วารสารวิทยาศาสตร์ลาดกระบัง ปีที่ 34 ฉบับที่ 2 เดือนกรกฎาคม - ธันวาคม 2568
Journal of Science Ladkrabang Vol. 34 No. 2 July - December 2025

บทความวิจัย (Research article)

การทดสอบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูล สำหรับคัดกรองผู้ป่วยโรคมะเร็งเต้านม

Assessing the Performance of Data Mining Techniques for Breast Cancer Patient Screening

อนุพงศ์ สุขประเสริฐ* ศิรินภา พรมโสภา และ ยศภัทร ศรีโม๊ะ
Anupong Sukprasert* Sirinapa Phomsopa and Yossapat Srimo
สาขาคอมพิวเตอร์ธุรกิจ คณะการบัญชีและการจัดการ มหาวิทยาลัยมหาสารคาม จ.มหาสารคาม ประเทศไทย
Department of Business Computer Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand
วันที่ส่งบทความ : 14 เมษายน 2568 วันที่แก้ไขบทความ : 31 ตุลาคม 2568 วันที่ตอบรับบทความ : 10 พฤศจิกายน 2568
Received: 14 April 2025, Revised: 31 October 2025, Accepted: 10 November 2025

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพของเทคนิคการทำเหมืองข้อมูลในการสร้างแบบจำลองสำหรับคัดกรองผู้ป่วยโรคมะเร็งเต้านม โดยเปรียบเทียบเทคนิคการจำแนกข้อมูล 7 วิธี ได้แก่ โครงข่ายประสาทเทียม (Neural networks) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) นาอีฟเบย์ (Naïve Bayes) เพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors: k-NN) ต้นไม้ตัดสินใจ (Decision tree) การเรียนรู้เชิงลึก (Deep learning) และวิธีการรวมกลุ่ม (Ensemble vote) ข้อมูลที่ใช้ในการวิเคราะห์เป็นชุดข้อมูลผู้ป่วยจำนวน 569 ราย จากฐานข้อมูลของมหาวิทยาลัยวิสคอนซินซึ่งเผยแพร่ผ่านเว็บไซต์ www.kaggle.com โดยดำเนินการวิเคราะห์ข้อมูลตามกระบวนการ CRISP-DM ซึ่งประกอบด้วยขั้นตอนการคัดเลือกตัวแปร การจัดการข้อมูลสูญหาย และการกำหนดบทบาทของแอตทริบิวต์แต่ละตัว ผลการศึกษาแสดงให้เห็นว่า เทคนิคโครงข่ายประสาทเทียมให้ผลลัพธ์ดีที่สุด โดยมีค่าความแม่นยำ 98.07% ความไว 99.15% ความจำเพาะ 96.21% และประสิทธิภาพโดยรวมที่ 98.47% ผลลัพธ์นี้ชี้ให้เห็นถึงศักยภาพของเทคนิคดังกล่าว ในการช่วยสนับสนุนกระบวนการวินิจฉัยโรคมะเร็งเต้านมในระยะเริ่มต้นได้อย่างมีนัยสำคัญ

คำสำคัญ : การทำเหมืองข้อมูล มะเร็งเต้านม การจำแนกประเภทข้อมูล

Abstract

This research aimed to evaluate the performance of various data mining techniques

*ที่อยู่ติดต่อ E-mail address: anupong.s@acc.msu.ac.th https://doi.org/10.55003/scikmitl.2025.267178

in constructing predictive models for breast cancer screening. Seven classification methods were compared, namely Neural networks, Support Vector Machine (SVM), Naïve Bayes, k-Nearest Neighbors (k-NN), Decision tree, Deep learning, and Ensemble vote. The dataset used in this study comprised 569 patient records obtained from the University of Wisconsin and made publicly available on www.kaggle.com. The analysis was conducted following the CRISP-DM process, which included variable selection, handling of missing data, and defining the roles of each attribute. The results revealed that the Neural Network technique yielded the best performance, achieving an accuracy of 98.07%, sensitivity of 99.15%, specificity of 96.21%, and an overall efficiency of 98.47%. These findings demonstrate the potential of this technique to significantly support the early detection and diagnosis of breast cancer.

Keywords: Data Mining, Breast Cancer, Classification

1. บทนำ

มะเร็งเต้านมเกิดจากความผิดปกติของเซลล์ท่อน้ำนม ซึ่งพบได้มากที่สุดถึงร้อยละ 80 ในขณะที่มะเร็งที่เกิดจากความผิดปกติของเซลล์ต่อมน้ำนมจะสามารถพบได้น้อยกว่าโดยพบประมาณ 10% (World Health Organization, 2024) โดยเซลล์ที่ผิดปกติจะเริ่มแบ่งตัวและลุกลามไปยังเนื้อเยื่อข้างเคียง ซึ่งหากไม่ได้รับการรักษาอาจแพร่กระจายไปยังเซลล์ส่วนอื่น ๆ ในร่างกายผ่านทางเดินน้ำเหลือง (Wacharaphapaiboon, 2021)

ปัจจุบันมะเร็งเต้านมถือเป็นปัญหาด้านสาธารณสุขที่สำคัญของผู้หญิงทั่วโลก จากข้อมูลขององค์การอนามัยโลก (WHO) ระบุว่าในแต่ละปีมีผู้ป่วยมะเร็งเต้านมรายใหม่มากถึง 2.3 ล้านรายทั่วโลก และมีผู้เสียชีวิตจากโรคนี้ประมาณ 685,000 รายในแต่ละปี (World Health Organization, 2024) สำหรับประเทศไทย สถิติจากกรมการแพทย์ กระทรวงสาธารณสุข รายงานว่ามีผู้ป่วยมะเร็งเต้านมรายใหม่เฉลี่ย 18,000 รายต่อปี และเสียชีวิตประมาณ 4,800 ราย โดยแนวโน้มดังกล่าวยังคงเพิ่มสูงขึ้น ดังจะเห็นได้จากสถิติปี พ.ศ. 2566 พบผู้ป่วยใหม่ประมาณ 22,000 รายต่อปี หรือเฉลี่ยวันละ 60 ราย (Hfocus, 2024) นอกจากนี้ ยังพบว่ามะเร็งเต้านมมีอัตราการกลับมาเป็นซ้ำหรือการแพร่กระจายของโรคที่สูงกว่ามะเร็งชนิดอื่น ๆ (Taiwiriyawet, 2023) สถานการณ์นี้สะท้อนให้เห็นถึงความจำเป็นในการตรวจคัดกรองที่รวดเร็ว แม่นยำ และมีประสิทธิภาพตั้งแต่ในระยะเริ่มต้น

อย่างไรก็ตาม ในการตรวจคัดกรองผู้ป่วยมะเร็งเต้านมด้วยวิธีการทางการแพทย์แบบเดิมอาจมีข้อจำกัดหลายประการ อาทิ ความล่าช้าในการวิเคราะห์ ความผิดพลาดในการวินิจฉัย หรือความไม่สม่ำเสมอของข้อมูลผู้ป่วย การประยุกต์ใช้เทคโนโลยีด้านข้อมูลจึงเป็นอีกแนวทางหนึ่งที่สามารถสนับสนุนการคัดกรองได้อย่างมีประสิทธิภาพ โดยเฉพาะเทคนิคการทำเหมืองข้อมูล (Data mining) ซึ่งสามารถสร้างแบบจำลองที่เรียนรู้จากข้อมูลขนาดใหญ่และจำแนกผู้ป่วยที่มีความเสี่ยงได้อย่างแม่นยำ

งานวิจัยโดย Srisuk & Thongkam (2021) รายงานว่าโครงข่ายประสาทเทียม (Neural network) ให้ผลการจำแนกผู้ป่วยมะเร็งเต้านมที่แม่นยำกว่าวิธีอื่น เช่น วิธีนาอีฟเบย์ (Naïve Bayes) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) และวิธีต้นไม้ตัดสินใจ (Decision tree) ในขณะที่ Prema & Jagadeesh (2023) แสดงให้เห็นถึงศักยภาพของเทคนิคโครงข่ายประสาทเทียมในการพยากรณ์โรคมะเร็งเต้านม ผลลัพธ์เหล่านี้สะท้อนแนวทางการใช้เทคนิคเหมืองข้อมูลหลากหลายแบบเพื่อเปรียบเทียบประสิทธิภาพของโมเดล ซึ่งสอดรับกับวัตถุประสงค์ของการศึกษานี้

จากปัญหาที่กล่าวมา การศึกษาครั้งนี้มุ่งเน้นการประเมินประสิทธิภาพของแบบจำลองที่ใช้สำหรับคัดกรองผู้ป่วยโรคมะเร็งเต้านม โดยใช้เทคนิคการจำแนกประเภท (Classification) ภายใต้กระบวนการทำเหมืองข้อมูลเป็นเครื่องมือในการวิเคราะห์ ทั้งนี้การดำเนินการวิจัยอยู่ภายใต้กรอบแนวทาง Cross-Industry Standard Process for Data Mining (CRISP-DM) ซึ่งประกอบด้วยขั้นตอนสำคัญ ได้แก่ การทำความเข้าใจธุรกิจ การทำความเข้าใจข้อมูล การเตรียมข้อมูล การพัฒนาแบบจำลอง การประเมินผล และการนำแบบจำลองไปประยุกต์ใช้งาน (Wirth & Hipp, 2000) สำหรับข้อมูลที่ใช้ในการศึกษาได้มาจากฐานข้อมูลของมหาวิทยาลัยวิสคอนซิน ซึ่งถูกรวบรวมไว้ในเว็บไซต์ www.kaggle.com ซึ่งมีจำนวนผู้ป่วยทั้งหมด 569 ราย (Muhammad, 2024) โดยแบบจำลองที่พัฒนาขึ้นมีเป้าหมายเพื่อช่วยลดข้อผิดพลาด เพิ่มความเร็วและความแม่นยำในการคัดกรอง และสนับสนุนการวินิจฉัยของแพทย์อย่างมีประสิทธิภาพ

2. วิธีดำเนินการวิจัย

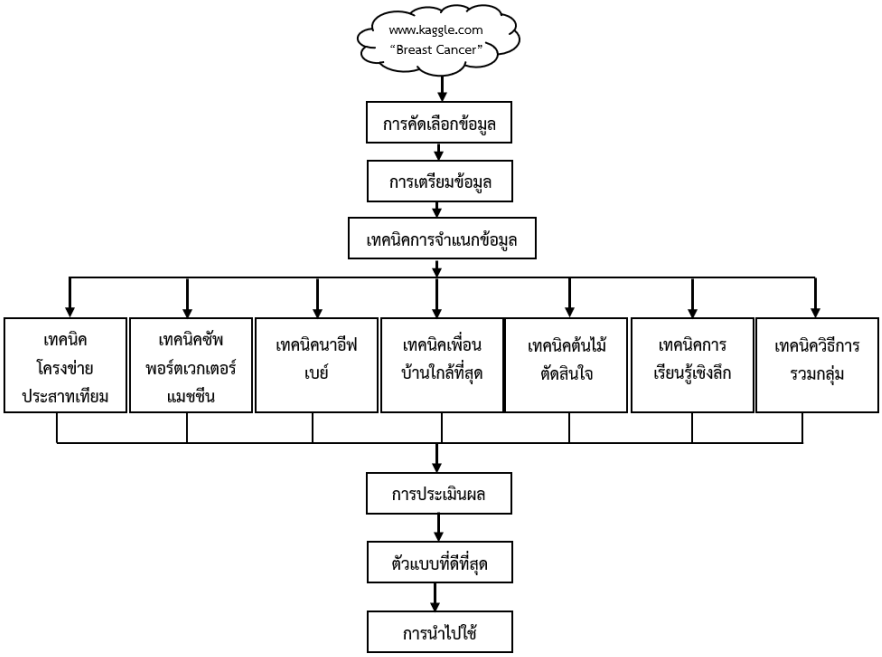
ข้อมูลที่ใช้ในการศึกษาครั้งนี้นำมาจากชุดข้อมูลที่เผยแพร่โดย Naira Saeed Muhammad แห่งมหาวิทยาลัยวิสคอนซิน ผ่านเว็บไซต์ www.kaggle.com ภายใต้ชื่อไฟล์ Breast Cancer (Muhammad, 2024) โดยผู้วิจัยได้ดำเนินการคัดกรองและปรับปรุงข้อมูลให้มีความเหมาะสมต่อการวิเคราะห์ เพื่อให้สามารถนำไปใช้พัฒนาแบบจำลองสำหรับการตรวจคัดกรองผู้ป่วยโรคมะเร็งเต้านมได้อย่างแม่นยำ ด้วยการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล กระบวนการดังกล่าวมีส่วนช่วยเพิ่มประสิทธิภาพและความแม่นยำในการคัดกรองผู้ป่วย โดยลำดับขั้นตอนการดำเนินงานวิจัยแสดงไว้ในรูปที่ 1

การดำเนินงานวิจัยในการทดสอบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูลสำหรับการพยากรณ์โรคมะเร็งเต้านมในครั้งนี้อ้างอิงจากแนวทางมาตรฐาน CRISP-DM ซึ่งเป็นกระบวนการหลักในการทำเหมืองข้อมูล ประกอบด้วย 6 ขั้นตอน แสดงดังรูปที่ 2

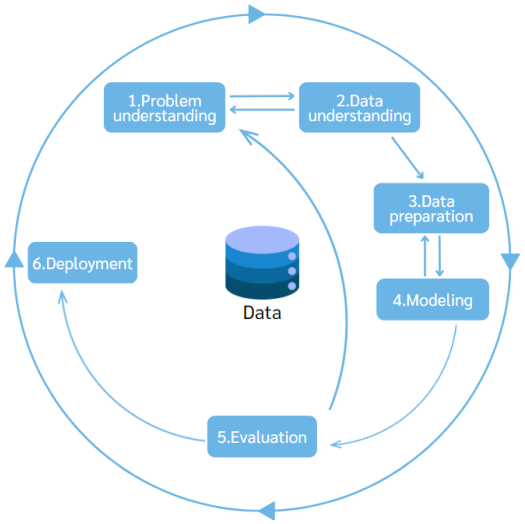
2.1 การทำความเข้าใจปัญหา (Problem understanding)

จากการศึกษาของผู้วิจัย พบว่า โรคมะเร็งเต้านมเป็นโรคที่พบได้บ่อยที่สุดในเพศหญิง และเป็นสาเหตุของการเสียชีวิตอันดับต้น ๆ ของผู้หญิงทั่วโลก ความเสี่ยงของโรคจะเพิ่มสูงขึ้นหากไม่ได้รับการตรวจคัดกรองอย่างสม่ำเสมอ เนื่องจากหากตรวจพบในระยะลุกลาม โรคอาจแพร่กระจายไปยังอวัยวะอื่น และทำให้โอกาสในการรักษาลดลงอย่างมีนัยสำคัญ การตรวจพบโรคในระยะแรกเป็นสิ่งสำคัญเพราะสามารถรักษาได้หายขาด แต่ปัญหาคือผู้ป่วยบางคนไม่รู้ตัวว่ามีโรคในระยะเริ่มต้นเนื่องจากอาการไม่ชัดเจน ดังนั้น การพัฒนาเทคโนโลยีเพื่อสนับสนุนการตรวจคัดกรองโรคมะเร็งเต้านมจึงมีความสำคัญอย่างยิ่ง โดยเฉพาะการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลร่วมกับกระบวนการจำแนกประเภท ซึ่งช่วยให้บุคลากรทางการ

แพทย์ตรวจพบผู้ที่มีความเสี่ยงสูงได้อย่างรวดเร็วและแม่นยำมากยิ่งขึ้น การสร้างแบบจำลองที่มีประสิทธิภาพสำหรับการคัดกรองจะช่วยเพิ่มโอกาสในการรักษาตั้งแต่ระยะแรกเริ่ม และลดอัตราการเสียชีวิตจากโรคดังกล่าวได้อย่างมีนัยสำคัญ



รูปที่ 1. ขั้นตอนการดำเนินงานวิจัย



รูปที่ 2. กระบวนการมาตรฐานการทำเหมืองข้อมูล (Wirth & Hipp, 2000)

2.2 การทำความเข้าใจเกี่ยวกับข้อมูล (Data understanding)

ข้อมูลที่ใช้ในการศึกษาครั้งนี้นำมาจากเว็บไซต์ www.kaggle.com ซึ่งอยู่ในรูปแบบไฟล์ CSV ภายใต้ชื่อ Breast Cancer (Muhammad, 2024) ซึ่งประกอบด้วยข้อมูลผู้ป่วยจำนวน 569 ราย และตัวแปรที่ใช้ในการวิเคราะห์ทั้งหมด 12 ตัวแปร รายละเอียดของชุดข้อมูลแสดงไว้ในตารางที่ 1

ตารางที่ 1. ข้อมูลที่ใช้ในการวิเคราะห์

ลำดับ	ตัวแปร	คำอธิบาย	ประเภทข้อมูล
1	ID number	รหัสผู้ป่วย (ID)	String
2	radius	ค่าเฉลี่ยของระยะห่างจากจุดศูนย์กลางไปยังจุดบนขอบ	real
3	texture	ลักษณะพื้นผิว	real
4	perimeter	เส้นรอบวง	real
5	area	พื้นที่	real
6	smoothness	ความเรียบ	real
7	compactness	ความแน่น	real
8	concavity	ความโค้งเข้าของขอบ	real
9	concave points	จำนวนที่เว้า	real
10	symmetry	ความสมมาตร	real
11	Fractal dimension	การวิเคราะห์เซลล์เนื้อเยื่อ	real
12	Diagnosis	ผลวินิจฉัย (Label)	binominal

2.3 การเตรียมข้อมูล (Data preparation)

ในกระบวนการเตรียมข้อมูล ผู้วิจัยได้ดำเนินการตรวจสอบข้อมูลเพื่อให้มีความถูกต้องและครบถ้วน โดยแบ่งออกเป็น 3 ขั้นตอน ดังนี้

2.3.1 การคัดเลือกข้อมูล (Data selection) ผู้วิจัยได้ศึกษาปัจจัยที่อาจส่งผลต่อการเกิดโรคมะเร็งเต้านม และได้คัดเลือกตัวแปรที่เกี่ยวข้องมาใช้ในการวิเคราะห์ โดยแบ่งออกเป็น ตัวแปรอิสระ (Independent variables) จำนวน 11 ตัว ดังนี้ 1) รหัสผู้ป่วย 2) ค่าเฉลี่ยของระยะห่างจากจุดศูนย์กลางไปยังจุดบนขอบ 3) ลักษณะพื้นผิว 4) เส้นรอบวง 5) พื้นที่ 6) ความเรียบ 7) ความแน่น 8) ความโค้งเข้าของขอบ 9) จำนวนที่เว้า 10) ความสมมาตรและ 11) การวิเคราะห์เซลล์เนื้อเยื่อ และกำหนดตัวแปรตาม (Dependent variable) คือ ผลวินิจฉัย

2.3.2 การกลั่นกรองข้อมูล (Data cleaning) ภายหลังการสำรวจข้อมูล (Data exploration) พบว่ามีค่าที่ไม่ทราบแน่ชัดปรากฏในบางแอตทริบิวต์ ได้แก่ แอตทริบิวต์ค่าลักษณะพื้นผิว (texture) และค่าความเรียบ (smoothness) ซึ่งแสดงผลเป็นค่า “N/A” ผู้วิจัยจึงพิจารณาให้ค่าดังกล่าวเป็นข้อมูลสูญหาย (Missing values) และตัดออกจากกระบวนการวิเคราะห์เพื่อหลีกเลี่ยงความคลาดเคลื่อนในการสร้างแบบจำลอง

2.3.3 การกำหนดหน้าที่ให้กับแอตทริบิวต์รหัสผู้ป่วย (ID number) หน้าที่ของแอตทริบิวต์นี้ใช้เป็น ID เพื่อระบุข้อมูลแต่ละรายการไม่ซ้ำกัน โดย ID นี้ทำหน้าที่ระบุตัวตนของข้อมูลในแต่ละแถวในระบบเพื่อป้องกันความซ้ำซ้อน ซึ่งข้อมูลในแอตทริบิวต์นี้จะไม่ถูกนำมาใช้ในกระบวนการวิเคราะห์ในครั้งนี้ เพราะมีหน้าที่หลักเป็นตัวระบุข้อมูลมากกว่าการเป็นข้อมูลเชิงวิเคราะห์ และกำหนดหน้าที่ให้กับแอตทริบิวต์ผลวินิจฉัย (Diagnosis) แอตทริบิวต์นี้มีหน้าที่เป็น Label เพื่อระบุคลาสของข้อมูลว่าเป็น “M” แสดงว่าผิดปกติ หรือ “B” แสดงว่าไม่พบความผิดปกติ โดยค่าของ Label นี้จะใช้ในการแบ่งประเภทหรือระบุคลาสคำตอบในการวิเคราะห์

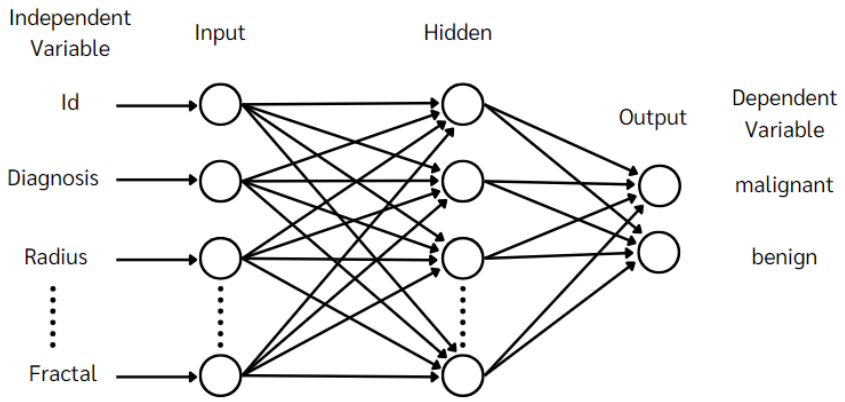
2.4 การสร้างแบบจำลอง (Modeling)

ในการสร้างแบบจำลองและประเมินประสิทธิภาพของแต่ละเทคนิค ผู้วิจัยได้ใช้โปรแกรม RapidMiner Studio Version 10.3 เป็นเครื่องมือหลักในการพัฒนาและวิเคราะห์ข้อมูล โดยประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลจำนวน 7 เทคนิค ได้แก่ โครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ เพื่อนบ้านใกล้ที่สุด ต้นไม้ตัดสินใจ การเรียนรู้เชิงลึก และวิธีการรวมกลุ่ม นอกจากนี้ ผู้วิจัยได้ดำเนินการปรับแต่งค่าพารามิเตอร์ (Hyperparameter optimization) ของแต่ละโมเดลอย่างเป็นระบบผ่านกระบวนการ Optimization เพื่อค้นหาค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับการเรียนรู้ของแบบจำลอง โดยมีวัตถุประสงค์เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์และคัดกรองผู้ป่วยโรคมะเร็งเต้านม ทั้งนี้กระบวนการดังกล่าวช่วยลดความคลาดเคลื่อนของการทำนายและเพิ่มความแม่นยำของผลลัพธ์อย่างมีนัยสำคัญ (Sukprasert, 2023) รายละเอียดแต่ละเทคนิค แสดงดังนี้

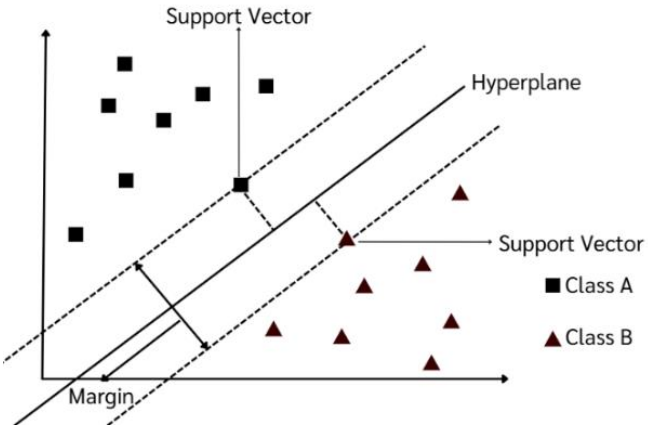
1) เทคนิคโครงข่ายประสาทเทียม เป็นแบบจำลองทางคณิตศาสตร์ที่เลียนแบบกระบวนการทำงานของโครงข่ายเซลล์ประสาทในสมองมนุษย์ ซึ่งมีความสามารถในการเรียนรู้ การจดจำรูปแบบ และการคาดการณ์ข้อมูล โดยมีพื้นฐานจากการศึกษาข่ายงานไฟฟ้าในสมองที่ประกอบด้วยเซลล์ประสาท (นิวรอน) ซึ่งเชื่อมต่อกันผ่านจุดประสานประสาท (Synapses) การทำงานของนิวรอนจะมีการรับสัญญาณผ่าน “เดนไดรท์ (Dendrite)” และส่งสัญญาณออกผ่าน “แอคซอน (Axon)” ซึ่งเป็นการประมวลผลข้อมูล โครงข่ายประสาทเทียมจะเรียนรู้โดยการปรับน้ำหนักของการเชื่อมต่อระหว่างนิวรอนเมื่อผลลัพธ์ที่ทำนายไม่ตรงกับข้อมูลจริง ซึ่งทำให้ระบบสามารถปรับตัวและเพิ่มความแม่นยำได้เมื่อได้รับข้อมูลใหม่ ดังแสดงในรูปที่ 3 การประยุกต์ใช้งานของโครงข่ายประสาทเทียม ได้แก่ การจำแนกข้อมูล การประมวลผลภาพ การรู้จำเสียง และการวินิจฉัยโรค ทำให้สามารถวิเคราะห์และทำนายผลได้อย่างมีประสิทธิภาพในหลายสาขา (Siphating et al., 2023)

2) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน ทำงานโดยการนำค่าของกลุ่มข้อมูลมาวางลงในฟีเจอร์สเปซ (Feature space) และหาเส้นแบ่ง (Hyperplane) ที่สามารถแยกข้อมูลทั้งสองกลุ่มออกจากกันได้ โดยเส้นแบ่งนี้จะถูกเลือกให้เป็นเส้นที่ดีที่สุดในการแยกข้อมูล สำหรับข้อมูลเชิงเส้น SVM สามารถใช้งานได้โดยตรง แต่ในกรณีข้อมูลที่ไม่เป็นเชิงเส้น จะมีการใช้ Kernel Function เพื่อทำให้ข้อมูลสามารถแยกได้ง่ายขึ้นบนระนาบหลายมิติ ในกระบวนการนี้จะมีการใช้โครงสร้างในการคัดเลือก (Feature selection) เพื่อหาเวกเตอร์ที่เหมาะสม ซึ่งเรียกว่า Support Vectors ตัวแบบ SVM มีเป้าหมายในการแยกกลุ่มของเวกเตอร์

ให้ชัดเจน โดยให้กลุ่มหนึ่งอยู่ฝั่งหนึ่งของระนาบและอีกกลุ่มอยู่ฝั่งตรงข้าม (Tongkunwong & Sawatkamon, 2024) ดังแสดงในรูปที่ 4



รูปที่ 3. เทคนิคโครงข่ายประสาทเทียม



รูปที่ 4. เทคนิคซัพพอร์ตเวกเตอร์แมชชีน

3) เทคนิคนาอีฟเบย์ เป็นเทคนิคในการจำแนกประเภทที่ใช้หลักการความน่าจะเป็น โดยอ้างอิงกฎของเบย์ (Bayes' Law) เพื่อประเมินว่าสมมติฐานใดมีโอกาสถูกต้องมากที่สุด โดยแนวคิดนี้มองว่า ปริมาณที่สนใจอยู่ภายใต้การแจกแจงความน่าจะเป็น (Probability distribution) ทำให้สามารถใช้ความน่าจะเป็นของตัวอย่างต่าง ๆ เพื่อประกอบการตัดสินใจที่มีเหตุผลในงานจำแนกประเภท กฎของเบย์หรือทฤษฎีบทของเบย์ (Bayes' Theorem) ตั้งชื่อตามโทมัส เบย์ (Thomas Bayes) นักสถิติชาวอังกฤษ อธิบายความน่าจะเป็นแบบมีเงื่อนไข โดยเชื่อมโยงเหตุการณ์ปัจจุบันกับสิ่งที่เคยเกิดขึ้นก่อนหน้า ทำให้สามารถเพิ่มความเชื่อมั่นในสมมติฐานต่าง ๆ ได้ เมื่อมีข้อมูลใหม่ โดยมีความน่าจะเป็นแบบมีเงื่อนไขแสดงดังสมการที่ (1) และการใช้กฎของเบย์ในการจำแนกประเภท (Phikulsri & Chanamarn, 2023) แสดงดังสมการที่ (2)

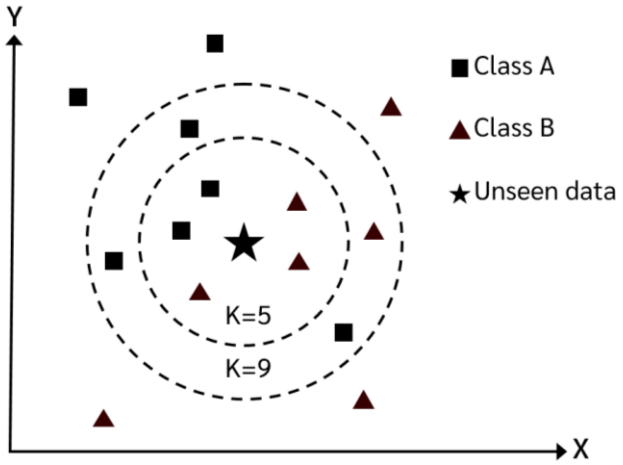
$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (1)$$

สามารถแสดงการคำนวณโดยใช้ Bayes' Theorem ได้ดังสมการที่ (2)

$$P(h|D_1, D_{2\ldots}, D_n = \frac{P(D_1|h) * P(D_2|h) \ldots P(D_n h) * P(h)}{P(D_1|h) * P(D_2|h) \ldots * P(D_n|h)} \quad (2)$$

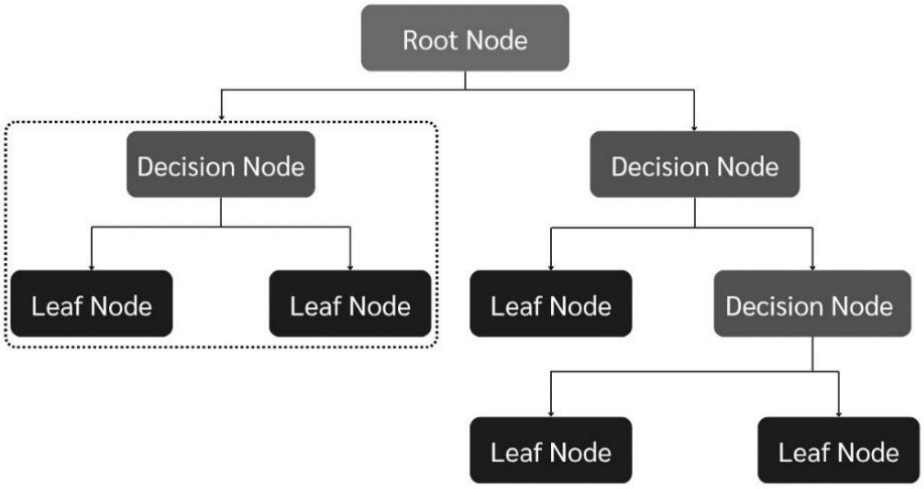
เมื่อ $P(D)$ หมายถึง ความน่าจะเป็นก่อน (Prior probability) ของเซตตัวอย่างฝึกฝน D
$P(h)$ หมายถึง ความน่าจะเป็นก่อนของสมมติฐาน $h \in H$
$P(h|D)$ หมายถึง ความน่าจะเป็นภายหลัง (Posterior probability) ของสมมติฐาน h เมื่อกำหนดเซตตัวอย่าง ฝึกฝน D
$P(D|h)$ หมายถึง ความน่าจะเป็นภายหลังของเซตตัวอย่างฝึกฝน D เมื่อกำหนดสมมติฐาน h

4) เทคนิคเพื่อนบ้านใกล้ที่สุด เป็นเทคนิคการจำแนกประเภทข้อมูลที่มีขั้นตอนการทำงานไม่ซับซ้อนและเข้าใจง่าย โดยใช้หลักการเปรียบเทียบความคล้ายคลึงกันระหว่างข้อมูลที่ต้องการจำแนกกับข้อมูลที่มีอยู่เดิมในชุดข้อมูล หลักการทำงานของ k-NN คือ การเลือกข้อมูลที่ใกล้เคียงมากที่สุด k ตัว จากข้อมูลทั้งหมด โดยพิจารณาความคล้ายคลึงหรือระยะห่างระหว่างข้อมูล ซึ่งจะใช้เป็นเกณฑ์ในการกำหนดว่าข้อมูลที่นำมาพิจารณาควรมีผลลัพธ์เหมือนกับกลุ่มข้อมูลใกล้เคียงที่สุดจำนวน k ตัวที่เลือกไว้ (Sukprasert, 2023) ซึ่งขั้นตอนการทำงานของเทคนิคเพื่อนบ้านใกล้ที่สุด ดังแสดงในรูปที่ 5



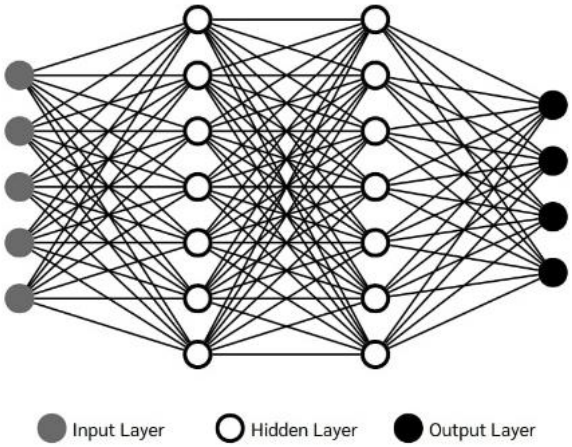
รูปที่ 5. เทคนิคเพื่อนบ้านใกล้ที่สุด

5) เทคนิคต้นไม้ตัดสินใจ เป็นการจำแนกประเภทที่แสดงผลลัพธ์ในรูปแบบโครงสร้างต้นไม้ โดยข้อมูลที่ต้องการจัดกลุ่มจะถูกนำไปเปรียบเทียบกับคุณลักษณะ (Attribute) ต่าง ๆ ภายในต้นไม้ จนกระทั่งถึงคลาสปลายทางซึ่งแทนกลุ่มของข้อมูลที่มีลักษณะคล้ายกัน โครงสร้างของต้นไม้ประกอบด้วย โหนด (Node) ซึ่งแต่ละโหนดมีคุณสมบัติที่ใช้สำหรับการทดสอบข้อมูล กิ่ง (Branch) แสดงค่าที่เป็นไปได้ของคุณสมบัตินั้น ส่วนใบ (Leaf) แสดงกลุ่มของข้อมูลหรือผลลัพธ์จากการทำนาย โดยโหนดที่อยู่สูงสุดในโครงสร้างเรียกว่า โหนดราก (Root node) ซึ่งเป็นจุดเริ่มต้นของการตัดสินใจในต้นไม้ตัดสินใจนี้ (Rideach et al., 2022) ขั้นตอนการทำงานขอเทคนิคต้นไม้ตัดสินใจ ดังแสดงในรูปที่ 6



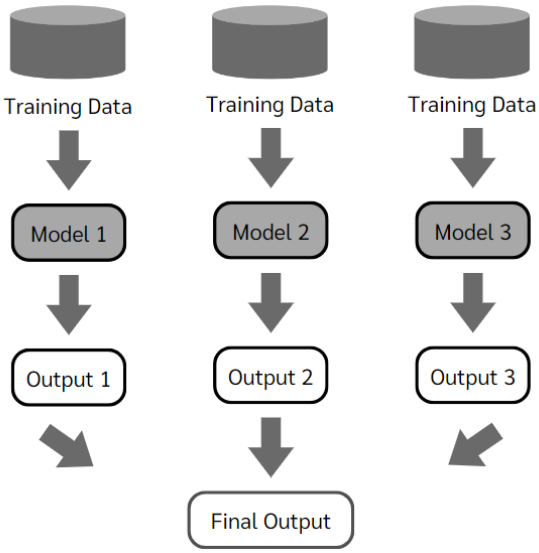
รูปที่ 6. เทคนิคต้นไม้ตัดสินใจ

6) เทคนิคการเรียนรู้เชิงลึก เป็นหนึ่งในเทคนิคของโครงข่ายประสาทเทียมที่มีโครงสร้างซับซ้อน ประกอบด้วยนิวรอนและชั้นซ่อนจำนวนมาก โดยอัลกอริธึมนี้ถูกออกแบบมาเพื่อการเรียนรู้ของเครื่องจักร ซึ่งต่างจากโครงข่ายประสาทเทียมที่มีเพียงเลเยอร์เดียว การเรียนรู้เชิงลึกประกอบไปด้วยหลาย Hidden layer ที่แต่ละชั้นจะทำหน้าที่ประมวลผลข้อมูลด้วยเซลล์ประสาทจำนวนมาก โดยเลเยอร์แรก (Input layer) จะรับข้อมูลเข้ามาและส่งข้อมูลที่ประมวลผลแล้วไปยังเลเยอร์สุดท้าย (Output layer) กระบวนการนี้ช่วยให้ข้อมูลในแต่ละเลเยอร์มีการประมวลผลที่แตกต่างกัน เช่น ค่าถ่วงน้ำหนัก (Weight) ค่าความเอนเอียง (Bias) และฟังก์ชันการกระตุ้น (Activation function) ซึ่งช่วยให้แต่ละเลเยอร์สามารถสกัดลักษณะของข้อมูลที่ซับซ้อนขึ้นเรื่อย ๆ ทำให้โมเดลสามารถตัดสินใจได้อย่างแม่นยำและใกล้เคียงกับการตัดสินใจของมนุษย์ (Kumjit et al., 2022) ดังแสดงในรูปที่ 7



รูปที่ 7. เทคนิคการเรียนรู้เชิงลึก

7) เทคนิควิธีการรวมกลุ่ม เป็นเทคนิคการเรียนรู้ของเครื่อง (Machine learning) ที่รวมแบบจำลองการจำแนกประเภทข้อมูลพื้นฐานหลายตัวเข้าด้วยกัน ซึ่งอาจเกิดจากการผสานเทคนิคการจำแนกข้อมูลตั้งแต่ 2 วิธีขึ้นไป เพื่อสร้างโมเดลที่มีความเหมาะสมและแม่นยำมากยิ่งขึ้นในการทำนายผลลัพธ์ โดยประสิทธิภาพของวิธีนี้จะขึ้นอยู่กับระดับความหลากหลายและความถูกต้องของตัวจำแนกแต่ละตัว การรวมกลุ่มช่วยลดข้อผิดพลาดที่เกิดจากความแปรปรวน (Variance) และเพิ่มความน่าเชื่อถือของผลลัพธ์ แนวคิดหลักของการเรียนรู้แบบนี้ คือ การรวมแบบจำลองหลายตัวเข้าด้วยกันเพื่อจัดการกับปัญหาเดียวกัน ซึ่งโดยทั่วไปจะให้ผลลัพธ์ที่แม่นยำกว่าการใช้แบบจำลองเดียว (Sukprasert, 2023) ดังแสดงในรูปที่ 8



รูปที่ 8. เทคนิควิธีการรวมกลุ่ม

สำหรับการสร้างตัวแบบเพื่อคัดกรองผู้ป่วยโรคมะเร็งเต้านมในครั้งนี้ ผู้วิจัยได้ดำเนินการปรับแต่งแบบจำลองโดยใช้เทคนิคการหาค่าที่เหมาะสมที่สุด (Optimization) เพื่อค้นหาชุดพารามิเตอร์ที่ส่งผลต่อประสิทธิภาพสูงสุดในแต่ละเทคนิค (Ruangsawud et al., 2023) โดยในกรณีของเทคนิคโครงข่ายประสาทเทียม มีการกำหนด Hidden layer size เท่ากับ 5 Training cycles เท่ากับ 200 Learning rate เท่ากับ 0.01 และ Momentum เท่ากับ 0.9 ส่งผลให้โมเดลมีประสิทธิภาพสูงสุดในการทำนาย สำหรับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน ค่าที่เหมาะสม ได้แก่ Kernel cache เท่ากับ 200 และค่าคงที่ C เท่ากับ 10 เทคนิคนาอีฟเบย์ได้ทำการปรับประสิทธิภาพด้วยการใช้ Laplace correction เพื่อป้องกันปัญหาค่าความน่าจะเป็นเป็นศูนย์ในกรณีที่ไม่มีตัวอย่างในบางคลาส ในกรณีของเทคนิคเพื่อนบ้านใกล้ที่สุด พบว่า ค่า k ที่เหมาะสมที่สุด คือ 5 และเทคนิคต้นไม้ตัดสินใจให้ประสิทธิภาพสูงสุด เมื่อกำหนดค่าความลึกของต้นไม้ (Maximal depth) เท่ากับ 10

2.5 การประเมินผล (Evaluation)

ในการประเมินประสิทธิภาพของแบบจำลอง ผู้วิจัยได้นำวิธี 10-fold cross-validation มาใช้ โดยแบ่งชุดข้อมูลออกเป็น 10 ส่วนเท่า ๆ กัน จากนั้นทำการประเมินแบบจำลองจำนวน 10 รอบ โดยในแต่ละรอบข้อมูลจะเป็นได้ทั้งชุดทดสอบ (Test set) และชุดฝึกสอน (Training set) วิธีนี้ทำให้ข้อมูลทุกตัวอย่างในชุดข้อมูลมีโอกาสถูกใช้ทั้งในการฝึก (Training) และในการทดสอบ (Testing) อย่างน้อยหนึ่งครั้ง ช่วยลดความเอนเอียงที่อาจเกิดจากการแบ่งข้อมูลแบบสุ่มเพียงครั้งเดียว และส่งผลให้ค่าประเมินที่ได้มีความแม่นยำและน่าเชื่อถือมากขึ้น สำหรับการวัดประสิทธิภาพของแบบจำลอง ผู้วิจัยใช้ตัวชี้วัดหลัก ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความไว (Sensitivity) ค่าความจำเพาะ (Specificity) และค่าประสิทธิภาพโดยรวม (F-measure) (Sukprasert, 2023) โดยค่าความไวที่สูงแสดงถึงความสามารถของโมเดลในการตรวจจับผู้ป่วยที่มีโรคได้อย่างถูกต้อง ในขณะที่ค่าจำเพาะบ่งบอกถึงความสามารถในการระบุผู้ที่ไม่มีโรคเพื่อลดการวินิจฉัยเกินจำเป็น ค่าทั้งสองจึงมีความสำคัญร่วมกันในการประเมินประสิทธิภาพของระบบคัดกรอง รายละเอียดตัวชี้วัด แสดงได้ดังต่อไปนี้

1) ค่าความแม่นยำ คือ จำนวนข้อมูลที่ทำนายถูกของทุกคลาส คำนวณได้ดังสมการที่ (3)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

2) ค่าความไว หมายถึง ความสามารถของโมเดลในการตรวจจับผู้ป่วยที่มีโรคได้อย่างถูกต้อง โดยแสดงสัดส่วนของผู้ป่วยที่เป็นโรคจริงซึ่งถูกพยากรณ์ว่าเป็นโรคได้อย่างถูกต้อง คำนวณได้จากสมการที่ (4)

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

3) ค่าจำเพาะ หมายถึง ความสามารถของโมเดลในการระบุผู้ที่ไม่ได้ป่วยว่าไม่ป่วยได้อย่างถูกต้อง โดยคำนวณจากสัดส่วนของผู้ที่ไม่มีโรคจริงซึ่งถูกจำแนกได้อย่างถูกต้อง คำนวณได้จากสมการที่ (5)

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

4) ค่าประสิทธิภาพโดยรวม คือ ค่าที่ได้จากการผสานระหว่างค่า Precision และ Recall เพื่อประเมินความแม่นยำโดยรวมของโมเดลในการจำแนกข้อมูล สามารถคำนวณได้จากสมการที่ (6) และ (7)

$$F-measure \text{ คลาสเป้าหมาย } YES = \frac{(2 * Precision(YES) * Recall\ (YES))}{(Precision(YES) + Recall\ (YES))} \quad (6)$$

$$F-measure \text{ คลาสเป้าหมาย } NO = \frac{(2*Precision(NO) * Recall\ (NO))}{(Precision(NO) + Recall\ (NO))} \quad (7)$$

เพื่อประเมินประสิทธิภาพของแบบจำลอง ผู้วิจัยได้ใช้ตาราง Confusion Matrix ดังแสดงในตารางที่ 2 ซึ่งช่วยให้สามารถระบุค่าความแม่นยำ ค่าความไว ค่าความจำเพาะ ของแบบจำลองได้อย่างชัดเจน

ตารางที่ 2. Confusion Matrix

	ผลทำนายว่าเป็นมะเร็ง (YES)	**ผลทำนายว่าไม่เป็นมะเร็ง (NO)**
ผลตรวจว่าเป็นมะเร็ง (YES)	True Positive (TP)	False Negative (FN)
ผลตรวจว่าไม่เป็นมะเร็ง (NO)	False Positive (FP)	True Negative (TN)

เมื่อ True Positive (TP) เกิดขึ้นเมื่อข้อมูลจริงเป็น YES และระบบทำนายได้ถูกต้องว่าเป็น YES
False Negative (FN) เกิดขึ้นเมื่อข้อมูลจริงเป็น YES แต่ระบบทำนายผิดว่าเป็น NO
True Negative (TN) เกิดขึ้นเมื่อข้อมูลจริงเป็น NO และระบบทำนายได้ถูกต้องว่าเป็น NO
False Positive (FP) เกิดขึ้นเมื่อข้อมูลจริงเป็น NO แต่ระบบทำนายผิดว่าเป็น YES

2.6 การนำไปใช้งาน (Deployment)

จากการดำเนินงานตามกระบวนการมาตรฐานของการทำเหมืองข้อมูล พบว่า เทคนิคโครงข่ายประสาทเทียมเป็นเทคนิคที่เหมาะสมที่สุด สำหรับการสร้างแบบจำลองที่ใช้ในการคัดกรองผู้ป่วยโรคมะเร็งเต้านม โดยแบบจำลองที่มีความแม่นยำสูงช่วยให้การประเมินความเสี่ยงของผู้ป่วยเป็นไปอย่างรวดเร็วและเชื่อถือได้ ส่งผลให้แพทย์สามารถตรวจพบความผิดปกติตั้งแต่ระยะเริ่มต้น ซึ่งมีบทบาทสำคัญในการลดความรุนแรงของโรค ป้องกันการลุกลาม และเพิ่มโอกาสในการรักษาให้หายขาด

นอกจากนี้ การใช้แบบจำลองที่มีประสิทธิภาพยังช่วยลดภาระค่าใช้จ่ายในการรักษาโรคในระยะลุกลาม รวมถึงส่งเสริมให้ระบบบริการสุขภาพสามารถจัดสรรทรัพยากรทางการแพทย์ได้อย่างมีประสิทธิผลมากขึ้น ทำให้สามารถดูแลผู้ป่วยกลุ่มเสี่ยงอื่น ๆ ได้อย่างทันท่วงที ทั้งยังช่วยให้กระบวนการคัดกรองและวินิจฉัยของแพทย์มีความแม่นยำและคล่องตัวมากยิ่งขึ้น

3. ผลการวิจัยและอภิปรายผล

ผู้วิจัยได้ทำการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกผู้ป่วยโรคมะเร็งเต้านม จำนวน 7 เทคนิค ได้แก่ เทคนิคโครงข่ายประสาทเทียม เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคนาอีฟเบย์ เทคนิคเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ เทคนิคการเรียนรู้เชิงลึก และเทคนิคการรวมกลุ่ม โดยทำการทดสอบประสิทธิภาพของแบบจำลองด้วยตัวชี้วัด 4 ค่า ได้แก่ ค่าความแม่นยำ ค่าความไว ค่าจำเพาะ และค่าประสิทธิภาพโดยรวม ผลการวิเคราะห์ประสิทธิภาพของแบบจำลองสำหรับคัดกรองผู้ป่วยมะเร็งเต้านมตามเทคนิคต่าง ๆ แสดงดังตารางที่ 3 ซึ่งให้รายละเอียดเชิงเปรียบเทียบของแต่ละเทคนิคเพื่อช่วยในการเลือกแบบจำลองที่เหมาะสมที่สุดสำหรับการคัดกรองผู้ป่วยโรคมะเร็งเต้านม

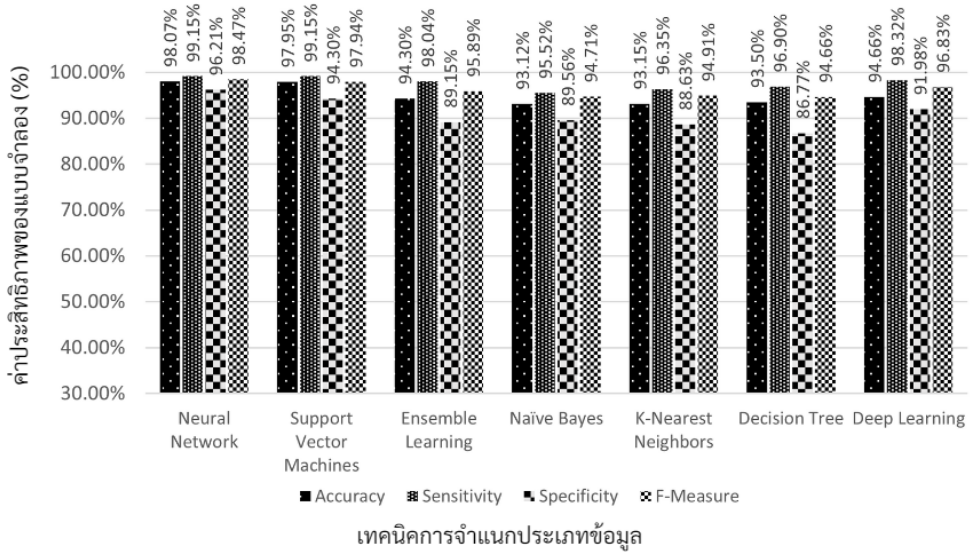
ตารางที่ 3. การเปรียบเทียบประสิทธิภาพของแบบจำลองจำแนกประเภทในการคัดกรองผู้ป่วยโรคมะเร็งเต้านม

เทคนิคการจำแนกประเภทข้อมูล	ค่าทดสอบประสิทธิภาพการจำแนกประเภทข้อมูล			
	ค่าความแม่นยำ	ค่าความไว	ค่าจำเพาะ	ค่าประสิทธิภาพโดยรวม
เทคนิคโครงข่ายประสาทเทียม*	**98.07%**	**99.15%**	**96.21%**	**98.47%**
เทคนิคซัพพอร์ตเวกเตอร์แมชชีน	97.36%	99.15%	94.30%	97.94%
เทคนิคการรวมกลุ่ม	94.73%	98.04%	89.15%	95.89%
เทคนิคนาอีฟเบย์	93.31%	95.52%	89.56%	94.71%
เทคนิคเพื่อนบ้านใกล้ที่สุด	93.49%	96.35%	88.63%	94.91%
เทคนิคต้นไม้ตัดสินใจ	93.15%	96.90%	86.77%	94.66%
เทคนิคการเรียนรู้เชิงลึก	95.96%	98.32%	91.98%	96.83%

* หมายถึง เทคนิคที่เหมาะสมที่สุดในการพัฒนาแบบจำลองสำหรับการคัดกรองผู้ป่วยโรคมะเร็งเต้านม

จากตารางที่ 3 พบว่า เทคนิคโครงข่ายประสาทเทียมให้ผลการจำแนกประเภทข้อมูลที่ดีที่สุด โดยมีค่าความแม่นยำเท่ากับ 98.07% รองลงมา คือ เทคนิคซัพพอร์ตเวกเตอร์แมชชีนที่มีค่าความแม่นยำเท่ากับ 97.36% ในขณะที่เทคนิคต้นไม้ตัดสินใจให้ค่าความแม่นยำต่ำที่สุดที่ 93.15% สำหรับค่าความไว ที่สะท้อนความสามารถในการระบุผู้ป่วยที่เป็นโรคได้อย่างถูกต้อง พบว่า เทคนิคโครงข่ายประสาทเทียมและเทคนิคซัพพอร์ตเวกเตอร์แมชชีนให้ค่าเท่ากันที่ 99.15% รองลงมา คือ เทคนิคการเรียนรู้เชิงลึกที่ 98.32% และเทคนิคนาอีฟเบย์ให้ค่าน้อยที่สุดที่ 95.52% และค่าจำเพาะที่แสดงถึงความสามารถในการจำแนกผู้ที่ไม่เป็นโรค พบว่า เทคนิคโครงข่ายประสาทเทียมให้ค่าสูงสุดที่ 96.21% ตามด้วย เทคนิคซัพพอร์ตเวกเตอร์แมชชีน

ที่ 94.30% ในขณะที่เทคนิคต้นไม้ตัดสินใจให้ค่าน้อยที่สุดที่ 86.77% เมื่อพิจารณาค่าประสิทธิภาพโดยรวม ซึ่งสะท้อนภาพรวมของความแม่นยำและความครอบคลุมแบบจำลอง พบว่า เทคนิคโครงข่ายประสาทเทียม ยังคงมีค่าสูงสุดที่ 98.47% รองลงมา คือ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน ที่ 97.94% และเทคนิคต้นไม้ตัดสินใจให้ค่าน้อยที่สุดที่ 94.66% ผลการเปรียบเทียบการทดสอบประสิทธิภาพแสดงดังรูปที่ 9



รูปที่ 9. เปรียบเทียบการทดสอบประสิทธิภาพของการจำแนกข้อมูลของตัวแบบคัดกรองผู้ป่วยมะเร็งเต้านม

ผลการเปรียบเทียบพบว่า โครงข่ายประสาทเทียมมีประสิทธิภาพสูงสุดในทุกตัวชี้วัด อันเป็นผลจากความสามารถในการเรียนรู้รูปแบบข้อมูลที่ไม่เป็นเชิงเส้นและมีความซับซ้อนสูง ประกอบกับโครงสร้างเครือข่ายและการปรับพารามิเตอร์ที่เหมาะสม (Hidden layer size = 5, Learning rate = 0.01 และ Momentum = 0.9) ซึ่งสอดคล้องกับลักษณะตัวแปรเชิงปริมาณของชุดข้อมูล ทำให้โมเดลสามารถจำแนกผู้ป่วยได้ครอบคลุมทั้งการตรวจพบผู้ป่วยและลดโอกาสวินิจฉัยเกินจำเป็น ซึ่งผลลัพธ์ดังกล่าวสอดคล้องกับผลการศึกษาในงานของ Srisuk & Thongkam (2021) ซึ่งรายงานว่าโครงข่ายประสาทเทียมให้ความแม่นยำที่ 97.66% ในการจำแนกข้อมูลของผู้ป่วยกลุ่มเดียวกัน เช่นเดียวกับการศึกษาของ Prema & Jagadeesh (2023) ซึ่งเปรียบเทียบระหว่างโครงข่ายประสาทเทียมและซัพพอร์ตเวกเตอร์แมชชีน โดยพบว่า เทคนิคโครงข่ายประสาทเทียมให้ผลลัพธ์ที่ดีกว่า ด้วยค่าความแม่นยำเท่ากับ 96.60% ขณะที่ Kumjit et al. (2022) ได้นำเทคนิคดังกล่าวมาใช้ในการพยากรณ์การกลับมาเป็นซ้ำของโรค และได้ค่าความแม่นยำอยู่ที่ 79.00%

ในการประเมินประสิทธิภาพของแบบจำลองจำแนกผู้ป่วยโรคมะเร็งเต้านม ผู้วิจัยใช้วิธีตรวจสอบแบบไขว้ 10 เท่า (10-fold cross-validation) ร่วมกับการวิเคราะห์ตารางความสับสน (Confusion

matrix) ของเทคนิคโครงข่ายประสาทเทียม เพื่อระบุจำนวนตัวอย่างที่จำแนกได้ถูกต้องและผิดพลาด ทั้งในกลุ่มผู้ป่วยและผู้ไม่ป่วย โดยมีรายละเอียดแสดงในตารางที่ 4

ตารางที่ 4. ผลลัพธ์ Confusion matrix จากการทดสอบเทคนิคโครงข่ายประสาทเทียมด้วยวิธีตรวจสอบแบบไขว้ 10 เท่า (10-fold cross validation)

	ทำนายว่าเป็นมะเร็ง (YES)	**ทำนายว่าไม่เป็นมะเร็ง (NO)**	**รวม**
ผลการตรวจว่าเป็นมะเร็ง (YES)	210	2	**212**
ผลการตรวจว่าไม่เป็นมะเร็ง (NO)	11	346	**357**
รวม	**221**	**348**	**569**

จากตารางที่ 4 แสดงผลลัพธ์ของ Confusion matrix ซึ่งมีจำนวน False Positive (FP) และ False Negative (FN) ต่ำที่สุดเมื่อเทียบกับเทคนิคอื่น ซึ่งเป็นดัชนีสำคัญต่อการคัดกรองโรคในทางคลินิก เนื่องจากช่วยลดความผิดพลาดในการตัดสินใจรักษา โดยใช้ข้อมูลจากผู้ป่วยจำนวน 569 ราย พบว่า แบบจำลองโครงข่ายประสาทเทียมสามารถจำแนกผู้ที่เป็นมะเร็งเต้านมได้อย่างถูกต้องจำนวน 210 ราย จากจำนวนผู้ป่วยทั้งหมด 212 ราย คิดเป็นค่าความไวเท่ากับ 99.06% และสามารถจำแนกผู้ที่ไม่เป็นมะเร็งได้ถูกต้องจำนวน 346 ราย จาก 357 ราย คิดเป็นค่าความจำเพาะเท่ากับ 96.92% ค่าความแม่นยำเท่ากับ 97.72% แสดงให้เห็นถึงประสิทธิภาพของแบบจำลองในการจำแนกกลุ่มตัวอย่างได้อย่างถูกต้องในภาพรวม นอกจากนี้ ค่าประสิทธิภาพโดยรวมอยู่ที่ 97.00% ซึ่งสะท้อนว่าแบบจำลองสามารถรักษาสมดุลระหว่าง Precision และ Recall ได้อย่างมีประสิทธิภาพ

นอกจากนี้ผู้วิจัยได้นำเสนอแผนที่ความร้อน (Heatmap) แสดงค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ดังแสดงในรูปที่ 10 ซึ่งเหมาะสำหรับการวิเคราะห์ข้อมูลเชิงต่อเนื่องและการตรวจสอบความสัมพันธ์เชิงเส้นตรง โดยระดับความสัมพันธ์ของตัวแปรต้นที่มีต่อตัวแปรตามแบ่งออกเป็น 5 ระดับตามเกณฑ์ ได้แก่ สูงมาก (0.80–1.00) สูง (0.60–0.79) ปานกลาง (0.40–0.59) ต่ำ (0.20–0.39) และต่ำมาก (0.00–0.19) (Papageorgiou, 2022)

Attributes	diagnosis	radius	texture	perimeter	area	smoothness	compactness	concavity	concave_points	symmetry	fractal_dimension
diagnosis	1	0.730	0.415	0.743	0.709	0.359	0.597	0.696	0.777	0.330	-0.013
radius	0.730	1	0.324	0.998	0.987	0.171	0.506	0.677	0.823	0.148	-0.312
texture	0.415	0.324	1	0.330	0.321	-0.023	0.237	0.302	0.293	0.071	-0.076
perimeter	0.743	0.998	0.330	1	0.987	0.207	0.557	0.716	0.851	0.183	-0.261
area	0.709	0.987	0.321	0.987	1	0.177	0.499	0.686	0.823	0.151	-0.283
smoothness	0.359	0.171	-0.023	0.207	0.177	1	0.659	0.522	0.554	0.558	0.585
compactness	0.597	0.506	0.237	0.557	0.499	0.659	1	0.883	0.831	0.603	0.565
concavity	0.696	0.677	0.302	0.716	0.686	0.522	0.883	1	0.921	0.501	0.337
concave_points	0.777	0.823	0.293	0.851	0.823	0.554	0.831	0.921	1	0.462	0.167
symmetry	0.330	0.148	0.071	0.183	0.151	0.558	0.603	0.501	0.462	1	0.480
fractal_dimension	-0.013	-0.312	-0.076	-0.261	-0.283	0.585	0.565	0.337	0.167	0.480	1

รูปที่ 10. แผนภาพความร้อน แสดงค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรต่าง ๆ

จากรูปที่ 10 แสดงค่าสัมประสิทธิ์สหสัมพันธ์แบบ Pearson's Correlation Coefficient ซึ่งสามารถแสดงค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรได้ดังนี้ concave points, perimeter, radius, area และ concavity มีความสัมพันธ์กับ diagnosis อยู่ในระดับสูง โดยมีค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.777, 0.743, 0.730, 0.709 และ 0.696 ตามลำดับ compactness และ texture มีความสัมพันธ์กับ diagnosis อยู่ในระดับปานกลาง โดยมีค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.597 และ 0.415 ตามลำดับ smoothness และ symmetry มีความสัมพันธ์กับ diagnosis อยู่ในระดับต่ำ โดยมีค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.359 และ 0.330 และ fractal dimension มีความสัมพันธ์กับ diagnosis อยู่ในระดับต่ำมาก โดยมีค่าสัมประสิทธิ์สหสัมพันธ์ เท่ากับ -0.013 ดังนั้นตัวแปรที่มีความสัมพันธ์ในระดับสูงกับการวินิจฉัย ได้แก่ concave points, perimeter, radius, area และ concavity ซึ่งมีศักยภาพในการเป็น feature สำคัญของแบบจำลอง จึงควรใช้เป็นตัวแปรหลักในการปรับปรุงโมเดลในอนาคต

การทดลองเปรียบเทียบแบบจำลองด้านการตรวจคัดกรองโรคมะเร็งเต้านม พบว่า เทคนิคโครงข่ายประสาทเทียมทำงานได้ดีที่สุด โดยมีค่าความแม่นยำเท่ากับ 98.07% ค่าความไวเท่ากับ 99.15% ค่าความจำเพาะเท่ากับ 96.21% และค่าประสิทธิภาพโดยรวมเท่ากับ 98.47% ความโดดเด่นของเทคนิคนี้สะท้อนถึงศักยภาพในการจับรูปแบบข้อมูลที่ซับซ้อนได้สูงกว่ารูปแบบอื่นในบริบทข้อมูลนี้ อย่างไรก็ตาม งานวิจัยนี้มีข้อจำกัดที่ควรพิจารณา ได้แก่ ขนาดตัวอย่างที่มาจากแหล่งเดียว (569 รายจาก Kaggle) และความไม่สมดุลระหว่างกลุ่มที่เป็น (Malignant) และไม่เป็นโรค (Benign) ซึ่งอาจจำกัดความสามารถในการประยุกต์ใช้กับประชากรจริง นอกจากนี้ ปัจจัยด้านคุณลักษณะของตัวแปร ขั้นตอนการเตรียมข้อมูล การเลือกพารามิเตอร์ และวิธีประเมินผล อาจส่งผลต่อประสิทธิภาพการเรียนรู้ของโมเดล

ดังนั้น งานวิจัยในอนาคตควรใช้ข้อมูลที่มีความหลากหลายและครอบคลุมหลายภูมิภาค ควบคู่กับการบูรณาการข้อมูลภาพทางการแพทย์ เช่น Mammogram หรือ Ultrasound เพื่อเพิ่มความแม่นยำ พร้อมทั้งพิจารณาใช้ปัญญาประดิษฐ์ที่อธิบายได้ (Explainable artificial intelligence: XAI) เพื่อช่วยให้ผลลัพธ์ของโมเดลสามารถอธิบายได้ (Interpretability) ช่วยให้แพทย์เข้าใจและยอมรับในกระบวนการตัดสินใจของระบบ จากผลการวิเคราะห์โมเดลนี้มีศักยภาพสูงในการเป็นเครื่องมือสนับสนุนการคัดกรองเบื้องต้นในสถานพยาบาลระดับชุมชนได้อย่างมีประสิทธิภาพ ช่วยลดภาระงานของบุคลากร เพิ่มการเข้าถึงการตรวจ และสนับสนุนเป้าหมายนโยบายสาธารณสุขในการยกระดับคุณภาพชีวิตประชาชนอย่างยั่งยืน

4. สรุปผลการวิจัย

การศึกษานี้ผู้วิจัยได้นำชุดข้อมูลจริงจากมหาวิทยาลัยวิสคอนซิน ผ่านเว็บไซต์ www.Kaggle ซึ่งมีจำนวนผู้ป่วย 569 ราย ประกอบด้วยตัวแปรเชิงปริมาณจำนวน 11 ตัวแปร และ 1 ตัวแปรเป้าหมาย ซึ่งจำแนกผู้ป่วยเป็นกลุ่มที่มีความผิดปกติ และไม่พบความผิดปกติ มาใช้เป็นฐานข้อมูลในการวิเคราะห์ โดยดำเนินการภายใต้กรอบกระบวนการมาตรฐานของการทำเหมืองข้อมูล ทั้งนี้ได้ประยุกต์ใช้เทคนิคการจำแนกประเภทจำนวน 7 เทคนิค ได้แก่ โครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ เพื่อนบ้านใกล้ที่สุด ต้นไม้ตัดสินใจ การเรียนรู้เชิงลึก และวิธีการรวมกลุ่ม นอกจากนี้ ผู้วิจัยได้ดำเนินการปรับแต่งค่าพารามิเตอร์ เพื่อค้นหาค่าที่เหมาะสมที่สุดในแต่ละเทคนิค อันนำไปสู่การพัฒนาแบบจำลองที่มีประสิทธิภาพและแม่นยำในการคัดกรองผู้ป่วยอย่างสูงสุด และทำการเปรียบเทียบประสิทธิภาพการจำแนก

ประเภทข้อมูลทั้ง 7 เทคนิคนี้ด้วยการวัดประสิทธิภาพทั้ง 4 ค่า ได้แก่ ค่าความแม่นยำ ค่าความไว ค่าจำเพาะ ค่าประสิทธิภาพโดยรวม จากผลการวิจัยพบว่า เทคนิคโครงข่ายประสาทเทียมเป็นเทคนิคที่มีความเหมาะสมที่สุดในการสร้างตัวแบบคัดกรองผู้ป่วยมะเร็งเต้านม โดยการตั้งค่า Hidden layer size = 5 และ Learning rate = 0.01 ส่งผลให้โมเดลมีประสิทธิภาพสูงสุดในการทำนาย ซึ่งมีค่าความแม่นยำเท่ากับ 98.07% ค่าความไวเท่ากับ 99.15% ค่าจำเพาะเท่ากับ 96.21% และค่าประสิทธิภาพโดยรวมเท่ากับ 98.47% แม้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) จะให้ค่าความไวเทียบเท่ากับเทคนิคโครงข่ายประสาทเทียม และค่าความแม่นยำสูงถึง 97.36% แต่ค่าจำเพาะ และค่าประสิทธิภาพโดยรวมยังต่ำกว่าแบบจำลองโครงข่ายประสาทเทียม จากผลการวัดประสิทธิภาพนี้ชี้ชัดว่า โครงข่ายประสาทเทียมเป็นเทคนิคที่เหมาะสมอย่างยิ่งในการพัฒนาแบบจำลองคัดกรองผู้ป่วยโรคมะเร็งเต้านม เนื่องจากสามารถประมวลผลและคาดการณ์ได้อย่างแม่นยำและเชื่อถือได้ การนำเทคนิคนี้ไปใช้สนับสนุนการทำงานของแพทย์ในสถานพยาบาลระดับชุมชนจะช่วยเพิ่มความรวดเร็วและความแม่นยำในการคัดกรอง ลดภาระค่าใช้จ่าย และขยายโอกาสการเข้าถึงการตรวจ โดยเฉพาะในพื้นที่ที่มีข้อจำกัดด้านทรัพยากร ทั้งยังยกระดับประสิทธิภาพในการตรวจจับโรคตั้งแต่ระยะแรก ลดความเสี่ยงจากการวินิจฉัยผิด และสนับสนุนการตัดสินใจของแพทย์ได้อย่างมีประสิทธิภาพ ซึ่งเป็นปัจจัยสำคัญที่ช่วยเพิ่มโอกาสการรักษาและลดอัตราการเสียชีวิตจากโรคมะเร็งเต้านมได้อย่างเป็นรูปธรรม

เอกสารอ้างอิง (References)

Hfocus. (2024, November 21). *MOPH emphasizes breast cancer threat: 49 Thai women diagnosed daily, 13 deaths per day*. Hfocus. https://www.hfocus.org/content/2024/11/32298 (in Thai)

Kumjit, K., Jaikoomkao, D., Phumirang, W., Sattanako, A., & Sukprasert, A. (2022). The efficiency of data mining technique for the prognosis of cerebrovascular disease. *Journal of Applied Informatics and Technology*, *4*(2), 87-98. https://doi.org/10.14456/jait.2022.7 (in Thai)

Muhammad, N. S. (2024). *Breast cancer dataset.* Kaggle. https://www.kaggle.com/datasets/nairasaeedmuhammad/breast-cancer

Papageorgiou, S. N. (2022). On correlation coefficients and their interpretation. *Journal of Orthodontics*, *49*(3), 359–361. https://doi.org/10.1177/14653125221076142

Phikulsri, A., & Chanamarn, N. (2023). Efficiency comparison of classification methods for kidney disease with data mining techniques. *Journal of Science Engineering and Technology*, *3*(1), 1-17. https://ph02.tci-thaijo.org/index.php/JSET/article/view/247493 (in Thai)

Prema, K. M., & Jagadeesh, P. (2023). Detection of breast cancer using artificial neural network classifier and comparing with support vector machine classifier.

Proceedings of the 4th International Conference on Material Science and Applications (pp. 020105). AIP Publishing LLC. https://doi.org/10.1063/5.0173034

Rideach, N., Khaoead, A., & Srisomboon, P. (2022). A behavioral analysis model and the cause of alcohol dependence with the decision tree technique. *Journal of Kasetsart Educational Review*, *37*(3), 202-211. https://so04.tci-thaijo.org/index.php/eduku/article/view/251166 (in Thai)

Ruangsawud, A., Sukprasert, A., Sinthukoot, T., & Kaiwinit, S. (2023). Comparison of predictive models for the prognosis of lung cancer. *Kalasin University Journal of Science Technology and Innovation*, *2*(2), 39-52. https://doi.org/10.14456/ksti.2023.8 (in Thai)

Siphating, K., Peranam, N., Sawangloke, W., & Sukprasert, A. (2023). Classification of MRI images for brain tumor patient screening. *Journal of Applied Informatics and Technology*, *5*(2), 100-115. https://doi.org/10.14456/jait.2023.8 (in Thai)

Srisuk, U., & Thongkam, J. (2021). The efficiency comparison of data mining techniques for patient incidence. *Journal of Science and Technology Mahasarakham University*, *40*(2), 157-163. https://li01.tci-thaijo.org/index.php/scimsujournal/article/view/247870 (in Thai)

Sukprasert, A. (2023). *Data Mining with RapidMiner Studio* (5th ed.). Department of Business Computer, Mahasarakham Business School, Mahasarakham University, Mahasarakham. (in Thai)

Taiwiriyawet, W. (2023). *The number 1 cancer among women worldwide: Understand breast cancer before it spreads*. Thammasat University. https://tu.ac.th/thammasat-090566-breast-cancer-no1-cancer-among-women-worldwide (in Thai)

Tongkunwong, S., & Sawatkamon, P. (2024). Comparing the performance of machine learning models for classifying lung cancer patients. *UTK Research Journal*, *18*(1), 33-42. https://ph02.tci-thaijo.org/index.php/rmutk/article/view/252954 (in Thai)

Wacharaphapaiboon, W. (2021). *What causes breast cancer?* Praram 9 Hospital. https://www.praram9.com/breast-cancer-staging/ (in Thai)

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29-40), Manchester, UK. http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

World Health Organization. (2024). *Breast cancer.* https://www.who.int/news-room/fact-sheets/detail/breast-cancer