็นว่าการแก้ไขแกมมาช่วยให้ความเข้มสีในบริเวณที่สว่างยิ่งมีความเข้มสีที่มากขึ้น และในทางตรง

กันข้ามบริเวณที่เป็นพื้นหลังที่มีความเข้มสีที่ค่อนข้างมืดก็จะถูกปรับลดความเข้มสีลง ส่งผลให้บริเวณที่เป็นก้อนเนื้อมีความสำคัญชัดขึ้นมา โดยสอดคล้องกับที่นักรังสีวิทยาได้กล่าวไว้ว่า บริเวณที่เป็นก้อนเนื้อที่มีความผิดปกติ ความเข้มสีบริเวณนั้นจะมีมากกว่าบริเวณอื่น ๆ ซึ่งวิธีการนี้จะเป็นประโยชน์ในการนำไปเข้ากระบวนการแบ่งขอบเขตภาพหรือ ROI ต่อไป



(ก) (ข) (ค) (ง)

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

3.1.3 การแบ่งขอบเขตภาพ (Image Segmentation)

ขั้นตอนการแบ่งขอบเขตภาพนั้นเป็นขั้นตอนที่สำคัญอีกขั้นตอนหนึ่งก่อนนำภาพไปทำการจำแนก เนื่องจากภาพแมมโมแกรมมีขนาดใหญ่ประมาณ 3000×4000 พิกเซล ซึ่งบริเวณที่สนใจนั้นจะเป็นบริเวณก้อนเนื้อที่มีความผิดปกติเท่านั้น ดังนั้นพื้นหลังในภาพแมมโมแกรม และบริเวณภาพที่เป็นไขมันจึงไม่จำเป็นต้องนำเข้าสู่กระบวนการจำแนก กระบวนการนี้จะช่วยลดขนาดข้อมูล ทำให้การจำแนกทำได้รวดเร็วขึ้นและไม่เกิดปัญหาหน่วยความจำเต็ม ในขั้นตอนนี้เราจะใช้วิธีการการขยายพื้นที่ของส่วนภาพ หรือ Region Growing ซึ่งวิธีนี้เป็นวิธีที่นิยมใช้ในการแบ่งส่วนภาพภายในภาพแมมโมแกรม โดยในขั้นตอนแรกนั้นจะทำ

การคัดเลือกจุดกึ่งกลางของก้อนเนื้อ (Seed Point) โดยใช้วิธีการหาจุดเซนทรอยด์ (Centroid) จากก้อนเนื้อในภาพ โดยใช้สมการที่ (3-1) และ (3-2)

$$\text{Area} = \sum_{i=1}^{m}\sum_{j=1}^{n} W[i,j] \tag{3-1}$$

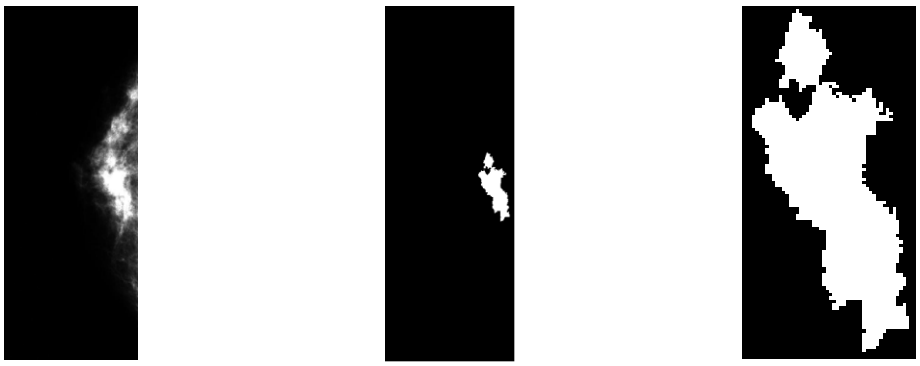
$$\text{Centroid} = (\bar{x}, \bar{y}) \tag{3-2}$$

โดยที่ $$\bar{y} = \frac{\sum_i \sum_j iW[i,j]}{\text{Area}}$$

$$\bar{x} = \frac{\sum_i \sum_j jW[i,j]}{\text{Area}}$$

จากสมการต้องทำการหาพื้นที่ (Area) ของบริเวณที่เป็นก้อนเนื้อก่อนโดย W คือจำนวนพิกเซลสีขาวทั้งหมดในภาพ และ i, j คือตำแหน่งของพิกเซลสีขาว หลังจากนั้นจึงทำการหาตำแหน่งจุดเซนทรอยด์ของก้อนเนื้อ หลังจากได้จุดเซนทรอยด์แล้วจะนำจุดเซนทรอยด์มาใช้เป็นพิกัดในการคัดเลือกจุดกึ่งกลางของก้อนเนื้อ เพื่อให้กระบวนการขยายพื้นที่ของส่วนภาพดำเนินการต่อไป

กระบวนการขยายพื้นที่ของภาพหรือ Region Growing ทำได้โดยเริ่มต้นพิจารณาจุดกึ่งกลาง เมื่อพบจุดภาพที่เป็นบริเวณขอบของ จุดกึ่งกลางหรือ Seed Region ก็จะพิจารณาจุดภาพข้างเคียง (Neighbor) ด้วยการวางหน้าต่างขนาด 3×3 ครอบจุดภาพนั้น หากจุดภาพข้างเคียงใดมีค่าระดับสีเทาอยู่ในขอบเขตของการขยายพื้นที่ ก็จะทำการรวมหรือขยายพื้นที่ส่วนภาพไปยังจุดข้างเคียงนั้น แต่ถ้าไม่ ก็จะพิจารณาจุดข้างเคียงถัดไป กระบวนการขยายส่วนพื้นที่ของภาพนี้ จะกระทำกับทุก ๆ Seed Region แบบวนซ้ำไปเรื่อย ๆ จนกระทั่งไม่สามารถขยายพื้นที่ได้ ผลจากการใช้กระบวนการขยายส่วนพื้นที่ในภาพแมมโมแกรมที่ผ่านการแก้ไขแกมมาแล้ว แสดงผลดังรูปที่ 3.5



(ก) (ข) (ค)

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

3.1.4 การดึงลักษณะสำคัญในภาพ (Image Feature Extraction)

หลังจากได้ ROI จากการทำการขยายส่วนพื้นที่ของภาพมาแล้วนั้น หากนำเพียงค่าความเข้มสีไปเข้ากระบวนการจำแนก อาจทำให้การจำแนกได้ผลไม่ดีนัก ดังนั้นการดึงลักษณะสำคัญในภาพ โดยนำคุณสมบัติของ ROI ที่เราตัดออกมาจากภาพแมมโมแกรมมาหาลักษณะสำคัญ ดังนั้นกระบวนนี้จึงเป็นวิธีการที่น่าสนใจและทำให้ประสิทธิภาพในการจำแนกดีขึ้นด้วย สำหรับในงานวิจัยนี้ได้ใช้ลักษณะสำคัญทั้งหมด 3 ลักษณะ คือ ลักษณะสำคัญของลวดลายภายในก้อนเนื้อ (Texture Feature) ลักษณะสำคัญของรูปร่างก้อนเนื้อ (Shape Feature) และ ลักษณะสำคัญของความเข้มสีของก้อนเนื้อ (Intensity Histogram Feature)

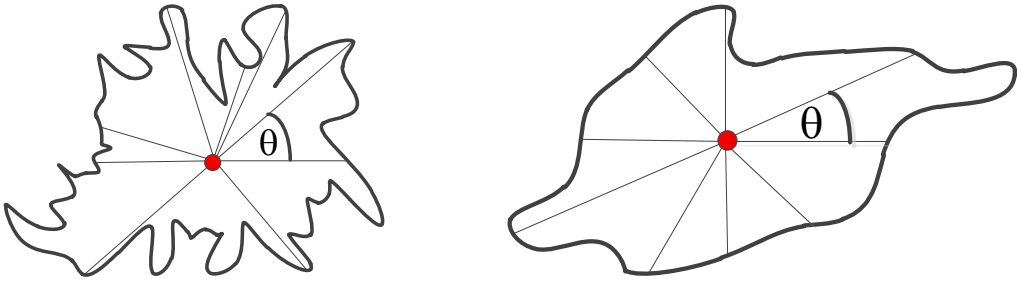
ลักษณะสำคัญของลวดลาย หรือ Texture Feature เป็นการหารูปแบบของลวดลายจาก ROI เทคนิคที่นิยมใช้ในการดึง Texture Feature จากภาพนั้นคือการใช้ Gray-Level Co-occurrence Matrix (GLCM) หรืออีกชื่อหนึ่งคือ Gray-level Spatial Dependence Matrix ฟังก์ชันใน GLCM นั้นจะทำการคำนวณและเปรียบเทียบการเกิดขึ้นของลวดลายหรือแพทเทินระหว่างพิกเซลในภาพ โดยใช้ความน่าจะเป็นในการแสดงผลของความชัดเจนของลวดลาย (Contrast) การเกิดขึ้นร่วมกันของลวดลาย (Correlation) รวมทั้งวัดค่าความเป็นเนื้อเดียวกัน (Homogeneity) ของลวดลายในภาพอีกด้วย (แสดงตัวอย่างค่าที่ได้จากการคำนวณในตารางที่ 3.1) ในงานวิทยานิพนธ์นี้จะใช้ลักษณะสำคัญของลวดลาย เหล่านี้ในการกระบวนการจำแนกมะเร็งเต้านมในภาพแมมโมแกรม

ตารางที่ 3.1 ตัวอย่างลักษณะสำคัญของลวดลายในภาพแมมโมแกรม

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

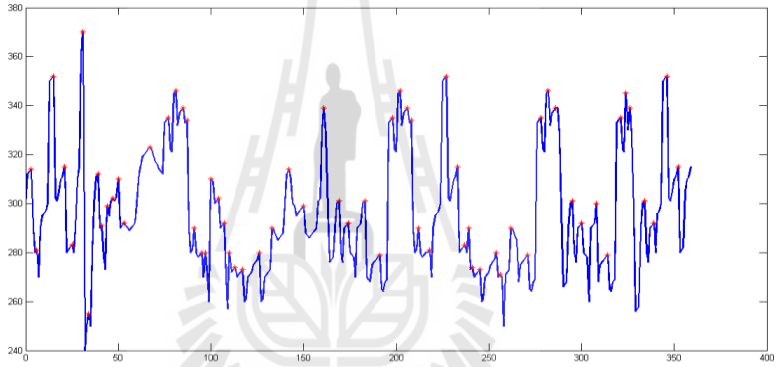
ลักษณะสำคัญของรูปร่าง หรือ Shape Feature เป็นการวัดความโค้งหรือความหยักของ ROI เนื่องจากนักรังสีวิทยาได้กล่าวว่า รูปร่างของก้อนเนื้อไม่เป็นอันตรายนั้นจะมีรูปร่างที่ค่อนข้างเป็นวงกลมหรือวงรีที่มีขอบค่อนข้างเรียบหรือมีรอยหยักน้อย และในทางกลับกันรูปร่างของเนื้อร้ายที่จะก่อตัวเป็นมะเร็งหรือเนื้อร้ายที่เป็นมะเร็งแล้วนั้นจะมีรูปร่างที่ค่อนข้างบิดเบี้ยวและขอบของก้อนเนื้อจะมีรอยหยักค่อนข้างมาก ดังนั้นการนำเอาลักษณะสำคัญของรูปร่างขอบของ ROI มาเป็นลักษณะสำคัญหรือ Feature ในการจำแนกจึงส่งผลให้โมเดลในการจำแนกมีประสิทธิภาพในการจำแนกค่อนข้างสูง

จากรูปที่ 3.6 แสดงการวัดความโค้งของ ROI ที่เป็นก้อนเนื้อไม่อันตรายและก้อนเนื้อร้าย โดยทำการโยงเส้นจากจุดเซนทรอยด์ของ ROI และแสดงออกมาเป็นกราฟแสดงความโค้งงอของเส้นขอบ ROI โดยจะสังเกตเห็นว่าก้อนเนื้อร้ายนั้นจะมีความหยักหรือการเปลี่ยนแปลงความโค้งของขอบภาพมากกว่าก้อนเนื้อไม่อันตราย และรูปที่ 3.7 แสดงความแตกต่างของกราฟแสดงความหยักของเส้นขอบแสดงรูปร่างของก้อนเนื้อร้ายและก้อนเนื้อไม่อันตราย แกน X แทนองศา ตั้งแต่ 0 องศา ถึง 360 องศา และ แกน Y แทน ระยะทางที่วัดจากจุดกึ่งกลางไปเส้นขอบ โดยมีการพิจารณาเฉพาะจำนวนจุดยอดของกราฟที่มีการเปลี่ยนแปลงให้มีค่าเป็น 1 ตามองศาการวัดระยะทางจากจุดเซนทรอยด์ไปยังเส้นขอบตั้งแต่ 0 ถึง 360 องศา ลักษณะของกราฟในรูปที่ 3.7 จะเป็น Background Knowledge ที่สำคัญในการจำแนกภาพมะเร็งเต้านม



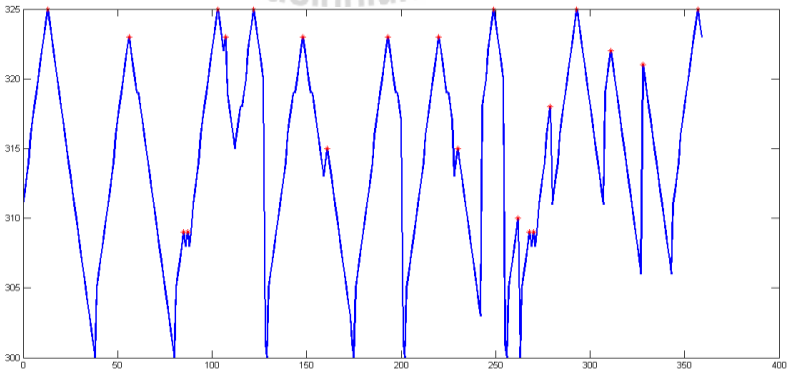
(ก) (ข)

รูปที่ 3.6 การวัดความหยักของเส้นขอบโดยวัดจากจุดเซนทรอยด์ (ก) เส้นขอบแสดงรูปร่างของก้อนเนื้อร้าย (ข) เส้นขอบแสดงรูปร่างของก้อนเนื้อไม่อันตราย



……..1001001100101010010001101110001110110001111……..

(ก)

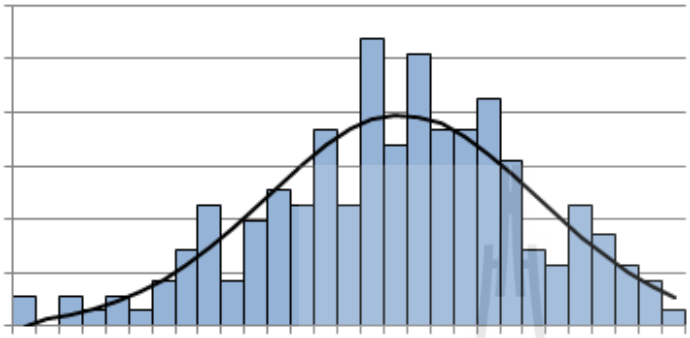


……..1000000000001000000000001000001000000000100……..

(ข)

รูปที่ 3.7 กราฟแสดงความหยักของ (ก) ก้อนเนื้อร้าย (ข) ก้อนเนื้อไม่อันตราย

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



รูปที่ 3.8 ตัวอย่างกราฟฮิสโตแกรมที่พิจารณาลักษณะสำคัญ 4 ค่า

3.1.5 การจำแนกมะเร็งเต้านม (Classification)

ในงานวิจัยนี้ทางผู้วิจัยได้ใช้ซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกโดยจะทดลองใช้เคอร์เนลฟังก์ชันหลายแบบ

สำหรับลักษณะสำคัญที่ใช้เป็นข้อมูลในการจำแนกนั้นจะใช้ลักษณะสำคัญ 3 แบบที่ได้จากหัวข้อ 3.1.4 ได้แก่ ลักษณะสำคัญของลวดลาย ลักษณะสำคัญของรูปร่าง และลักษณะสำคัญของความเข้มสีจากฮิสโตแกรม ในงานวิจัยครั้งนี้ผู้วิจัยได้ใช้ภาพจาก DDSM สำหรับการประมวลผลภาพก่อนนำมาจำแนกนั้น ภาพในขั้นตอนการสอนและการทดสอบทั้งหมดต้องผ่านการประมวลผลภาพในแบบเดียวกันคือ การกำจัดสัญญาณรบกวนภาพ การขยายส่วนพื้นที่ภาพ และการดึงลักษณะสำคัญ หลังจากขั้นตอนการจำแนกแล้ว ผู้วิจัยยังได้นำการจำแนกอีก 2 วิธีคือ นาอีฟเบย์ และ โครงข่ายประสาทเทียมมาเปรียบเทียบกับอัลกอริทึมที่นำเสนออีกด้วย

3.2 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการพัฒนางานวิจัยนี้ ประกอบด้วย

1) เครื่องคอมพิวเตอร์สำหรับการพัฒนา มีรายละเอียดดังนี้

- หน่วยประมวลผลกลาง : Intel Core i5
- หน่วยความจำหลัก : 4 GB

- หน่วยความจำสำรอง : 320 GB

2) ระบบปฏิบัติการและโปรแกรมประยุกต์สำหรับการพัฒนา ประกอบไปด้วย
   - ระบบปฏิบัติการ : Windows 7 Professional 64 bits
   - เครื่องมือที่ใช้ในการพัฒนา : R Language, Matlab2013

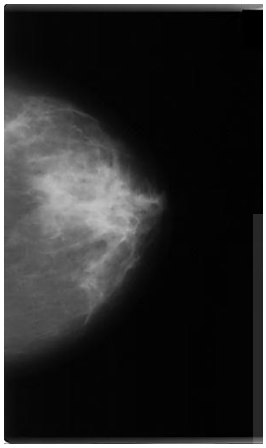
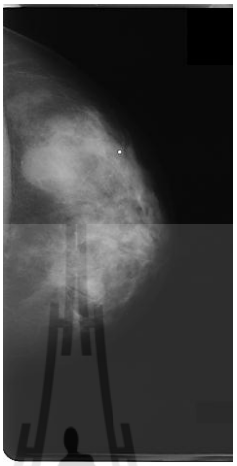
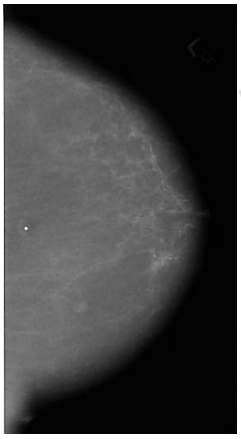
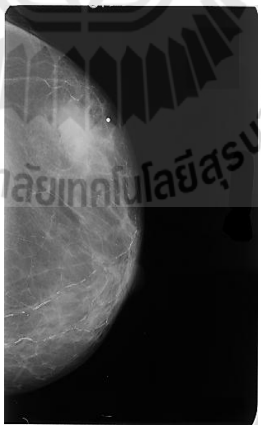
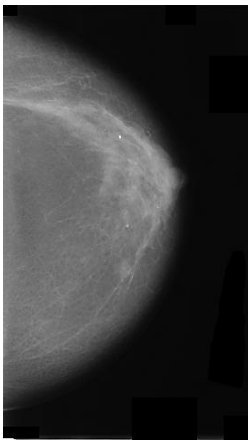
บทที่ 4
การทดสอบและอภิปรายผล

การทดสอบประสิทธิภาพของระบบนั้น จะทดสอบประสิทธิภาพความแม่น ค่า Sensitivity ค่า Specificity และ พื้นที่ใต้กราฟ ROC ในการจำแนกภาพแมมโมแกรมระหว่างก้อนเนื้ออันตรายและก้อนเนื้อไม่อันตราย โดยเปรียบเทียบกับอัลกอริทึมในการจำแนกอีก 2 อัลกอริทึม คือ โครงข่ายประสาทเทียม และ นาอีฟเบย์ สำหรับเนื้อหาในบทนี้จะประกอบด้วย ข้อมูลที่ใช้ในการทดสอบ การทดสอบประสิทธิภาพการจำแนกภาพแมมโมแกรม โดยแบ่งเป็นการทดสอบประสิทธิภาพการจำแนกด้วยค่าความแม่น ค่า Sensitivity ค่า Specificity และ พื้นที่ใต้กราฟ ROC และในหัวข้อสุดท้ายเป็นการอภิปรายผล

4.1 ข้อมูลที่ใช้ในการทดสอบ

การทดสอบการจำแนกภาพแมมโมแกรมด้วยการประมวลผลภาพร่วมกับซัพพอร์ตเวกเตอร์แมชชีนจะใช้ข้อมูลมาตรฐานภาพแมมโมแกรม (Digital Database for Screening Mammograpy: DDSM) จากเว็บไซต์ของมหาวิทยาลัยเซาท์ฟลอริดา (University of South Florida) ซึ่งเป็นข้อมูลภาพระดับสีเทา (Grey Scale Image) มีข้อมูลจากคนไข้ทั้งหมด 2,500 ข้อมูล โดยมีข้อมูลของคนไข้ที่มีก้อนเนื้ออันตรายและก้อนเนื้อไม่อันตราย ซึ่งประกอบด้วยข้อมูลภาพแมมโมแกรมใน 2 มุมมอง คือ ภาพแมมโมแกรมในมุมมองแบบ MLO และ ภาพแมมโมแกรมในมุมมองแบบ CC โดยในการทดสอบนั้นจะคัดเลือกเฉพาะภาพแมมโมแกรมในมุมมองแบบ CC มาทั้งหมด 190 ภาพ เนื่องจากภาพในมุมมองแบบ CC ไม่มีส่วนพื้นที่สีขาวในมุมบนด้านซ้ายและขวาทำให้สะดวกต่อการประมวลผลภาพ ภาพในมุมมองแบบ CC ที่เลือกมานั้นมีคลาสเป้าหมายสองกลุ่ม คือ Malignant และ Benignโดยแต่ละภาพจะมีความกว้างประมาณ 3,000 พิกเซล และความสูงประมาณ 5,000 พิกเซล สามารถดาวน์โหลดภาพแมมโมแกรมได้ที่เว็บไซต์ http://marathon.csee.usf.edu/Mammography/Database.html โดยรายละเอียดตัวอย่างข้อมูลแสดงดังตารางที่ 4.1

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
		

เนื่องจากภาพแมมโมแกรมที่ใช้มีค่อนข้างขนาดใหญ่ ดังนั้นจึงต้องผ่านกระบวนการการประมวลผลภาพก่อนด้วยโปรแกรม MATLAB R2013b โดยการผ่านกระบวนการ มีเดียนฟิลเตอร์ แกมมาคอเรคชัน และการขยายพื้นที่ของภาพ เพื่อคัดเลือกเฉพาะบริเวณภาพก้อนเนื้อที่สนใจ หลังจากนั้นจึงทำการดึงลักษณะสำคัญของภาพ 3 ประเภท ออกมาเป็นข้อมูลตัวเลขดังแสดงวิธีการ

ดำเนินการในบทที่ 3 โดยประกอบด้วยข้อมูลจำนวน 21 คอลัมน์ โดยคอลัมน์ที่ 1 –15 เป็นลักษณะสำคัญของลวดลาย คอลัมน์ที่ 16 – 19 เป็นลักษณะสำคัญของกราฟฮิสโตแกรม และ คอลัมน์ที่ 20 เป็นลักณษะสำคัญของรูปร่าง และคอลัมน์ที่ 21 เป็นหมายเลขคลาส ดังแสดงในตารางที่ 4.2 และตารางที่ 4.3 นั้น แสดงตัวอย่างข้อมูลลักษณะสำคัญของภาพแมมโมแกรมจำนวน 16 ตัวอย่าง

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 ตัวอย่างข้อมูลลักษณะสำคัญของภาพแมมโมแกรมจำนวน 16 ตัวอย่าง

Cont0	Cont45	Cont90	Cont135	Homo0	Homo45	Homo90	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

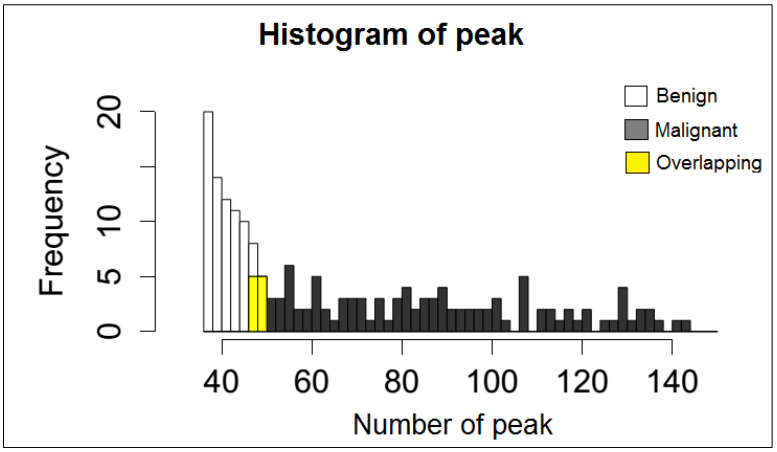
Corr135	Avg_Cont	Avg_Homo	Avg_Corr	Hist_Avg	Hist_Var	Hist_Skew	Hist_Kur	Peak_No	Class_No
0.91654	0.17814	0.91093	0.16738	105.98	1026.5	-0.54017	1.7046	41	0
0.91586	0.18032	0.90984	0.14649	119.4	1102	-0.62523	1.8918	42	0
0.91432	0.18241	0.9088	0.17577	98.062	1247.7	-0.78261	2.349	41	0
0.9177	0.17646	0.91177	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 โดยค่าตัวเลขเหล่านี้ได้มาจากวิธีการหาลักษณะสำคัญของรูปร่างในหัวข้อที่ 2.3.3 ซึ่งเป็นการวัดระยะห่างจากจุดเซนทรอยด์ไปยังเส้นขอบของรูปภาพ หลังจากนั้นจึงทำการนับจุดเปลี่ยนโค้งของเส้นขอบแสดงรูปร่างออกมาเป็นตัวเลขดังแสดงในตารางที่ 4.3 เนื่องจากลักษณะสำคัญของรูปร่างในคอลัมน์ Peak_No เมื่อนำไปเข้ากระบวนการจำแนกร่วมกับลักษณะสำคัญอื่น ๆ แล้วนั้น ประสิทธิภาพการจำแนกยังไม่ดีเท่าที่ควร ดังนั้นจึงทำการเพิ่มเติมลักษณะสำคัญของรูปร่างโดยทำการหาค่า Threshold ที่เหมาะสมของข้อมูลในคอลัมน์ Peak_No ระหว่างก้อนเนื้อไม่อันตรายกับก้อนเนื้ออันตราย

วิธีการหาค่า Threshold ที่เหมาะสมทำได้ดังนี้

1) นำข้อมูลจากคอลัมน์ Peak_No จากคลาส Benign มาทำการพล็อตกราฟฮิสโตแกรม
2) นำข้อมูลจากคอลัมน์ Peak_No จากคลาส Malignant มาทำการพล็อตกราฟฮิสโตแกรม
3) ทำการหาค่า Threshold ที่เหมาะสมโดยดูจากบริเวณที่กราฟฮิสโตแกรมทั้งสองฝั่งมีการซ้อนทับกัน (Overlapping) ดังแสดงในรูปที่ 4.1
4) เมื่อได้ค่า Threshold แล้ว นำค่า Threshold ไปเป็นค่าในการเพิ่มชุดข้อมูลลักษณะสำคัญของรูปร่าง โดยทำการพิจารณาจากคอลัมน์ Peak_no เทียบกับค่า Threshold ที่หาได้จากการพล็อตกราฟฮิสโตแกรม 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 คอลัมน์ ดังแสดงในรูปที่ 4.2



รูปที่ 4.1 กราฟฮิสโตแกรมแสดงความถี่ของจุดพีคระหว่างก้อนเนื้ออันตราย และก้อนเนื้อไม่อันตราย

Threshold = 50

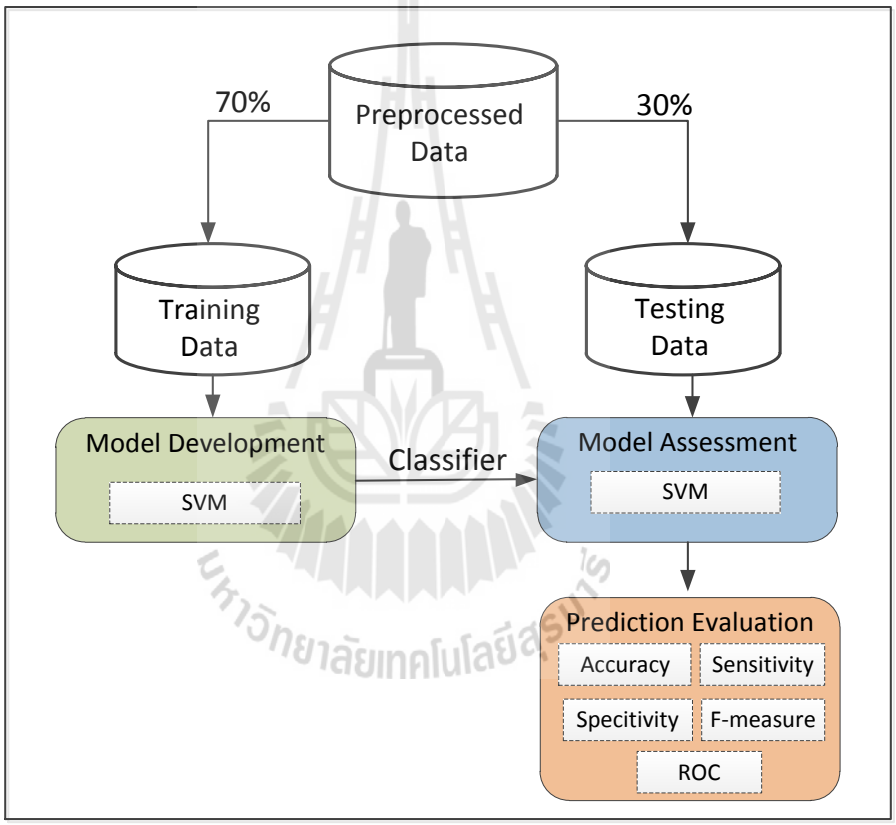
Peak_no	1	2	3	4	.....	17	18	19	20
48	1	1	1	1	.....	1	1	1	1
39	1	1	1	1	.....	1	1	1	1
42	1	1	1	1	.....	1	1	1	1
50	100	100	100	100	.....	100	100	100	100
63	100	100	100	100	.....	100	100	100	100
114	100	100	100	100	.....	100	100	100	100

รูปที่ 4.2 ภาพตัวอย่างแสดงการเพิ่มชุดข้อมูลจากการพิจารณาฮิสโตแกรมลักษณะสำคัญของรูปร่าง

4.3 การทดสอบประสิทธิภาพการจำแนกภาพแมมโมแกรม

การทดสอบประสิทธิภาพการจำแนกนี้จะทำการเปรียบเทียบประสิทธิภาพความแม่น (Accuracy) ค่า Sesitivity ค่า Specificity ค่า F-measure และพื้นที่ใต้กราฟ ROC ในการจำแนกข้อมูลภาพแมมโมแกรมของอัลกอริทึม 3 แบบ ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (เคอร์เนลฟังก์ชันเรเดียลเบสิส) โครงข่ายประสาทเทียม และ นาอีฟเบย์ ข้อมูลที่ใช้ทดสอบแบ่งเป็น ข้อมูลที่ใช้ในการฝึกสอน (Train Data) จำนวน 133 ข้อมูล (70% จาก 190 ภาพ) โดยแบ่งเป็นภาพก้อนเนื้ออันตราย

จำนวน 77 ภาพ (คลาส Malignant) และก้อนเนื้อไม่อันตรายจำนวน 56 ภาพ (คลาส Benign) และข้อมูลที่ใช้ทดสอบ (Test Data) จำนวน 57 ข้อมูล (30% จาก 190 ภาพ) โดยแบ่งเป็นภาพก้อนเนื้ออันตรายจำนวน 33 ภาพ และก้อนเนื้อไม่อันตรายจำนวน 24 ภาพ โดยนำข้อมูล 70% ผ่านขั้นตอนการฝึกสอน หลังจากนำข้อมูลผ่านการฝึกสอนแล้วจะได้โมเดลการจำแนกเพื่อนำไปจำแนกข้อมูลทดสอบจำนวน 30% หลังจากนั้นจะทำการประเมินประสิทธิภาพของระบบด้วยการประเมินประสิทธิภาพ 5 แบบ ดังรูปที่ 4.1 แสดงแผนภาพวิธีการทดสอบประสิทธิภาพการจำแนกภาพแมมโมแกรม



รูปที่ 4.3 แผนภาพวิธีการทดสอบประสิทธิภาพโมเดลสำหรับการจำแนกข้อมูลภาพแมมโมแกรม

4.4 ผลการทดสอบประสิทธิภาพ

สำหรับการทดสอบประสิทธิภาพของระบบนั้น จะใช้ข้อมูลลักษณะสำคัญของภาพ 39 คอลัมน์ (ลักษณะสำคัญของลวดลายจำนวน 15 คอลัมน์ ลักษณะสำคัญของกราฟฮิสโตแกรมจำนวน 4 คอลัมน์ และลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล 20 คอลัมน์) โดยทำการสุ่ม

ข้อมูลทดสอบจำนวน 57 ข้อมูลจากทั้ง 2 คลาส รายละเอียดการทดสอบประสิทธิภาพจะแบ่งออกเป็น 2 กรณี ดังต่อไปนี้

4.4.1 ผลการทดลองการเปรียบเทียบประสิทธิภาพความแม่นระหว่างซัพพอร์ตเวกเตอร์แมชชีน โครงข่ายประสาทเทียม และ นาอีฟเบย์

จากผลการทดลองได้ผลลัพธ์ดังต่อไปนี้ ตารางที่ 4.4 4.5 และ 4.6 แสดงตาราง Confusion Matrix ในการคำนวณหาค่าความแม่น ค่า Sensitivity ค่า Specificity และค่า F-measureในการจำแนกภาพแมมโมแกรมด้วยซัพพอร์ตเวกเตอร์แมชชีน โครงข่ายประสาทเทียม และนาอีฟเบย์ ตามลำดับ โดยใช้ลักษณะสำคัญทั้ง 3 แบบ คือ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม (ADSF,Texture, Histogram หรือ ADSF-TH) เป็นข้อมูลนำเข้า ผลของการทดสอบคือ การจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีนได้ผลความแม่นที่สูงสุดคือ 92.98% และยังให้ค่า Sensitivity ค่า Specificity และค่า F-measure สูงที่สุดอีกด้วยคือ 93.94% 91.67% และ 93.94% ตามลำดับ รองลงมาคือการจำแนกด้วยโครงข่ายประสาทเทียมให้ค่าความแม่น ค่า Sensitivity ค่า Specificity และค่า F-measure คือ 89.47% 90.91% 87.50% และ90.91% ตามลำดับ และสำหรับการจำแนกด้วยนาอีฟเบย์ ให้ค่าความแม่น ค่า Sensitivity ค่า Specificity และค่า F-measure คือ 82.45% 87.10% 79.92% และ 84.38% ตามลำดับ

ตารางที่ 4.4 Confusion Matrix ของการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญขอ'กราฟฮิสโตแกรม เป็นข้อมูลนำเข้า

<table>
<tr><td colspan="2" rowspan="2">SVM(ADSF-TH)</td><td colspan="2">ค่าความจริง (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 colspan="2">Precision = 31 / (31+2) = 0.9394</td><td>Sensitivity = (31/(31+2))*100 = 93.94%</td><td>Specificity = (22/(2+22))*100 = 91.67%</td><td>F-measure = (2×0.9394×0.93944) / (0.9394+0.9394) = 0.9394</td></tr>
</table>

ตารางที่ 4.5 Confusion Matrix ของการจำแนกด้วยโครงข่ายประสาทเทียม โดยใช้ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม เป็นข้อมูลนำเข้า

ANN(ADSF-TH)		ค่าความจริง (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 Confusion Matrix ของการจำแนกด้วยนาอีฟเบย์ โดยใช้ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม เป็นข้อมูลนำเข้า

Naïve(ADSF-TH)		**ค่าความจริง (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**

ตารางที่ 4.7 4.8 และ 4.9 แสดง Confusion Matrix ในการคำนวณหาค่าความแม่น ค่า Sensitivity ค่า Specificity และค่า F-measure ในการจำแนกภาพแมมโมแกรมด้วยซัพพอร์ตเวกเตอร์แมชชีน โครงข่ายประสาทเทียม และนาอีฟเบย์ ตามลำดับ โดยใช้ลักษณะสำคัญของรูปร่างแบบไม่เพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม (Shape, Texture, Histogram หรือ STH) เป็นข้อมูลนำเข้า ผลของการทดสอบคือ การจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีนได้ผลความแม่นคือ 87.27% ค่า Sensitivity ค่า Specificity และค่า F-measure คือ 90.63% 84.00% และ89.23% ตามลำดับ การจำแนกด้วยโครงข่ายประสาทเทียมให้ค่าความแม่น ค่า Sensitivity ค่า Specificity และค่า F-measure คือ 84.21% 87.50% 80.00% และ86.15% ตามลำดับ และสำหรับการจำแนกด้วยนาอีฟเบย์ ให้ค่าความแม่น ค่า Sensitivity ค่า Specificity และค่า F-measure คือ 78.95% 83.87% 73.08% และ 81.25% ตามลำดับ

ตารางที่ 4.7 Confusion Matrix ของการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ลักษณะสำคัญของรูปร่างแบบไม่เพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม เป็นข้อมูลนำเข้า

<table>
<tr><td colspan="2" rowspan="2">SVM(STH)</td><td colspan="2">ค่าความจริง (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) = 29</td><td>False Positive (FP) = 4</td><td rowspan="2">Accuracy = ((29+21) / (29+21+4+3))*100 = 87.72%</td></tr>
<tr><td>Negative</td><td>False Negative (FN) = 3</td><td>True Negative (TN) = 21</td></tr>
<tr><td colspan="2">Precision
= 29 / (29+4)
= 0.8788</td><td>Sensitivity
=(29/(29+3))*100
= 90.63%</td><td>Specificity
=(21/(4+21))*100
= 84.00%</td><td>F-measure
= (2×0.8788×0.9063) / (0.8788+0.9063)
= 0.8923</td></tr>
</table>

ตารางที่ 4.8 Confusion Matrix ของการจำแนกด้วยโครงข่ายประสาทเทียม โดยใช้ลักษณะสำคัญของรูปร่างแบบไม่เพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม เป็นข้อมูลนำเข้า

ANN(STH)		ค่าความจริง (Actual)		
		Positive	Negative	
ค่าทำนาย (Predict)	Positive	True Positive (TP) = 28	False Positive (FP) = 5	**Accuracy = ((28+20) / (28+20+5+4))*100 = 84.21%**
	Negative	False Negative (FN) = 4	True Negative (TN) = 20	
Precision = 28 / (28+5) = 0.8485		**Sensitivity = (28/(28+4))*100 = 87.50%**	**Specificity =(20/(5+20))*100 = 80.00%**	**F-measure = (2×0.8485×0.8750) / (0.8485+0.8750) = 0.8615**

ตารางที่ 4.9 Confusion Matrix ของการจำแนกด้วยนาอีฟเบย์ โดยใช้ลักษณะสำคัญของรูปร่างแบบไม่เพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม เป็นข้อมูลนำเข้า

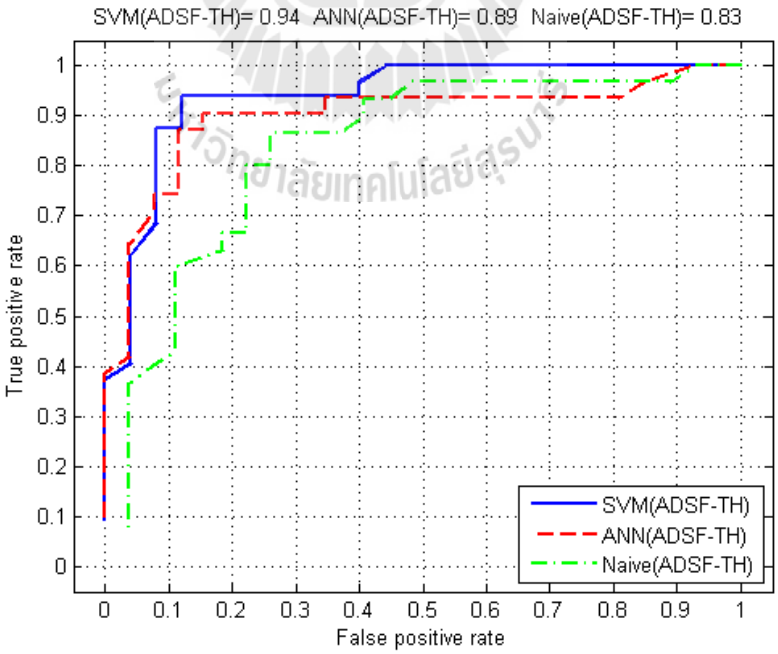
Naïve(STH)		**ค่าความจริง (Actual)**		
		Positive	Negative	
ค่าทำนาย (Predict)	Positive	True Positive (TP) = 26	False Positive (FP) = 7	**Accuracy = ((26+19) / (26+19+7+5))*100 = 78.95%**
	Negative	False Negative (FN) = 5	True Negative (TN) = 19	
Precision = 26 / (26+7) = 0.7879		**Sensitivity = (25/(25+7))*100 =83.87%**	**Specificity =(17/(8+17))*100 = 73.08%**	**F-measure = (2×0.7879×0.8387) / (0.7879+0.8387) = 0.8125**

4.4.2 ผลการทดลองการเปรียบเทียบพื้นที่ใต้กราฟ ROC ระหว่างอัลกอริทึม 3 แบบ

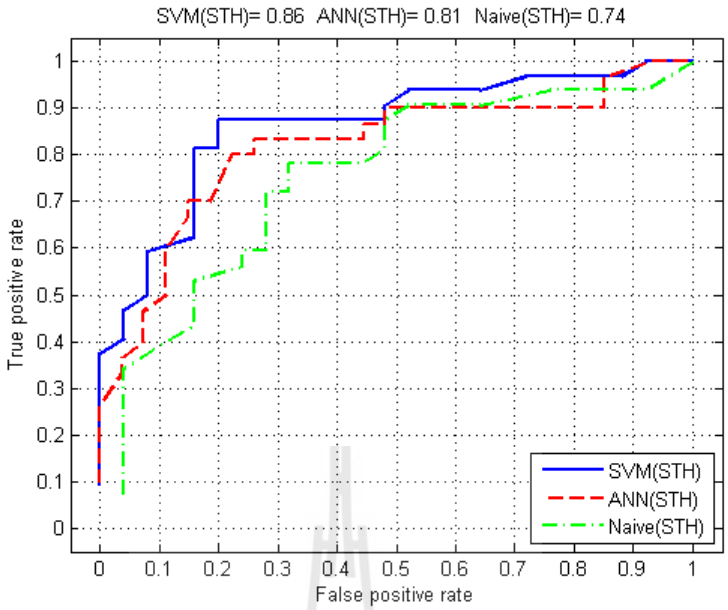
จากการทดลองได้ผลลัพธ์ดังต่อไปนี้ สำหรับการใช้ข้อมูลลักษณะสำคัญ แบบ ADSF-TH คือ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน เป็นอัลกอริทึมที่ให้ค่าพื้นที่ใต้กราฟ ROC มากที่สุด คือ 0.94 รองลงมาเป็นอัลกอริทึมโครงข่ายประสาทเทียมให้ค่าพื้นที่ใต้กราฟ คือ 0.89 และลำดับสุดท้ายคืออัลกอริทึมนาอีฟเบย์ ให้ค่าพื้นที่ใต้กราฟ คือ 0.83 ดังแสดงในรูปที่ 4.4

สำหรับการใช้ข้อมูลลักษณะสำคัญแบบ STH คือ ลักษณะสำคัญของรูปร่างแบบไม่เพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าพื้นที่ใต้กราฟ ROC มากที่สุด คือ 0.86 รองลงมาเป็นอัลกอริทึมโครงข่ายประสาทเทียมให้ค่าพื้นที่ใต้กราฟ คือ 0.81 และลำดับสุดท้ายคืออัลกอริทึมนาอีฟเบย์ ให้ค่าพื้นที่ใต้กราฟ คือ 0.74 ดังแสดงในรูปที่ 4.5

ดังนั้นจึงสรุปได้ว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพในการจำแนกดีที่สุดคือได้พื้นที่ใต้กราฟ ROC มากที่สุด และให้ค่า False Positive Rate ต่ำสุดซึ่งส่งผลให้กราฟ ROC เข้าชิดมุมซ้ายบนมากที่สุดจึงทำให้มีพื้นที่ใต้กราฟมาก



รูปที่ 4.4 พื้นที่ใต้กราฟ ROC โดยใช้ลักษณะสำคัญแบบ ADSF-TH

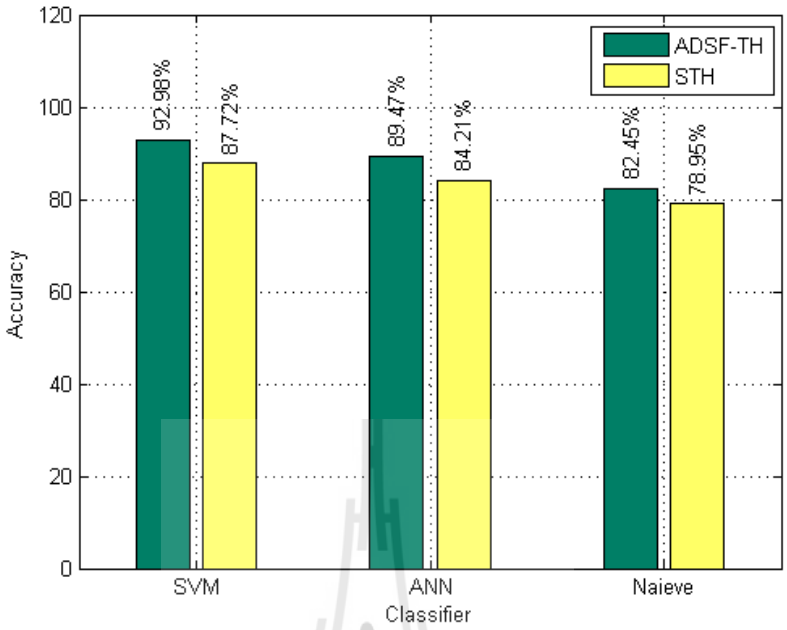


รูปที่ 4.5 พื้นที่ใต้กราฟ ROC โดยใช้ลักษณะสำคัญแบบ TSH

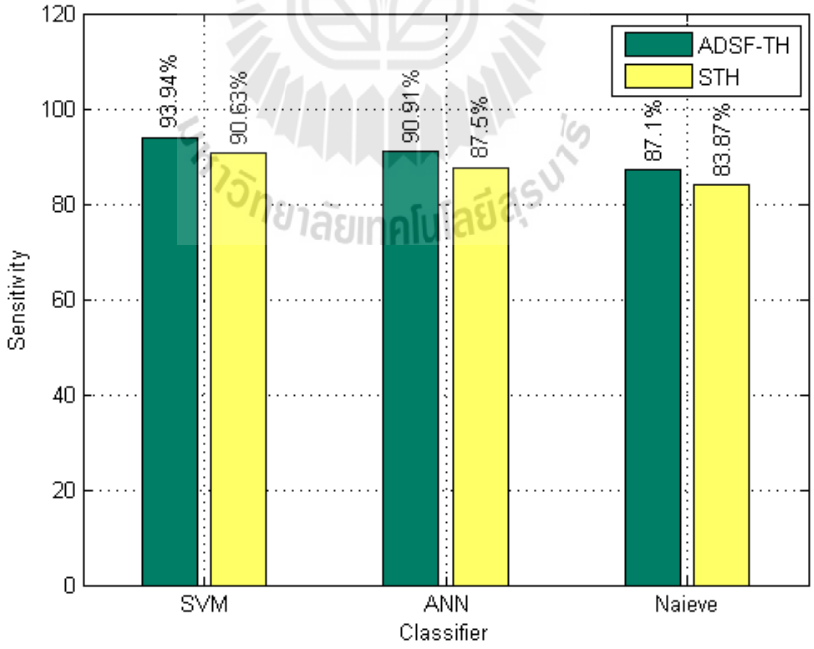
สำหรับตารางที่ 4.10 เป็นการแสดงการเปรียบเทียบค่าความแม่น ค่า Sensitivity ค่า Specificity ค่า F-measure และ พื้นที่ใต้กราฟ ROC ระหว่าง 3 อัลกอริทึม จากค่าที่แสดงในตารางจะให้เห็นว่าการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ข้อมูลจากลักษณะสำคัญแบบ ADSF-TH ให้ค่าสูงที่สุดในทุกด้าน และในรูปที่ 4.6 4.7 4.8 และ 4.9 แสดงกราฟเปรียบเทียบประสิทธิภาพการจำแนกโดยแสดงค่าความแม่น ค่า Sensitivity ค่า Specificity และ ค่า F-measure ของทั้ง 3 อัลกอริทึม

ตารางที่ 4.10 เปรียบเทียบค่า Accuracy Sensitivity Specificity F-measure และ AUC

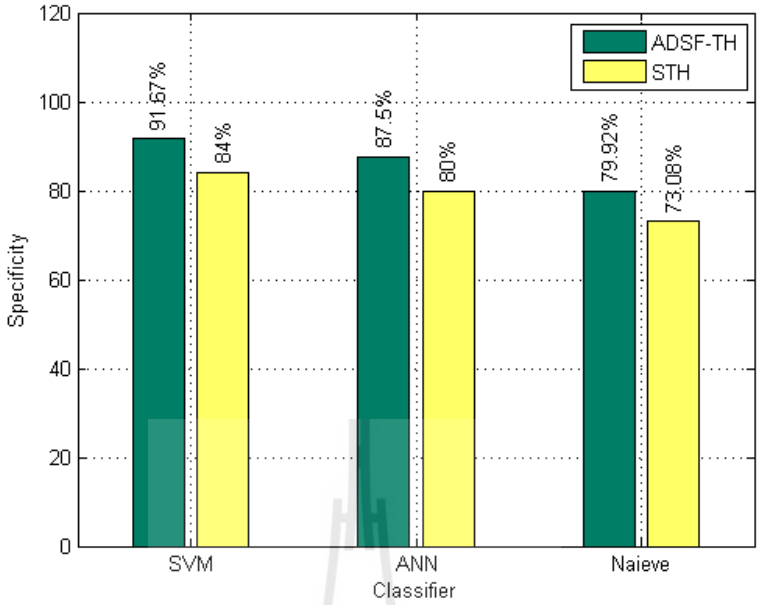
Algorithm	**Features**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F-measure (%)**	**AUC**
SVM	**ADSF-TH**	**92.98%**	**93.94%**	**91.67%**	**93.94%**	**0.94**
	STH	87.72%	90.63%	84.00%	89.23%	0.86
ANN	ADSF-TH	89.47%	90.91%	87.50%	90.91%	0.89
	STH	84.21%	87.50%	80.00%	86.15%	0.81
Naïve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



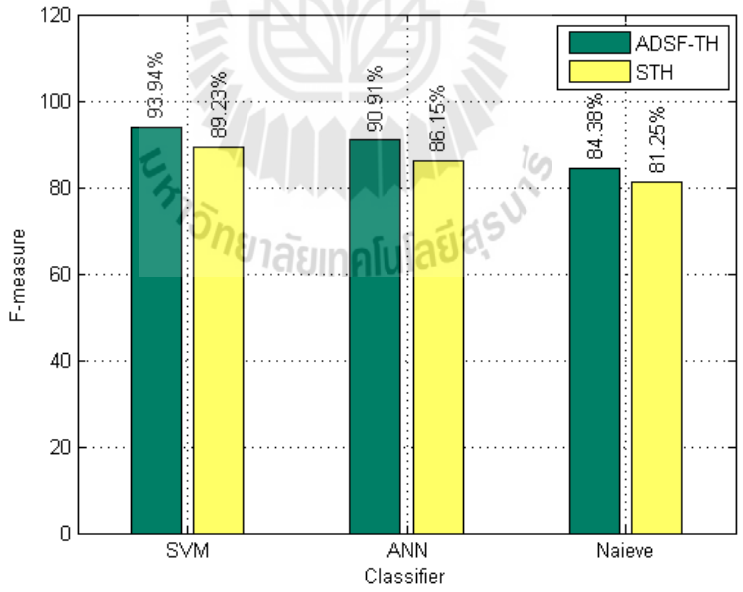
รูปที่ 4.6 กราฟเปรียบเทียบประสิทธิภาพความแม่นในการจำแนก 3 แบบ



รูปที่ 4.7 กราฟเปรียบเทียบค่า Sensitivity ในการจำแนก 3 แบบ



รูปที่ 4.8 กราฟเปรียบเทียบค่า Specificity ในการจำแนก 3 แบบ



รูปที่ 4.9 กราฟเปรียบเทียบค่า F-measure ในการจำแนก 3 แบบ

4.5 อภิปรายผล

จากผลการทดสอบประสิทธิภาพการจำแนกภาพแมมโมแกรมโดยใช้การประมวลผลภาพร่วมกับซัพพอร์ตเวกเตอร์แมชชีน ได้ทำการทดสอบกับข้อมูลภาพแมมโมแกรมจำนวน 190 ภาพ โดยประกอบด้วยข้อมูลภาพ 2 คลาส คือ ก้อนเนื้ออันตราย ก้อนเนื้อไม่อันตราย กระบวนการประมวลผลภาพได้ถูกนำมาใช้ เพื่อทำการดึงเฉพาะลักษณะสำคัญของภาพ ก่อนนำข้อมูลลักษณะสำคัญไปเข้ากระบวนการจำแนก และประเมินประสิทธิภาพ สามารถสรุปผลการทดสอบเปรียบเทียบได้ดังนี้

1) การประมวลภาพ โดยใช้วิธีการกำจัดสัญญาณรบกวนในภาพด้วยตัวกรองมัธยฐาน การแก้ไขค่าแกมมา และ การขยายส่วนของพื้นที่ มีผลทำให้ข้อมูลภาพมีขนาดเล็กลง และทำให้สามารถดึงลักษณะสำคัญของภาพออกมาได้ง่ายขึ้น เพื่อลดมิติข้อมูลในการนำเข้ากระบวนการจำแนก เนื่องจากข้อมูลความเข้มสีของภาพเป็นข้อมูลที่มีขนาดมิติใหญ่มาก ซึ่งส่งผลต่อประสิทธิภาพในการจำแนกโดยตรง

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

3) การจำแนกภาพแมมโมแกรมด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนโดยใช้เคอร์เนลฟังก์ชัน เรเดียลเบสิส โดยใช้ลักษณะสำคัญทั้ง 3 แบบ คือ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และ ลักษณะสำคัญของฮิสโตแกรม ให้ประสิทธิภาพในการจำแนกดีที่สุด เมื่อเปรียบเทียบกับอีก 2 อัลกอริทึม คือ โครงข่ายประสาทเทียม และ นาอีฟเบย์ โดยการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีนได้ค่าความแม่น ค่า Sesitivity ค่า Specificity ค่า F-measure และ พื้นที่ใต้กราฟ ROC คือ92.98% 93.94% 91.67% และ93.94% ตามลำดับ

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

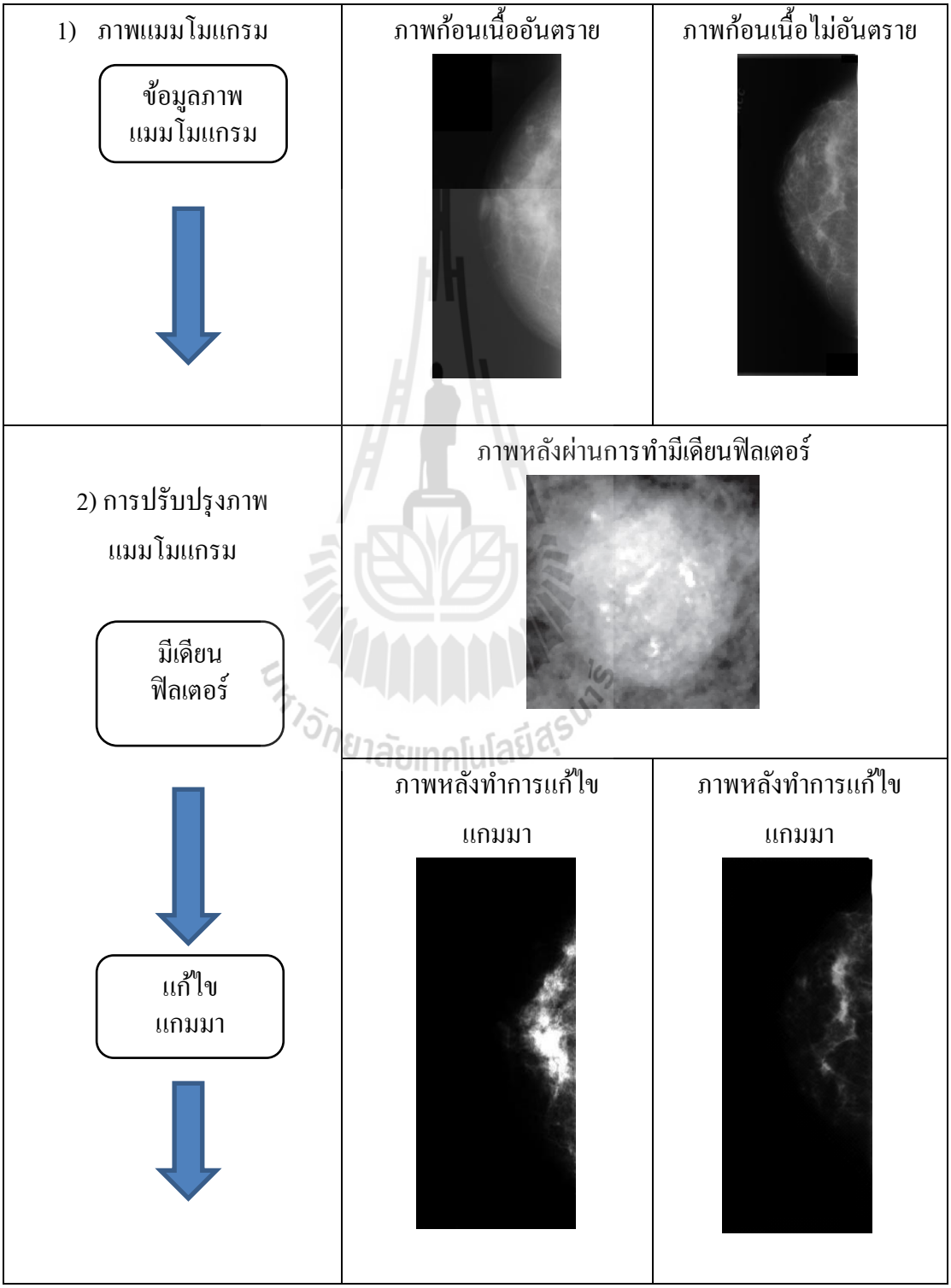
การจำแนกมะเร็งเต้านมจากภาพแมมโมแกรม มีวัตถุประสงค์เพื่อทำการจำแนกก้อนเนื้อภายในภาพแมมโมแกรมว่าเป็นก้อนเนื้อไม่อันตรายหรือก้อนเนื้ออันตราย เพื่อประโยชน์ในการช่วยนักรังสีวิทยาวินิจฉัยโรคมะเร็งเต้านม และยังช่วยทำให้ผู้ป่วยได้รู้ผลการวินิจฉัยเบื้องต้นจากภาพแมมโมแกรมโดยไม่จำเป็นต้องมีความเสี่ยงจากการผ่าตัดเพื่อนำชิ้นเนื้อในเต้านมไปตรวจสอบ

ในปัจจุบันมีนักวิจัยจำนวนมากพัฒนาประสิทธิภาพของการจำแนกภาพแมมโมแกรมโดยใช้เทคนิควิธีต่าง ๆ ของการประมวลผลภาพร่วมกับเทคนิควิธีการเรียนรู้ของเครื่อง เพื่อเพิ่มความแม่นในการจำแนกมะเร็งเต้านมจากภาพแมมโมแกรม การปรับปรุงภาพก่อนการนำไปจำแนกเป็นขั้นตอนที่สำคัญเนื่องจากภาพแมมโมแกรมอาจมีความไม่ชัดเจนหรือมีสัญญาณรบกวนในภาพ ทำให้การจำแนกได้ผลที่ไม่ดีนัก

ดังนั้นวัตถุประสงค์ของงานวิจัยวิทยานิพนธ์นี้คือ เสนอวิธีการปรับปรุงภาพ โดยการกำจัดหรือลดสัญญาณรบกวนภายในภาพออกไป แล้วจึงทำการปรับปรุงภาพโดยทำให้ความเข้มสีบริเวณก้อนเนื้อในภาพชัดเจนขึ้น จากนั้นจึงใช้เทคนิคการประมวลผลภาพด้วยวิธีการหาขอบเขตที่น่าสนใจ โดยใช้ขั้นตอนวิธีในการตัดเฉพาะบริเวณก้อนเนื้อในภาพแมมโมแกรมเพื่อนำมาประมวลผล หลังจากได้บริเวณขอบเขตที่สนใจแล้ว ขั้นตอนก่อนการจำแนกอีกขั้นตอนหนึ่งคือการหาลักษณะสำคัญภายในบริเวณขอบเขตที่สนใจ โดยงานวิจัยนี้จะพิจารณาลักษณะสำคัญ 3 ลักษณะ คือ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของฮิสโตแกรม และ และในขั้นตอนสุดท้าย ลักษณะสำคัญทั้ง 3 แบบจะถูกนำไปใช้ในการจำแนก ด้วยเทคนิควิธีในการจำแนกข้อมูลแบบมีผู้สอนที่ชื่อว่าซัพพอร์ตเวกเตอร์แมชชีน วิธีซัพพอร์ตเวกเตอร์แมชชีนนี้นิยมใช้ในการจำแนกข้อมูลภาพ เนื่องจากข้อมูลภาพเป็นข้อมูลที่มีมิติข้อมูลสูง และซัพพอร์ตเวกเตอร์แมชชีนสามารถใช้ร่วมกับเคอร์เนลฟังก์ชันหลายแบบ งานวิจัยนี้ใช้เคอร์เนลฟังก์ชันเรเดียลเบสิสซึ่งให้ผลความแม่นในการจำแนกดีที่สุด และได้ทำการเปรียบเทียบประสิทธิภาพการจำแนกระหว่างเทคนิควิธีซัพพอร์ตเวกเตอร์แมชชีนกับเทคนิคการจำแนกอื่น ๆ เช่น โครงข่ายประสาทเทียม และ นาอีฟเบย์ โดยพบว่าการนำลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูลเข้าไปใช้ในการจำแนกทำให้ประสิทธิภาพในการจำแนกดีในทั้ง 3 อัลกอริทึม

5.1 สรุปผลการวิจัย

จากผลการทดสอบประสิทธิภาพของการจำแนกภาพแมมโมแกรม ด้วยชุดข้อมูลมาตรฐาน DDSM นั้น สามารถสรุปผลการทดสอบเป็นขั้นตอนดังแสดงในรูปที่ 5.1



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

ข้อมูลที่ใช้ในการฝึกสอน (Train) จำนวน 133 ภาพ

ข้อมูลที่ใช้ในการทดสอบ (Test) จำนวน 157 ภาพ

6) ประเมินประสิทธิภาพ

Accuracy, Sensitivity, Specificity, F-measure และ ROC

รูปที่ 5.1 ผลสรุปขั้นตอนการดำเนินงานในงานวิจัยนี้

1) ข้อมูลภาพแมมโมแกรม (Mammography)

ข้อมูลภาพแมมโมแกรมที่มาจากชุดข้อมูลมาตรฐาน DDSM เป็นภาพระดับสีเทาซึ่งมีขนาดภาพค่อนข้างใหญ่ประมาณ 3000 ×4000 พิกเซล โดยผู้วิจัยได้ทำการคัดเลือกภาพแมมโมแกรมในมุมมองแบบ CC มาจำนวน 190 ภาพ ประกอบด้วยภาพก้อนเนื้ออันตรายจำนวน 110 ภาพ และก้อนเนื้อไม่อันตรายจำนวน 80 ภาพ ในงานวิจัยนี้ใช้จำนวนภาพในการฝึกสอนโมเดล (Test) จำนวน 133 ภาพ ประกอบด้วยภาพก้อนเนื้ออันตรายจำนวน 77ภาพ ก้อนเนื้อไม่อันตรายจำนวน 56 ภาพ และใช้ภาพในการทดสอบ (Test) จำนวน 57 ภาพ ประกอบด้วยภาพก้อนเนื้ออันตรายจำนวน 33 ภาพ ก้อนเนื้อไม่อันตรายจำนวน 24 ภาพ

2) การปรับปรุงภาพแมมโมแกรม (Preprocessing)

- การกำจัดสัญญาณรบกวนภายในภาพด้วยวิธีมีเดียนฟิลเตอร์ (Median Filter) ซึ่งสัญญาณรบกวนนี้เกิดจากคุณภาพของอุปกรณ์ในการรับภาพ หลักการของมีเดียนฟิลเตอร์คือการใช้หน้าต่างขนาดเล็กเช่น 3×3 พิกเซล หรือ 5×5 พิกเซลเลื่อนไปบนภาพที่ต้องการกำจัดสัญญาณรบกวน โดยในขณะที่เลื่อนนั้นในหน้าต่างขนาดเล็กก็ทำหน้าที่ให้การคำนวณและเปลี่ยนแปลงค่าพิกเซล ณ จุดใด ๆ โดยทำการเรียงค่าความเข้มสีของบริเวณหน้าต่างที่ครอบภาพ และทำการเรียงค่าจากน้อยไปมาก จากนั้นจึงทำการคัดเลือกค่าที่มัธยฐานแล้วนำค่ามัธยฐานแทนที่ลงในพิกเซลปัจจุบัน ดังนั้นเมื่อภาพผ่านกระบวนการของมีเดียนฟิลเตอร์แล้ว ภาพจะมีความชัดเจนขึ้นในระดับหนึ่ง

-วิธีแก้ไขแกมมา (Gamma Correction) เป็นการปรับปรุงความชัดเจนของภาพ โดยเพิ่มความเข้มสีในบริเวณที่คาดว่าเป็นก้อนเนื้อเพื่อทำให้เห็นภาพบริเวณก้อนเนื้อได้ชัดเจนยิ่งขึ้น ในขณะเดียวกันก็ทำการปรับลดความเข้มสีในบริเวณที่เป็นพื้นหลังลงด้วย การแก้ไขแกมมาช่วยให้ความเข้มสีในบริเวณที่สว่างยิ่งมีความเข้มสีที่มากขึ้น และในทางตรงกันข้ามบริเวณที่เป็นพื้นหลังที่มีความเข้มสีที่ค่อนข้างมืดก็จะถูกปรับลดความเข้มสีลง ส่งผลให้บริเวณที่เป็นก้อนเนื้อมีความสำคัญชัดขึ้นมา โดยสอดคล้องกับที่นักรังสีวิทยาได้กล่าวไว้ว่า บริเวณที่เป็นก้อนเนื้อที่มีความผิดปกติ ความเข้มสีบริเวณนั้นจะมีมากกว่าบริเวณอื่น ๆ ซึ่งวิธีการนี้จะเป็นประโยชน์ในการนำไปเข้ากระบวนการแบ่งขอบเขตภาพหรือ ROI ต่อไป

3) การแบ่งขอบเขตภาพ (Image Segmentation)

ขั้นตอนการแบ่งขอบเขตภาพนั้นเป็นขั้นตอนที่สำคัญอีกขั้นตอนหนึ่งก่อนนำภาพไปทำการจำแนก เนื่องจากภาพแมมโมแกรมมีขนาดใหญ่ประมาณ 3000×4000 พิกเซล ซึ่งบริเวณที่สนใจนั้นจะเป็นบริเวณก้อนเนื้อที่มีความเข้มสี (Intensity) ค่อนข้างสูงเท่านั้น ดังนั้นพื้นหลังในภาพแมมโมแกรม และบริเวณภาพที่เป็นไขมันจึงไม่จำเป็นต้องนำเข้าสู่กระบวนการจำแนก

กระบวนการนี้จะช่วยลดขนาดข้อมูล ทำให้การจำแนกทำได้รวดเร็วขึ้นและไม่ทำให้เกิดปัญหาหน่วยความจำเต็มระหว่างการจำแนก ในขั้นตอนนี้เราจะใช้วิธีการการขยายพื้นที่ของส่วนภาพ หรือ Region Growing ซึ่งวิธีนี้เป็นวิธีที่นิยมใช้ในการแบ่งส่วนภาพภายในภาพแมมโมแกรม Region Growing ทำได้โดยเริ่มต้นพิจารณาจุดกึ่งกลาง เมื่อพบจุดภาพที่เป็นบริเวณขอบของ จุดกึ่งกลางหรือ Seed Region ก็จะพิจารณาจุดภาพข้างเคียง (Neighbor) ด้วยการวางหน้าต่างขนาด 3×3 ครอบจุดภาพนั้น หากจุดภาพข้างเคียงใดมีค่าระดับสีเทาอยู่ในขอบเขตของการขยายพื้นที่ ก็จะทำการรวมหรือขยายพื้นที่ส่วนภาพไปยังจุดข้างเคียงนั้น แต่ถ้าไม่ ก็จะพิจารณาจุดข้างเคียงถัดไป กระบวนการขยายส่วนพื้นที่ของภาพนี้ จะกระทำกับทุก ๆ Seed Region แบบวนซ้ำไปเรื่อย ๆ จนกระทั่งไม่สามารถขยายพื้นที่ได้

4) การดึงลักษณะสำคัญในภาพ (Image Feature Extraction)

ลักษณะสำคัญทั้ง 3 แบบที่ใช้ในงานวิจัยนี้ คือ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล (Additional Data from Shape Feature หรือ ADSF) ลักษณะสำคัญของลวดลาย (Texture Feature) และลักษณะสำคัญของกราฟฮิสโตแกรม (Histogram Feature) จากการทดลองพบว่าลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล (ADSF) มีผลอย่างมากในการเพิ่มประสิทธิภาพการจำแนก เมื่อเปรียบเทียบกับ ลักษณะสำคัญของรูปร่างแบบไม่เพิ่มชุดข้อมูล ซึ่งลักษณะสำคัญของรูปร่างนั้นจะเป็นตัวบ่งบอกถึงความหยักของก้อนเนื้อในภาพแมมโมแกรม ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ทำให้การจำแนกมีประสิทธิภาพเพิ่มมากขึ้นอย่างมีนัยสำคัญ

5) การจำแนกภาพแมมโมแกรมด้วย SVM (SVM Classification)

ในงานวิจัยนี้ทางผู้วิจัยได้ใช้ซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกโดยใช้เคอร์เนลฟังก์ชันเรเดียลเบสิสซึ่งให้ประสิทธิภาพในการจำแนกดีกว่าเคอร์เนลฟังก์ชันอื่น สำหรับลักษณะสำคัญที่ใช้เป็นข้อมูลในการจำแนกนั้นจะใช้ลักษณะสำคัญ 3 แบบ ได้แก่ ลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของความเข้มสีจากฮิสโตแกรม ในงานวิจัยนี้ได้นำการจำแนกอีก 2 วิธีคือนาอีฟเบย์ และ โครงข่ายประสาทเทียมมาเปรียบเทียบกับอัลกอริทึมที่นำเสนออีกด้วย

6) การประเมินประสิทธิภาพ (Evaluation)

สำหรับเกณฑ์ที่ใช้ในการประเมินสิทธิภาพของโมเดลในงานวิจัยนี้ ได้แก่ เกณฑ์ความแม่น (Accuracy) ค่า Sensitivity ค่า Specificity ค่า F-measure และ กราฟ Receiver Operation Characteristic หรือ กราฟ ROCในกระบวนการจำแนกด้วยอัลกอริทึม 3 แบบ คือ ซัพพอร์ตเวกเตอร์แมชชีน โครงข่ายประสาทเทียม และนาอีฟเบย์ จากผลการทดสอบประสิทธิภาพพบว่า

อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (เคอร์เนลฟังก์ชัน เรเดียลเบสิส) ให้ค่าความแม่น ค่า Sensitivity ค่า Specificity ค่า F-measure และพื้นที่ใต้กราฟ ROC สูงที่สุด

จากการสรุปผลการทดสอบประสิทธิภาพของการจำแนกภาพแมมโมแกรมจะเห็นได้ว่า การประมวลผลภาพมีความจำเป็นอย่างมากเพื่อลดทอนข้อมูลภาพแมมโมแกรมให้มีขนาดเล็กลง รวมทั้งปรับปรุงคุณภาพของภาพแมมโมแกรมให้มีความชัดเจนมากยิ่งขึ้น จากการเปรียบเทียบอัลกอริทึมสำหรับจำแนกทั้ง 3 แบบ ผลปรากฏว่าซัพพอร์ตเวกเตอร์แมชชีน (เคอร์เนลฟังก์ชัน เรเดียลเบสิส) มีความเหมาะสมและมีประสิทธิภาพดีที่สุดในการจำแนกข้อมูลภาพแมมโมแกรม เมื่อใช้ข้อมูลลักษณะสำคัญร่วมกันทั้ง 3 แบบ คือ ลักษณะสำคัญของรูปร่าง แบบเพิ่มชุดข้อมูล ลักษณะสำคัญของลวดลาย และลักษณะสำคัญของกราฟฮิสโตแกรม โดยเฉพาะอย่างยิ่งลักษณะสำคัญของรูปร่างแบบเพิ่มชุดข้อมูล สามารถช่วยให้การจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพมากยิ่งขึ้น

5.2 ปัญหาและข้อเสนอแนะ

เนื่องจากภาพแมมโมแกรมมีขนาดภาพที่ใหญ่มาก ทำให้ในส่วนของการประมวลผลภาพมีความล่าช้า เนื่องจากในงานวิจัยนี้ได้ใช้การประมวลผลภาพหลากหลายวิธี เพื่อทำการปรับปรุงภาพให้มีคุณภาพที่ดีขึ้น ก่อนนำไปเข้ากระบวนการจำแนก ซึ่งงานวิจัยในอนาคตอาจจะพิจารณานำเอาเทคนิคการปรับปรุงภาพอื่น ๆ มาใช้ เพื่อทำให้กระบวนการปรับปรุงภาพมีความรวดเร็วมากยิ่งขึ้น

งานวิจัยนี้ในขั้นตอนการคัดเลือกลักษณะสำคัญนั้น เป็นการทดสอบการผสมผสานลักษณะสำคัญ 3 แบบ และนำมาเปรียบเทียบประสิทธิภาพเพื่อสรุปว่าลักษณะสำคัญใด มีผลทำให้ประสิทธิภาพในการจำแนกดีขึ้น ซึ่งการนำลักษณะสำคัญมาผสมผสานกันนั้นยังไม่เป็นอัตโนมัติ ทำให้หากมีลักษณะสำคัญหลาย ๆ แบบจะทำให้สิ้นเปลืองเวลาในการทดสอบและเปรียบเทียบ ดังนั้น ทางผู้วิจัยจึงเล็งเห็นว่าควรนำเทคนิคในการคัดเลือกลักษณะสำคัญแบบอัตโนมัติมาใช้ เพื่อเพิ่มประสิทธิภาพในการคัดเลือกลักษณะสำคัญของภาพแมมโมแกรม

รายการอ้างอิง

Aguiar, A., Pereira, G., Azevedo, I. and Gomes, L. (2015). Evaluation of axillary dose coverage following whole breast radiotherapy: variation with the breast volume and shape. **Radiotherapy and Oncology.** 114(1): 22-27.

Al-Najdawi, N., Biltawi, M. and Tedmori, S. (2015). Mammogram image visual enhancement mass segmentation and classification. **Applied Soft Computing.** 35(1): 175-185.

Berannek, R., Jakubowski, W., Mazurczak, A., Postolski, M. and Wiazel, W. (1998). Contrast enhanced evaluation of the solid lesions in the breast-own experience**. European Journal of Ultrasound.** 7(1): S13.

Braz, G., Paiva, C., Silva, A. and Oliveira, A. (2009). Classification of breast tissues using moran's index and geary's coefficient as texture signatures and SVM. **Computers in Biology and Medicine.** 39(1): 1063-1072.

Chen, T., Gao, H., Guo, W., Qing, X., Gao, K., Yu, J. and Deng, Y. (2015). A novel application of the automated breast volume scanner (abvs) in the diagnosis of soft tissue tumors. **Clinical Imaging.** 39(3): 401-407.

Dhahbi. S., Barhoumi, W. and Zagrouba, E. (2015). Breast cancer diagnosis in digitized mammograms using curvelet moments. **Computers in Biology and Medicine.** 64(1): 79-90.

Dietzel, M., Baltzer, P., Dietzel, A., Zoubi, R., Groschel, T., Burmeister, H., Bogdan, M. and Kaiser, W. (2012). Artificial neural networks for differential diagnosis of breast lesions in mr-mammography: a systematic approach addressing the influence of network architecture on diagnostic performance using a large clinical database. **European Journal of Radiology.** 81(7): 1508-1513.

Hassanien, A. and Kim, T. (2012). Breast cancer mri diagnosis approach using support vector machine and pulse coupled neural networks. **Journal of Applied Logic.** 10(1): 277-284.

Helena, G., Miranda, B. and Felipe, J. (2015). Computer-aided diagnosis system based on fuzzy logic for breast cancer categorization. **Computers in Biology and Medicine.** 64(1): 334-346.

Huo, Z., Giger, M., Vyborny, C., Wolverton, D. and Metz C. (2000). Computerized classification of benign and malignant on digitized mammograms: a study of robustness. **Academic Radiology.** 7(12): 1077-1084

Jalalian, A., Mashohor, S., Mahmud, H., Saripan, M., Rahman, A., Ramli, B. and Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. **Clinical Imaging.** 37(3): 420-426.

Wajid, S. and Hussain, A. (2015). Local energy-based shape histogram feature extraction technique for breast cancer diagnosis. **Expert Systems with Applications.** 42(20): 6990-6999.

Karmilasari, K., Widodo, S., Hermita, M., Agustiyani, N. and Hanum, Y. (2014). Sample k-means clustering method for determining the stage of breast cancer malignancy based on cancer size on mammogram image basis. **International Journal of Advanced Computer Science and Application.** 5(3): 86-90.

Lee, H. and Chen, Y. (2015). Image based computer aided diagnosis system for cancer detection. **Expert Systems with Applications.** 42(1): 5356-5365.

Lo, C. and Wang, C. (2012). Support vector machine for breast mr image classification. **Computers and Mathematics with Applications.** 64(1): 1153-1162.

Oliver, A., Llado, X., Perez, E., Pont, J., Denton, E., Freixenet, J. and Marti J. (2010). A statistical approach for breast density segmentation. **Journal of Digital Imaging.** 23(5): 527-537.

Oliver, A., Freixenet, J., Marti, J., Perez, E., Pont, J. and Denton E. (2010). A review of automatic mass detection and segmentation in mammographic imges. **Medical Image Analysis.** 14(1): 87-110.

Pak, F., Kanan, H. and Alikhassi, A. (in press). Breast cancer detection and classification in digital mammography based on non-subsampled contourlet transform (nsct) and super resolution. **Computer Methods and Programs in Biomedicine**.

Perez, N., Lopez, M. and Silva, A. (2015). Improving the mann-whitney statistical test for feature seletion: An approach in breast cancer diagnosis on mammography. **Artificial Intelligence in Medicine.** 63(1): 19-31.

Robert, M., Shanmugam, K. and Dinstein, I. (1973). Texture features for image classification. **IEEE Transactions on Systems, Man, and Cybernetics.** 3(6): 610-621

Rouhi, R., Jafari, M., Kasaei, S. and Keshavarzian, P. (2015). Benign and malignant breast tumors classification based on region growing and CNN segmentation. **Expert Systems with Applications.** 42(1): 990-1002.

Shi, X., Cheng, H.D., Hu, L., Ju, W. and Tian, J. (2010). Detection and classification of masses in breast ultrasound images. **Digital Signal Processing.** 20(1): 824-836.

Singh, B., Verma, K. and Thoke, A.S. (2015). Adaptive gradient descent backpropagation for classification of breast tumors in ultrasound imaging. **Procedia Computer Science**. 46(1): 1601-1609.

Szeliski, R. (2010). **Computer Vision Algorithms and Applications**. Springer.

Wu, W., Lin, S. and Moon, W. (2012). Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. **Computerized Medical Imaging and Graphpics.** 36(1): 627-633.

Wu, Y., Freedman, M., Hasegawa, A., Zurbier, R., Lo, S. and Mun, S. (1995). Classification of microcalcifications in radiographs of pathologic specimens for diagnosis of breast cancer. **Academic Radiology**. 2(3): 199-204.

Xie, W., Li, Y. and Ma, Y. (in press). Breast mass classification in digital mammography based on extreme learning machine. **Neurocomputing**.

Zaki, M. (2014). Support Vector Machine. **Data Mining and Analysis: Fundamental Concepts and Algorithms**. 566-601.

Zhou, S., Shi, J., Cai, J. and Wang, R. (2013). Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image. **Biomedical Signal Processing and Control**. 8(6): 688-696.

Zyout, I., Czajkowska, J. and Grzegorzek, M. (in press). Multi-scale texture feature extraction and particle swarm optimization based model selection for false positive reduction in mammography. **Computerized Medical Imaging and Graphics.**

ภาคผนวก ก

รหัสต้นฉบับของโปรแกรม

%ฟังก์ชันการปรับปรุงแกมมา (MATLAB)

```
% Author: Pranam Janney                    02/11/2003    14:30
% Outputs:
%       Correction = gamma correction for the input image
% Inputs:
%       Image = input image file
%       GammaValue= gamma correction factor, if not specified gamma = 1

function Y=gamma_correction(X, in_interval, out_interval, gamma);
% Default return value
% X is an input image. Y is an output image
Y=[];
% Parameter check
if nargin~=4
  disp('Error: The function takes exactly four arguments.');
  return;
end
% Init. Operations
X=double(X);
[a,b]=size(X);
% Map to input interval
if ~isempty(in_interval)
   if length(in_interval)==2
      X=adjust_range(X,in_interval);
   else
     disp('Error: Input interval needs to be a two-component vector.');
     return;
   end
end
% Do gamma correction
X=X.^gamma;
```

```
% Map to output interval
if ~isempty(out_interval)
   if length(out_interval)==2
      Y=adjust_range(X,out_interval);
   else
      disp('Error: Output interval needs to be a two-component vector.');
      return;
   end
end
```

%ฟังก์ชันการขยายส่วนพื้นที่ (MATLB)

```
% Author: D. Kroon, University of Twente
% Outpur
%       J : logical output image of region

%Input
%       I : input image
%       x,y : the position of the seedpoint (if not given uses function getpts)
%       t : maximum intensity distance (defaults to 0.2)


function J=regiongrowing(I,x,y,reg_maxdist)
if(exist('reg_maxdist','var')==0), reg_maxdist=0.2; end
if(exist('y','var')==0), figure, imshow(I,[]); [y,x]=getpts; y=round(y(1)); x=round(x(1)); end
J = zeros(size(I)); % Output
Isizes = size(I); % Dimensions of input image
reg_mean = I(x,y); % The mean of the segmented region
reg_size = 1; % Number of pixels in region

% Free memory to store neighbours of the (segmented) region
neg_free = 10000; neg_pos=0;
neg_list = zeros(neg_free,3);
pixdist=0; % Distance of the region newest pixel to the regio mean

% Neighbor locations (footprint)
neigb=[-1 0; 1 0; 0 -1;0 1];

% Start regiogrowing until distance between regio and posible new pixels become
% higher than a certain threshold
while(pixdist<reg_maxdist&&reg_size<numel(I))
   % Add new neighbors pixels
   for j=1:4,
```

```
        % Calculate the neighbour coordinate
        xn = x +neigb(j,1); yn = y +neigb(j,2);
            % Check if neighbour is inside or outside the image
        ins=(xn>=1)&&(yn>=1)&&(xn<=Isizes(1))&&(yn<=Isizes(2));
            % Add neighbor if inside and not already part of the segmented area
        if(ins&&(J(xn,yn)==0))
            neg_pos = neg_pos+1;
            neg_list(neg_pos,:) = [xn yn I(xn,yn)]; J(xn,yn)=1;
        end
    end
    % Add a new block of free memory
    if(neg_pos+10>neg_free), neg_free=neg_free+10000; neg_list((neg_pos+1):neg_free,:)=0; end
    % Add pixel with intensity nearest to the mean of the region, to the region
    dist = abs(neg_list(1:neg_pos,3)-reg_mean);
    [pixdist, index] = min(dist);
    J(x,y)=2; reg_size=reg_size+1;
    % Calculate the new mean of the region
    reg_mean= (reg_mean*reg_size + neg_list(index,3))/(reg_size+1);
     % Save the x and y coordinates of the pixel (for the neighbour add proccess)
    x = neg_list(index,1); y = neg_list(index,2);
      % Remove the pixel from the neighbour (check) list
    neg_list(index,:)=neg_list(neg_pos,:); neg_pos=neg_pos-1;
end
% Return the segmented area as logical matrix
J=J>1;
```

%ฟังก์ชันลักษณะสำคัญของฮิสโตแกรม (MATLAB)

```
function [meanGL varianceGL skew kurtosis] = GetSkewAndKurtosis(GLs, pixelCounts)
  try
    % Get the number of pixels in the histogram.
    numberOfPixels = sum(pixelCounts);
    % Get the mean gray lavel.
    meanGL = sum(GLs .* pixelCounts) / numberOfPixels;
    % Get the variance, which is the second central moment.
    varianceGL = sum((GLs - meanGL) .^ 2 .* pixelCounts) / (numberOfPixels-1);
    % Get the standard deviation.
    sd = sqrt(varianceGL);
    % Get the skew.
    skew = sum((GLs - meanGL) .^ 3 .* pixelCounts) / ((numberOfPixels - 1) * sd^3);
    % Get the kurtosis.
    kurtosis = sum((GLs - meanGL) .^ 4 .* pixelCounts) / ((numberOfPixels - 1) * sd^4);
  catch ME
    errorMessage = sprintf('Error in GetSkewAndKurtosis().\nThe error reported by MATLAB
is:\n\n%s', ME.message);
    uiwait(warndlg(errorMessage));
    set(handles.txtInfo, 'String', errorMessage);
  end
  return; % from GetSkewAndKurto
```

%ฟังก์ชันลักษณะสำคัญของรูปร่าง (MATLAB)

```
% Eli Billauer, 3.4.05
% Output
%       maxtab: maximum peak
%       mintab: minimum peak
% Input
%       v: Vector  delta:   x:


function [maxtab, mintab]=detectpeak(v, delta, x)
%Detect peaks in a vector
maxtab = [];
mintab = [];
 v = v(:); % Just in case this wasn't a proper vector
 if nargin < 3
  x = (1:length(v))';
else
  x = x(:);
  if length(v)~= length(x)
    error('Input vectors v and x must have same length');
  end
end


if (length(delta(:)))>1
  error('Input argument DELTA must be a scalar');
end


if delta <= 0
  error('Input argument DELTA must be positive');
end


mn = Inf; mx = -Inf;
```

```
mnpos = NaN; mxpos = NaN;


lookformax = 1;
 for i=1:length(v)
  this = v(i);
  if this > mx, mx = this; mxpos = x(i); end
  if this < mn, mn = this; mnpos = x(i); end

  if lookformax
    if this < mx-delta
      maxtab = [maxtab ; mxpos mx];
      mn = this; mnpos = x(i);
      lookformax = 0;
    end
  else
    if this > mn+delta
      mintab = [mintab ; mnpos mn];
      mx = this; mxpos = x(i);
      lookformax = 1;
    end
  end
end
```

%ฟังก์ชันซัพพอร์ตเวกเตอร์แมชชีน (MATLAB)

```
function [outclass, f]=my_svm()
outclass = []; %output class
f = []; %output score
rrow = 190;
classcol = 20;
dat = csvread('total_shape.csv');
data = dat(1:rrow,1:classcol);
groups = dat(1:rrow,classcol+1);
groups2 = vertcat(dat(1:56,classcol+1),dat(81:157,classcol+1));
groups3 = vertcat(dat(57:80,classcol+1),dat(158:190,classcol+1));
% Matlab build-in SVM
 [g, gn]              = grp2idx(dat(1:rrow,classcol+1));    % Nominal class to numeric

% Split training and testing sets
[trainIdx, testIdx]    = crossvalind('HoldOut', dat(1:rrow,classcol+1),0.3,'classes',{'0','1'});

pairwise             = nchoosek(1:length(gn),2);          % 1-vs-1 pairwise models
svmModel               = cell(size(pairwise,1),1);          % Store binary-classifers
predTest             = zeros(sum(testIdx),numel(svmModel)); % Store binary predictions

% classify using one-against-one approach, SVM with rbf
for k=1:numel(svmModel)
   % get only training instances belonging to this pair
   idx = trainIdx
   % train
   svmModel{k}            = svmtrain(data(idx,:), g(idx), ...
   'Autoscale',true, 'Showplot',false, 'Method','QP', ...
            'BoxConstraint',2e-1, 'Kernel_Function','rbf', 'RBF_Sigma',1);
  % test
```

```
   predTest(:,k)          = svmclassify(svmModel{k}, data(testIdx,:));
   [outclass,f]           = svmclassify2(svmModel{k}, data(testIdx,:));
end


pred = mode(predTest,2);   % Voting: clasify as the class receiving most votes


% performance
cmat                      = confusionmat(g(testIdx),pred);
acc                       = 100*sum(diag(cmat))./sum(cmat(:));
fprintf('SVM (1-against-1):\naccuracy = %.2f%%\n', acc);
fprintf('Confusion Matrix:\n'), disp(cmat)
```

%ฟังก์ชันกราฟ ROC (MATLAB)

```
% Output
%         Tps: True positive rate    Fps: False positive rate
% Input
%         Scores : Score from classification   labels: Label of class from classification

function [Tps, Fps] = ROC(scores, labels)
% Sort Labels and Scores by Scores
sl = [scores; labels];
[d1 d2] = sort(sl(1,:));
sorted_sl = sl(:,d2);
s_scores = sorted_sl(1,:);
s_labels = round(sorted_sl(2,:));

 % Constants
counts = histc(s_labels, unique(s_labels));
Tps = zeros(1, size(s_labels,2) + 1);
Fps = zeros(1,  size(s_labels,2) + 1);
negCount = counts(1);
posCount = counts(2);

 % Shift threshold to find the ROC
for thresIdx = 1:(size(s_scores,2)+1)
   % for each Threshold Index
   tpCount = 0;
   fpCount = 0;
    for i = [1:size(s_scores,2)]
      if (i >= thresIdx)          % We think it is positive
         if (s_labels(i) == 1)   % Right!
            tpCount = tpCount + 1;
         else                    % Wrong!
            fpCount = fpCount + 1;
```

```
        end
      end
    end
    Tps(thresIdx) = tpCount/posCount;
    Fps(thresIdx) = fpCount/negCount;
end

 % Draw the Curve
% Sort [Tps;Fps]
x = Tps;
y = Fps;
% Interplotion to draw spline line
count = 100;
dist = (x(1) - x(size(x,2)))/100;
xx = [x(1):-dist:x(size(x,2))];

 % In order to get the interpolations, we remove all the unique numbers
[d1 d2] = unique(x);
uni_x = x(1,d2);
uni_y = y(1,d2);
yy = spline(uni_x,uni_y,xx);

 % No value should exceed 1
yy = min(yy, 1);
plot(x,y,'x',xx,yy);
```

#ภาษา R เรียกใช้ฟังก์ชัน plotOverlappingHist

```
library(EbayesThresh)
library(ggplot2)
df <- read.csv(file= "C:/Users/Kedkarn/Documents/R_SVR/total_shape.csv",header =
TRUE,sep=",")
peakk <- df[1:190,20]
tfromx(peakk, prior="laplace")
bb <- data.frame(df[1:77,20])
mm <- data.frame(df[81:136,20])
threshold <- quantile(peakk, probs=0.48)
threshold
ggplot(df, aes(x=peakk, color="blue")) + geom_density() + geom_vline(xintercept=threshold)
###################################3
#combine two dataframes into one.  First make a new column in each.
bb$veg <- 'benign'
mm$veg <- 'malignant'
colnames(bb) <- c("peak", "veg")
colnames(mm) <- c("peak", "veg")
#combine into the new data frame vegLengths
vegLengths <- rbind(bb, mm)
#make your plot
ggplot(vegLengths, aes(peak, fill = veg),xlim=NULL,ylim=NULL) + geom_density(alpha = 0.2)
###################################
ov <- plotOverlappingHist(bb$peak, mm$peak, colors=c("white","gray20","yellow"),
                breaks=NULL, xlim=NULL, ylim=NULL)
```

#ฟังก์ชันภาษา R แสดงการพล็อตฮิสโตแกรมที่ซ้อนทับกัน

```
# Output
#          overlap: Overlap value from two histogram
# Input
#          a: vector a    b: vector b    colors : Colors for histogram1 histogram2  and overlapping
area

library(ggplot2)
plotOverlappingHist <- function(a, b, colors=c("white","gray20","gray50"),
breaks=NULL, xlim=NULL, ylim=NULL){
 ahist=NULL
 bhist=NULL
 if(!(is.null(breaks))){
  ahist=hist(a,breaks=breaks,plot=F)
  bhist=hist(b,breaks=breaks,plot=F)
 } else {
  ahist=hist(a,plot=F)
  bhist=hist(b,plot=F)
  dist = ahist$breaks[2]-ahist$breaks[1]
  breaks = seq(min(ahist$breaks,bhist$breaks),max(ahist$breaks,bhist$breaks),dist
  ahist=hist(a,breaks=breaks,plot=F)
  bhist=hist(b,breaks=breaks,plot=F)
 }
 if(is.null(xlim)){
  xlim = c(min(ahist$breaks,bhist$breaks),max(ahist$breaks,bhist$breaks))
 }
 if(is.null(ylim)){
  ylim = c(0,max(ahist$counts,bhist$counts))
 }
 overlap = ahist
 for(i in 1:length(overlap$counts)){
```

```
    if(ahist$counts[i] > 0 & bhist$counts[i] > 0){
      overlap$counts[i] = min(ahist$counts[i],bhist$counts[i])
    } else {
      overlap$counts[i] = 0
    }
  }
  xlim = c(30,150)
  ylim = c(0,20)
  plot(ahist, xlim=xlim, ylim=ylim, col=colors[1])
  plot(bhist, xlim=xlim, ylim=ylim, col=colors[2], add=T)
  plot(overlap, xlim=xlim, ylim=ylim, col=colors[3], add=T)
return(overlap)
}
```

ภาคผนวก ข

บทความวิชาการที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างการศึกษา

รายชื่อบทความวิชาการที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างการศึกษา

Kedkarn Chaiyakhan, Nittaya Kerdprasop and Kittisak Kerdprasop. (2015). **Mammography ImagesCategorization with k-Means Clustering.** SEATUC'2015 the 9th South East Asia Technical University Consortium (SEATUC) Symposium 2015, Suranaree University of Technology, Thailand. 27-30 July 2015.

Kedkarn Chaiyakhan, Nittaya Kerdprasop and Kittisak Kerdprasop. (2016). **Feature Selection Techniques for Breast Cancer Image Classification with Support Vector Machine**. IMECS'2016, International MultiConference of Engineers and Computer Scientists 2016, Hong Kong. 16-18 March 2016.

Kedkarn Chaiyakhan, Nittaya Kerdprasop and Kittisak Kerdprasop. (2016). **Mammography Image Classification and Clustering using Support Vector Machine and k-Means.** ICIC Express Letters, Part B: Applications, International Journal of Research and Surveys, 7(5) : 961-967.

MAMMOGRAPHY IMAGES CATEGORIZATION WITH K-MEANS CLUSTERING

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

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

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.

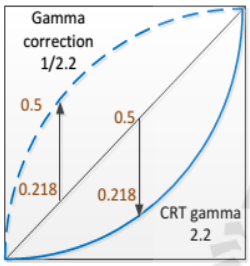


Fig. 1 Gamma correction model.

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 $D = \{x_1, \ldots, x_N\}$. Since the data is p-dimensional, then represent it as $X_n = \{x_{n,1}, \ldots, x_{n,p}\}$. The distance function is $d(X_n, X_m)$ between two data points. The k groups has distinguish the data into $\{z_1, \ldots, z_N\}$ where $x \in \{1, \ldots, K\}$.

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.

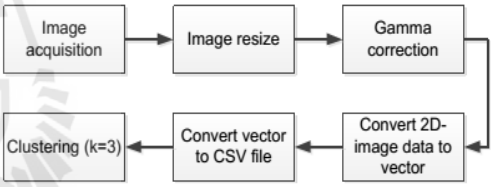


Fig. 2 The framework of the proposed work.

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.

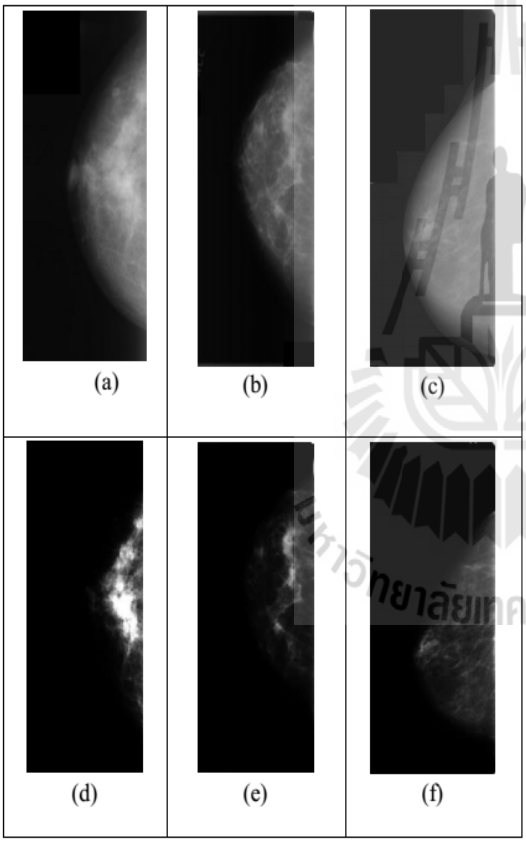


Fig. 3 (a) Original malignant, (b) original benign, (c) original normal, (d) corrected malignant, (e) corrected benign, (f) corrected normal.

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**

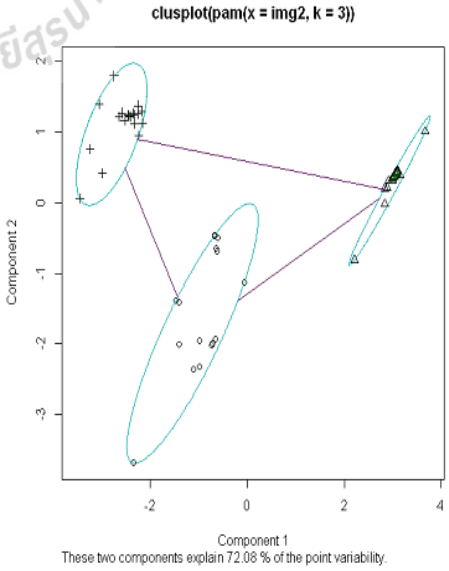


Fig. 4 Two components of clustering plot.

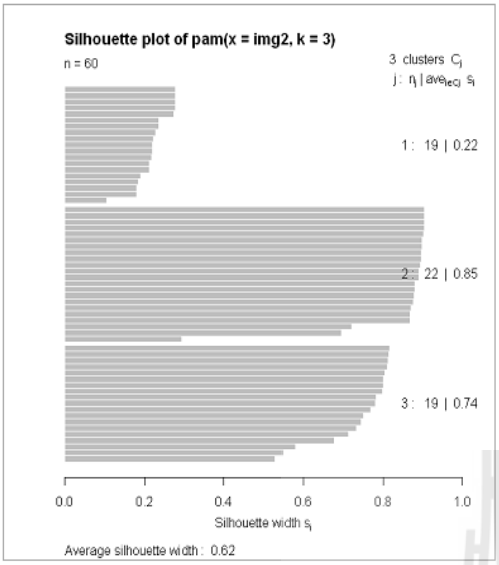


Fig. 5 The silhouette plot when k=3.

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.

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]. 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.

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).

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).

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

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.

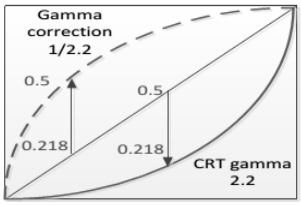


Fig. 1. Gamma correction model.

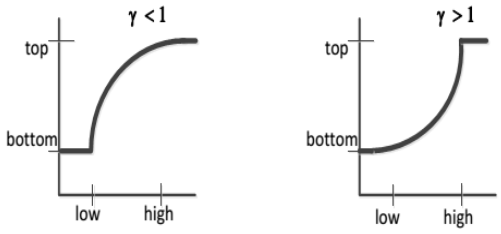


Fig. 2. Two different gamma correction settings.

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.

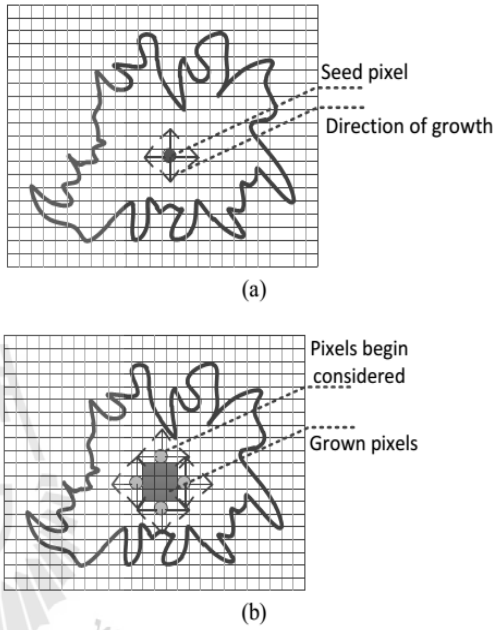


Fig. 3. The example of region growing.

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.

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

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

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

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

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).

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

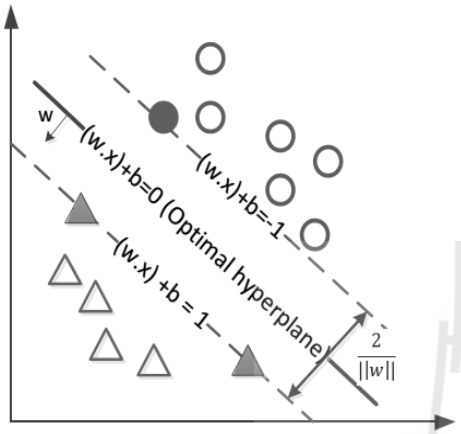


Fig. 4. Optimal hyperplane with maximum margin.

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.

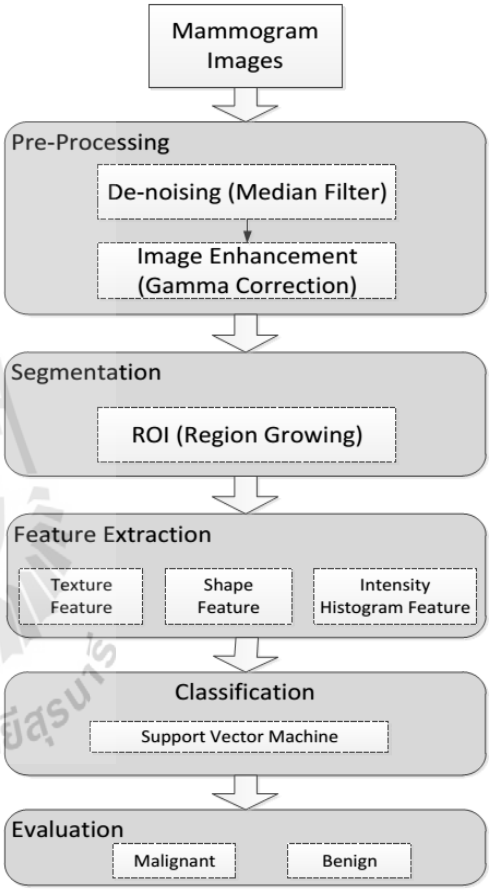


Fig. 5. The framework of the proposed system.

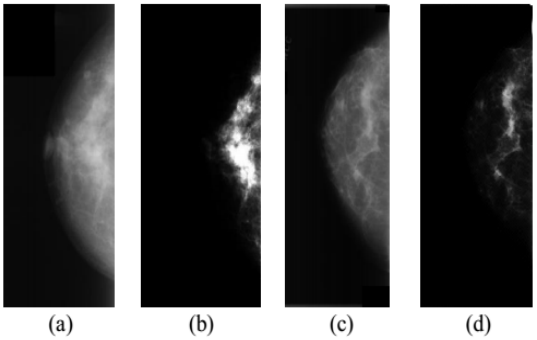


Fig. 6. Breast tumor images: (a) original malignant tumor, (b) malignant tumor after gamma correction, (c) original benign tumor, (d) benign tumor after gamma correction.

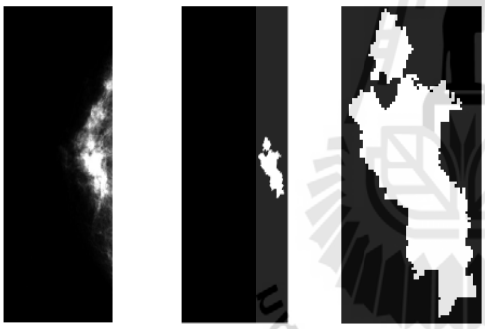
B. Segmentation

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

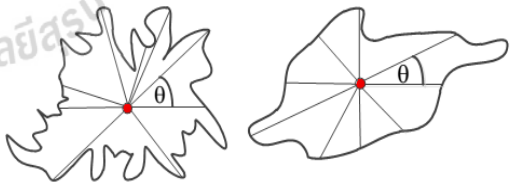
Texture is one of the important features used in identifying objects in an image. Texture features are based on gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The GLCM function characterizes the texture of an image by calculating how often pairs of pixels 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.

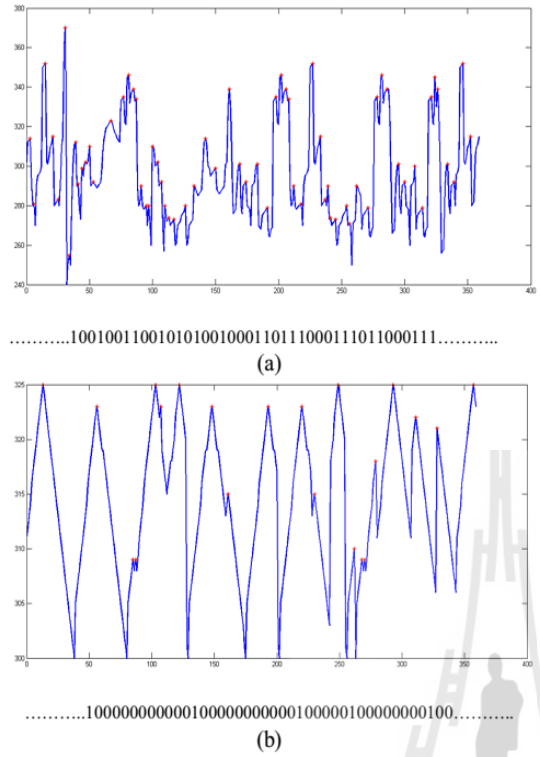


Fig. 9. Graph of curvature: (a) malignant contour (b) benign contour.

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 80 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.

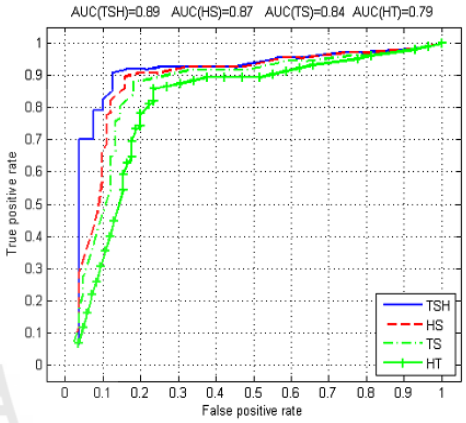


Fig. 10. Area under curve of four combination features.

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.

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.

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.

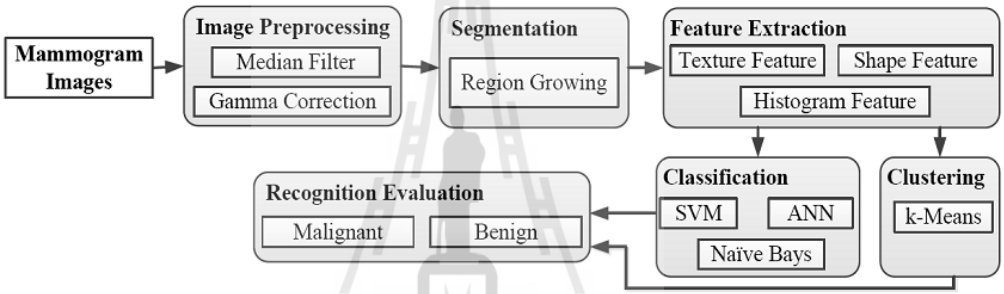


FIGURE 1. The framework of the proposed tumor recognition system

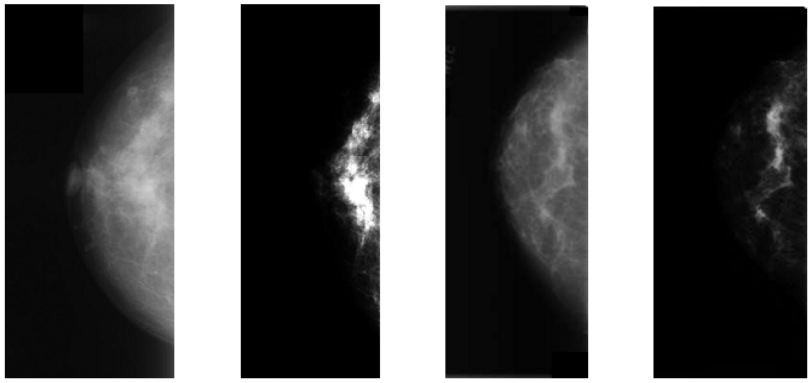
2.1. **Image preprocessing.** Mammogram images usually contain noises because of disturbances like Gaussian noise or some little darkness and brightness noise called salt and pepper noise. We use median filter to remove these noises. The output of this de-noising step is the clear images that are appropriate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

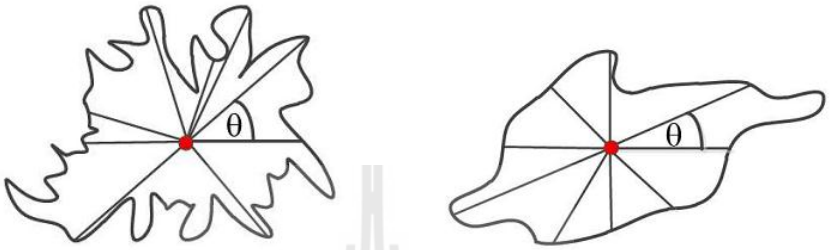
1). Texture Features

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

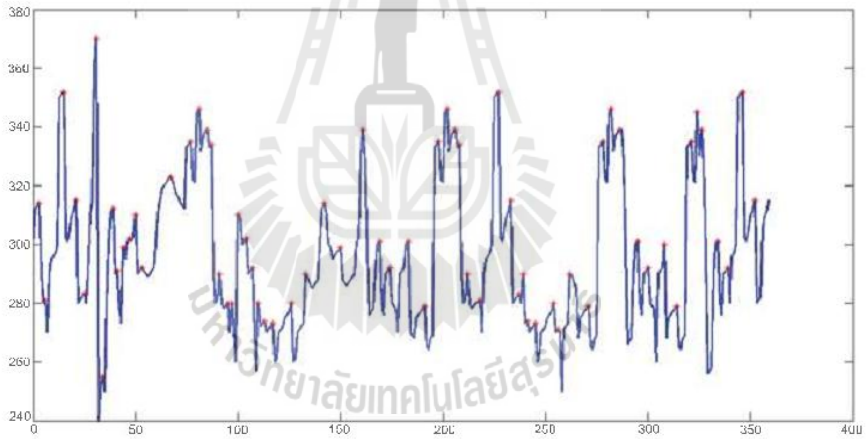
We extract shape feature using the percentage of curvature. First we drag lines from centroid to every edge pixel and then measure distance and angle from centroid to every edge pixel. After that, we plot the graph with angle along the x-axis and distance on the y-axis. From the graph, we can notice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

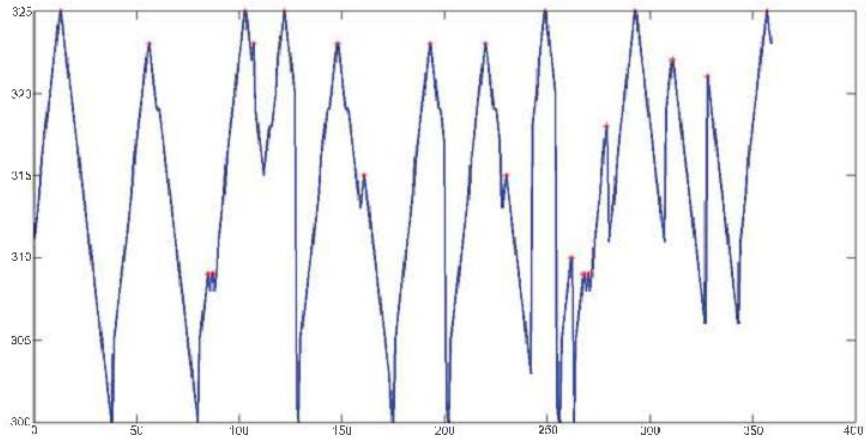


(a) (b)

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

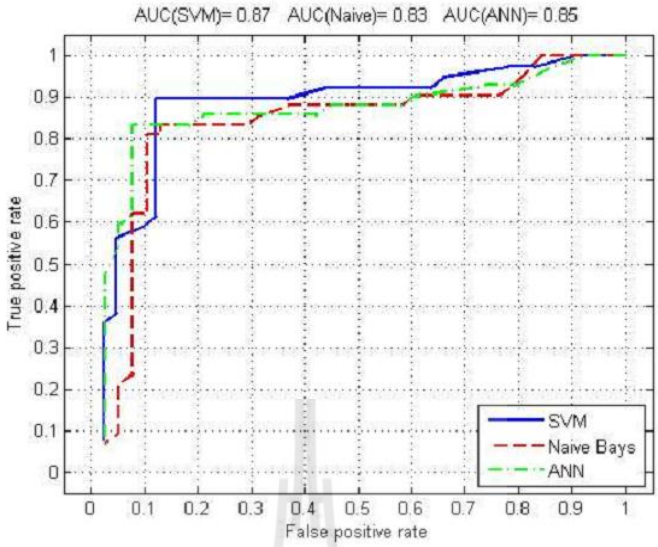
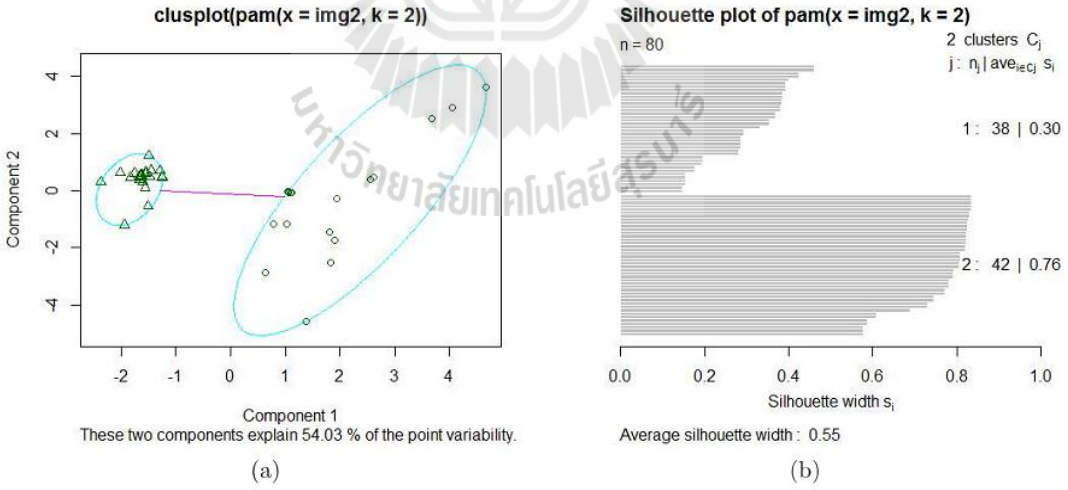


FIGURE 6. Area under curve of three classifiers

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.

ประวัติผู้เขียน

นางเกตุกาญจน์ ไชยขันธุ์ เกิดเมื่อวันที่ 7 กุมภาพันธ์ 2519 ที่อำเภอบัวใหญ่ จังหวัดนครราชสีมา เริ่มเข้าศึกษาชั้นประถมศึกษาปีที่ 1 ถึง 6 ที่โรงเรียนมารีย์วิทยา อำเภอเมือง จังหวัดนครราชสีมา จากนั้นศึกษาต่อในระดับมัธยมตอนต้นและตอนปลายที่โรงเรียนสุรนารีวิทยา อำเภอเมือง จังหวัดนครราชสีมา ในปีการศึกษา 2534 ได้เข้าศึกษาระดับประกาศนียบัตรวิชาชีพชั้นสูงในโปรแกรมวิชาเทคนิคคอมพิวเตอร์ คณะวิชาไฟฟ้า สถาบันเทคโนโลยีราชมงคลภาคตะวันออกเฉียงเหนือ และสำเร็จการศึกษาเมื่อปี พ.ศ. 2537 จากนั้นในปีการศึกษา 2538 ได้เข้าศึกษาต่อระดับปริญญาตรีในสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ ศูนย์กลางสถาบันเทคโนโลยีราชมงคลธัญบุรี และสำเร็จการศึกษาเมื่อไป พ.ศ. 2540

ภายหลังสำเร็จการศึกษาในระดับปริญญาตรี ได้เข้าทำงานในสำนักเทคโนโลยีสารสนเทศ ศูนย์กลางสถาบันเทคโนโลยีราชมงคลธัญบุรี ในตำแหน่งโปรแกรมเมอร์ เมื่อปี พ.ศ. 2541 – 2542 จากนั้นได้เข้าทำงานที่บริษัทดาต้าโกลบ ซอฟต์แวร์ จำกัด ในตำแหน่งโปรแกรมเมอร์ เมื่อปี พ.ศ. 2543 – 2544 หลังจากนั้นได้สอบบรรจุเข้ารับราชการในตำแหน่งอาจารย์ ประจำสาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ในปี พ.ศ. 2544

ในปี 2546 ได้รับทุนสนับสนุนการศึกษาจากมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ไปศึกษาต่อระดับปริญญาโท ที่มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี สาขาวิศวกรรมคอมพิวเตอร์ และสำเร็จการศึกษาในปี พ.ศ. 2548 และในปีการศึกษา 2557 ได้ศึกษาต่อระดับปริญญาเอกในสาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

ในระหว่างการศึกษาได้รับความอนุเคราะห์เป็นอย่างดีจากอาจารย์ที่ปรึกษาและอาจารย์ประจำวิชาต่าง ๆ และได้รับการตีพิมพ์เผยแพร่บทความวิชาการซึ่งรายละเอียดสามารถดูได้ที่ภาคผนวก ข