

การพยากรณ์โรคมะเร็งเต้านมด้วยอัลกอริทึมการจำแนกประเภทแบบเคมีน

ร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง

Breast Cancer Prediction Using K-mean Classification Algorithm with Self-adaptive Weight

อาริกา ธรรมโน^{1*}, มุทิตา หวังคิด¹ และ อาริต ธรรมโน²

Arika Thammano^{1}, Muthita Wangkid¹ and Arit Thammano²*

¹โรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ ศูนย์วิจัยและพัฒนาการศึกษา

มหาวิทยาลัยเกษตรศาสตร์ กรุงเทพฯ 10900

²ห้องปฏิบัติการวิจัยปัญญาเชิงคำนวณ คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ 10520

¹*Kasetsart University Laboratory School, Center for Educational Research and Development,*

Kasetsart University, Bangkok 10900

²*Computational Intelligence Laboratory, Faculty of Information Technology,*

King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520

Received: July 02, 2020; Revised: August 05, 2020; Accepted: August 11, 2020; Published: August 26, 2020

ABSTRACT – “Breast cancer” is the first-rank non-communicable disease found in women both in Thailand and the world at large. The statistic of the National Cancer Institute reveals that the number of Thai women suffering from breast cancer is likely to increase every year. If breast cancer can be found at its early stage and cured properly, the risk of mortality can be considerably reduced. This research presents K-mean classification algorithm with self-adaptive weight and the program for breast cancer prediction using Python programming language with an aim to help identify and cure the early-stage breast cancer patients in a timely manner. The algorithm presented in this research has been modified from K-mean clustering algorithm to be able to perform the classification task and to have the ability to self-adapt the weights of the features in the Euclidean distance equation. The efficiency of algorithm and breast cancer prediction program was tested using Breast Cancer Coimbra data set. The result shows that the breast cancer prediction using the proposed algorithm is more accurate than other artificial intelligence algorithms.

KEYWORDS: Breast cancer, Prediction, Artificial intelligence, Data science, Classification, K-mean algorithm

*Corresponding Author: arikatham@gmail.com

บทคัดย่อ - “มะเร็งเต้านม” เป็นโรคในกลุ่มโรคไม่ติดต่อเรื้อรังที่พบมากเป็นอันดับหนึ่งในผู้หญิงทั้งในประเทศไทยและต่างประเทศ สถิติของสถาบันมะเร็งแห่งชาติระบุว่า ผู้หญิงไทยที่ป่วยเป็นโรคมะเร็งเต้านมมีแนวโน้มที่จะเพิ่มมากขึ้นทุกปี ซึ่งถ้าหากตรวจพบเร็วและสามารถรักษาได้ถูกวิธี ก็จะสามารถลดอัตราการเสียชีวิตได้อย่างมาก งานวิจัยนี้จึงได้นำเสนออัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง รวมทั้งทำการพัฒนาโปรแกรมพยากรณ์โรคมะเร็งเต้านมด้วยภาษาไพทอน โดยมีจุดประสงค์เพื่อช่วยในการคัดกรองผู้ป่วยโรคมะเร็งเต้านมในเบื้องต้นเพื่อให้กระบวนการการวินิจฉัยเป็นไปได้อย่างรวดเร็วและสามารถรักษาได้อย่างทันเวลา โดยอัลกอริทึมที่นำเสนอเป็นอัลกอริทึมที่ถูกพัฒนาต่อยอดมาจากอัลกอริทึมการแบ่งกลุ่มแบบเคมีนให้มีความสามารถในการจำแนกประเภท และปรับเปลี่ยนค่าถ่วงน้ำหนักของคุณลักษณะเด่นในสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของประเภทแต่ละประเภทให้เหมาะสมได้ด้วยตัวเองในระหว่างการเรียนรู้ข้อมูล งานวิจัยนี้ได้ทำการทดสอบประสิทธิภาพของอัลกอริทึมและโปรแกรมการพยากรณ์โรคมะเร็งเต้านมที่นำเสนอโดยใช้ชุดข้อมูล Breast Cancer Coimbra ผลการทดสอบแสดงให้เห็นถึงความถูกต้องในการพยากรณ์โรคมะเร็งเต้านมของอัลกอริทึมที่นำเสนอที่สูงกว่าอัลกอริทึมทางปัญญาประดิษฐ์อื่นๆ

คำสำคัญ: โรคมะเร็งเต้านม, การพยากรณ์, ปัญญาประดิษฐ์, วิทยาการข้อมูล, การจำแนกประเภท, อัลกอริทึมเคมีน

1. บทนำ

มะเร็งเต้านมเป็นโรคในกลุ่มโรคไม่ติดต่อเรื้อรังที่พบมากเป็นอันดับหนึ่งในผู้หญิงทั่วโลก โดยสถิติของสถาบันมะเร็งแห่งชาติได้แสดงให้เห็นว่า ผู้หญิงไทยป่วยเป็นโรคมะเร็งเต้านม 13,000 คนต่อปี หรือ 35 คนต่อวัน และยังมีแนวโน้มที่จะเพิ่มมากขึ้นทุกปี [1] อีกทั้งองค์การอนามัยโลกได้เปิดเผยว่า มีผู้หญิงจากทั่วโลกป่วยเป็นโรคมะเร็งเต้านมเพิ่มขึ้น 2.1 ล้านคนต่อปี และมีแนวโน้มที่จะเพิ่มขึ้นในปีต่อไป นอกจากนี้โรคมะเร็งเต้านมยังเป็นสาเหตุการตายของผู้ป่วยมากกว่า 627,000 คนในปีพุทธศักราช 2561 [2] ซึ่งนับเป็นร้อยละ 15 ของผู้ป่วยโรคมะเร็งเต้านมทั้งหมด โดยหากโรคมะเร็งเต้านมถูกตรวจพบได้ในระยะเริ่มแรกและได้รับการรักษาอย่างทันเวลา ก็จะสามารถลดอัตราการตายดังกล่าวได้อย่างมาก

ในปัจจุบัน วิธีการวินิจฉัยโรคมะเร็งเต้านมที่ถูกใช้มากที่สุด คือการแมมโมแกรมซึ่งมีข้อจำกัด คือ สามารถแปลผลได้ยาก และในกรณีที่พบก้อน ก็อาจจะแยกได้ยากว่าก้อนนั้นเป็นก้อนน้ำหรือก้อนเนื้อ [3] หรือในบางรายที่มีความหนาแน่นของเนื้อเต้านมมาก เนื้อเต้านมก็มีโอกาสที่จะบังสิ่งผิดปกติทำให้ตรวจไม่พบ [4] วิธีการตรวจหาโรคมะเร็งเต้านมที่ดีอีกวิธีหนึ่งคือการตรวจเลือดเพื่อหาสารบ่งชี้มะเร็ง ซึ่งเมื่อใช้ร่วมกับการซักประวัติ การตรวจร่างกายและการตรวจทางกายภาพอื่นๆ เช่น การเอกซเรย์ หรือการอัลตราซาวด์ จะสามารถใช้ในการวินิจฉัยโรคมะเร็งเต้านมได้ แต่เนื่องจากโรคมะเร็งเต้านมใน

ระยะแรก ผู้ป่วยจะไม่มีอาการใดๆ ถ้าผู้ป่วยไม่ได้ตรวจเต้านมด้วยตัวเองเป็นประจำ หรือไม่ได้ไปตรวจคัดกรองที่สถานพยาบาล เซลล์มะเร็งอาจจะลุกลามไปยังอวัยวะข้างเคียง ทำให้การรักษาให้หายขาดเป็นไปได้ยากลำบากมากยิ่งขึ้น

ผู้วิจัยจึงมีแนวคิดที่จะใช้เทคโนโลยีสารสนเทศมาเป็นตัวช่วยในการคัดกรองโรคมะเร็งเต้านมในเบื้องต้นให้กับประชาชนทั่วไป โดยในบทความนี้ผู้วิจัยได้นำเสนออัลกอริทึมทางปัญญาประดิษฐ์ 2 อัลกอริทึม พร้อมทั้งได้ทำการพัฒนาโปรแกรมพยากรณ์โรคมะเร็งเต้านมด้วยอัลกอริทึมที่นำเสนอ โปรแกรมที่พัฒนาขึ้นมานี้อาศัยผลตรวจเลือดจากการตรวจสุขภาพประจำปีเพื่อเป็นข้อมูลในการพยากรณ์และจำแนกบุคคลทั่วไปที่ไม่ได้ป่วยเป็นโรคมะเร็งเต้านม และผู้ป่วยโรคมะเร็งเต้านมออกจากกัน ผลลัพธ์ของการพยากรณ์ที่ได้จากอัลกอริทึมที่นำเสนอทั้ง 2 อัลกอริทึมจะถูกนำไปเปรียบเทียบกับผลลัพธ์ของการพยากรณ์ที่ได้จากอัลกอริทึมทางปัญญาประดิษฐ์อื่นที่ได้รับการยอมรับอย่างกว้างขวาง 5 อัลกอริทึม เพื่อให้มั่นใจถึงประสิทธิภาพของอัลกอริทึมที่นำเสนอ ผู้วิจัยคาดหวังว่าเมื่อผู้ป่วยได้รับทราบผลจากการพยากรณ์ของโปรแกรมจะไปพบแพทย์ เพื่อรับการตรวจคัดกรองและได้รับการรักษาอย่างถูกต้องตามหลักวิชาการได้อย่างทันท่วงทีตั้งแต่ในระยะเริ่มต้น ซึ่งจะนำไปสู่การลดลงของอัตราการตายในผู้ป่วยโรคมะเร็งเต้านมได้

เนื้อหาของบทความนี้ถูกแบ่งเป็น 5 หัวข้อดังนี้ หัวข้อที่ 1 บทนำ หัวข้อที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง หัวข้อที่ 3

อัลกอริทึมในการพยากรณ์โรคมะเร็งเต้านมที่นำเสนอ หัวข้อที่ 4 ผลการทดลอง และหัวข้อที่ 5 จะเป็นบทสรุป

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 โรคมะเร็งเต้านม (Breast Cancer)

โรคมะเร็งเต้านม เกิดจากการที่เซลล์ในเต้านมมีความผิดปกติ เกิดการเจริญเติบโตและแบ่งเซลล์ได้เร็วกว่าเซลล์ปกติโดยที่ร่างกายไม่สามารถควบคุมได้ จนเกิดเป็นก้อนเนื้อหรือตุ่มที่โตขึ้นอย่างผิดปกติ โดยที่เซลล์เหล่านี้สามารถเจริญลุกลามและแพร่ไปยังส่วนอื่นๆ ของร่างกายได้ เช่น แพร่เข้าต่อมน้ำเหลือง หรือแพร่กระจายเข้าหลอดเลือด กระแสเลือด ส่งผลให้เซลล์หรือเนื้อเยื่อนั้นไม่สามารถทำงานได้อย่างปกติ และเป็นสาเหตุแห่งการเสียชีวิตได้ในที่สุด

โรคมะเร็งเต้านม เป็นโรคมะเร็งอันดับ 1 ที่พบในผู้หญิงทั้งในไทยและทั่วโลก และมีแนวโน้มที่จะเพิ่มสูงขึ้นทุกปี ในปี พ.ศ. 2557 ข้อมูลทางระบาดวิทยาในสหรัฐอเมริกาได้ระบุว่า มีผู้ป่วยโรคมะเร็งเต้านมรายใหม่กว่า 232,670 ราย [5] ในส่วนของประเทศไทยนั้น โรคมะเร็งเต้านมเป็นโรคมะเร็งที่เกิดมากที่สุดของผู้หญิง จากรายงานทะเบียนมะเร็งระดับโรงพยาบาลของสถาบันมะเร็งแห่งชาติในปี พ.ศ. 2557 ได้ระบุว่า ผู้ป่วยโรคมะเร็งเต้านมเพศหญิงถือเป็นร้อยละ 31.4 ของผู้ป่วยโรคมะเร็งเพศหญิงทั้งหมด

โรคมะเร็งเต้านมสามารถแบ่งออกได้เป็น 4 ระยะ [6] คือ ระยะที่ 1 ก้อนมะเร็งมีขนาดเล็กกว่า 2 เซนติเมตร และยังไม่มีการแพร่กระจายไปยังต่อมน้ำเหลืองที่รักแร้ ระยะที่ 2 ก้อนมะเร็งมีขนาดระหว่าง 2 – 5 เซนติเมตร อาจจะมีการแพร่กระจายไปยังต่อมน้ำเหลืองที่รักแร้ข้างเดียวหรือไม่กี่ได้ ระยะที่ 3 ก้อนมะเร็งมีขนาดใหญ่กว่า 5 เซนติเมตร และมีการแพร่กระจายไปยังต่อมน้ำเหลืองที่รักแร้ข้างเดียวอย่างมาก และระยะที่ 4 ก้อนมะเร็งมีขนาดเท่าไรก็ได้ แต่มีการแพร่กระจายไปยังส่วนอื่นของร่างกายที่อยู่ห่างออกไป โรคมะเร็งเต้านมสามารถพบได้ทั้งในเพศหญิงและเพศชาย แต่จากสถิติพบว่าเพศหญิงมีความเสี่ยงที่จะเป็นโรคมะเร็งเต้านมมากกว่าเพศชาย ในปี พ.ศ. 2552 มีผู้ป่วยโรคมะเร็งเต้านมเพศหญิงจำนวน 765 ราย แบ่งออกเป็นผู้ป่วยในระยะที่ 1 ร้อยละ 20 ระยะที่ 2 ร้อยละ 35 ระยะที่ 3 ร้อยละ 20 และระยะที่ 4 หรือระยะกระจายร้อยละ 7 ส่วนผู้ป่วยโรคมะเร็งเต้านมเพศชายในปี

เดียวกันมีเพียง 3 รายเท่านั้น เป็นผู้ป่วยในระยะที่ 1 ร้อยละ 33.33 และเป็นผู้ป่วยในระยะที่ 2 ร้อยละ 33.33 [7]

ในปัจจุบันวิธีในการตรวจหาโรคมะเร็งเต้านมนั้นจะทำโดยวิธีการถ่ายภาพรังสีเต้านม (mammogram) แต่ก็ยังมีอีกวิธีในการตรวจหาหะเร็ง ได้แก่ การตรวจเลือดเพื่อหาสารบ่งชี้มะเร็ง (tumor marker) สารบ่งชี้มะเร็งอาจจะเป็นแอนติเจน สารโปรตีน ฮอร์โมน หรือเอนไซม์ ที่จะตรวจพบได้ในปริมาณที่สูงในกรณีที่ผู้ป่วยเป็นมะเร็งเต้านม เช่น Carcinoembryonic antigen (CEA), CA 19-9 และ CA 15-3 เป็นต้น [8] โดยที่ปริมาณสารบ่งชี้มะเร็งจะแปรผันตรงกับระยะของโรคมะเร็ง [9] ถ้าหากพบปริมาณสารบ่งชี้มะเร็งมาก ก็จะคาดการณ์ได้ว่ามะเร็งได้มีขนาดใหญ่และลุกลามไปยังส่วนอื่นๆ ของร่างกายแล้ว ดังนั้นการตรวจเลือดเพื่อหาสารบ่งชี้มะเร็ง ร่วมกับการตรวจร่างกาย การซักประวัติ และการตรวจทางพยาธิวิทยา เช่น การถ่ายภาพรังสีเต้านม การเอกซเรย์ หรือการอัลตราซาวด์ ก็จะสามารถช่วยในการวินิจฉัยและพยากรณ์โรคมะเร็งเต้านมได้เป็นอย่างดี ซึ่งถ้าหากโรคมะเร็งได้รับการวินิจฉัยอย่างถูกต้องและรวดเร็ว ก็จะทำให้โอกาสในการลุกลามของมะเร็งและการเสียชีวิตของผู้ป่วยลดลงไปด้วย

2.2 งานวิจัยที่เกี่ยวข้อง

จากความจำเป็นที่ผู้ป่วยควรจะต้องได้รับการวินิจฉัยโรคมะเร็งเต้านมอย่างถูกต้องและรวดเร็ว จึงได้มีนักวิจัยนำเอาอัลกอริทึมทางปัญญาประดิษฐ์ต่างๆ มาทำการทดสอบประสิทธิภาพในการพยากรณ์โรคมะเร็งเต้านม

Sardouk, Duru และ Bayat [10] ได้ใช้โปรแกรม WEKA และ MATLAB ในการเปรียบเทียบประสิทธิภาพในการพยากรณ์โรคมะเร็งเต้านมของอัลกอริทึม 6 อัลกอริทึม ได้แก่ AdaBoost, Classification Via Regression, Random Forest, Jrip, Radial Basis Function Neural Network (RBFNN) และ J48 จากผลการทดลองพบว่า Classification Via Regression มีค่าความแม่นยำสูงที่สุดในกลุ่ม ตามมาด้วย AdaBoost, Random Forest, RBFNN, Jrip และ J48 ตามลำดับ

Ray และคณะ [11] ทำการทดลองโดยใช้ Decision Tree, k-nearest neighbors (k-NN), Random Forest และ Gaussian Naïve Bayes ในการพยากรณ์โรคมะเร็งเต้านม จากผลการทดลองพบว่า Decision Tree และ Random Forest มีความสามารถในการพยากรณ์สูงกว่าอีก 2 อัลกอริทึมที่เหลือ

Austria และคณะ [12] ได้นำเอาอัลกอริทึมทางปัญญาประดิษฐ์ 10 อัลกอริทึมมาพัฒนาโปรแกรมเพื่อพยากรณ์โรคมะเร็งเต้านม ผลการทดลองสรุปได้ว่าอัลกอริทึม Gradient Boosting สามารถพยากรณ์ได้ค่าความถูกต้องสูงสุด ส่วนอัลกอริทึม k-NN เป็นอัลกอริทึมที่มีความสามารถในการเรียนรู้ได้เร็วที่สุด และอัลกอริทึม Nonlinear Support Vector Machine ใช้เวลาในการทดสอบเร็วที่สุด

2.3 การแบ่งกลุ่ม (Clustering)

การแบ่งกลุ่มเป็นเทคนิคในการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) ข้อมูลจะถูกแบ่งออกเป็นกลุ่มๆ โดยที่ผู้ใช้ไม่ต้องกำหนดเกณฑ์ที่ใช้ในการคัดเลือกข้อมูลเข้ากลุ่ม และไม่มีมีการกำหนดเป้าหมายว่าข้อมูลควรจะอยู่ในกลุ่มใด แต่จะใช้เพียงคุณลักษณะของข้อมูลเป็นแนวทางในการแบ่งกลุ่ม โดยข้อมูลที่มีคุณลักษณะคล้ายกันจะถูกจัดให้อยู่ในกลุ่มเดียวกัน

2.4 การจำแนกประเภท (Classification)

การจำแนกประเภทเป็นเทคนิคในการเรียนรู้แบบมีผู้สอน (Supervised learning) การจำแนกประเภทจะทำการสร้างเกณฑ์เพื่อใช้ในการแบ่งข้อมูลออกเป็นกลุ่มๆ โดยมีแอมพริบิวท์เป้าหมายเป็นตัวกำหนดให้ว่าข้อมูลไหนควรจะอยู่ในกลุ่มใด ดังนั้นการจำแนกประเภทจึงจำเป็นที่จะต้องมียข้อมูล 2 กลุ่ม คือกลุ่มข้อมูลที่ใช้ในการเรียนรู้และกลุ่มข้อมูลที่ใช้ในการทดสอบ กลุ่มข้อมูลที่ใช้ในการเรียนรู้จะถูกใช้ในการสร้างเกณฑ์ในการแบ่งข้อมูล ส่วนกลุ่มข้อมูลที่ใช้ในการทดสอบจะถูกใช้ในการตรวจสอบว่าเกณฑ์ที่สร้างขึ้นมามีประสิทธิภาพมากน้อยเพียงใด โดยดูจากตัววัดประสิทธิภาพต่างๆ ได้แก่ ความถูกต้อง (Accuracy) ความแม่นยำ (Precision) ความครบถ้วน (Recall) และ F1 Score

2.5 อัลกอริทึมการแบ่งกลุ่มแบบเคมีน (K-mean clustering algorithm)

เคมีนเป็นอัลกอริทึมในการแบ่งกลุ่มที่ได้รับความนิยมสูงสุดเนื่องจากมีประสิทธิภาพสูงและมีวิธีการดำเนินงานที่ง่าย โดยเคมีนจะทำการแบ่งข้อมูลออกเป็น K กลุ่ม ข้อมูลที่มีคุณลักษณะคล้ายกันจะถูกจัดเอาไว้ในกลุ่มเดียวกัน หลังจากนั้นจะทำการหาค่าเฉลี่ยของข้อมูลภายในกลุ่มเพื่อกำหนดเป็นตัวแทนของกลุ่ม เป้าหมายสูงสุดของเคมีนคือต้องการให้ข้อมูลที่อยู่ในกลุ่ม

เดียวกันมีระยะห่างระหว่างกันน้อยที่สุด ส่วนข้อมูลที่อยู่ต่างกลุ่มกันให้มีระยะห่างระหว่างกันมากที่สุด ขั้นตอนการทำงานของอัลกอริทึมการแบ่งกลุ่มแบบเคมีนเป็นดังนี้

- ก) ในขั้นตอนเริ่มต้นให้ทำการสุ่มข้อมูลขึ้นมา K ข้อมูลจากข้อมูลที่จะใช้ในการแบ่งกลุ่มทั้งหมด เพื่อนำมาใช้เป็นจุดศูนย์กลางของกลุ่มแต่ละกลุ่ม โดยค่า K จะถูกกำหนดโดยผู้ใช้งานต้องการจะแบ่งข้อมูลออกเป็นกี่กลุ่ม
- ข) จัดข้อมูลทั้งหมดเข้าในกลุ่ม โดยหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของแต่ละกลุ่ม ถ้าข้อมูลนั้นมีระยะห่างจากจุดศูนย์กลางกลุ่มใดน้อยที่สุด ก็จะได้ว่าข้อมูลดังกล่าวอยู่ในกลุ่มนั้น
- ค) หาค่าเฉลี่ย (Mean) ของข้อมูลภายในแต่ละกลุ่มเพื่อนำมาเป็นจุดศูนย์กลางใหม่ของกลุ่ม
- ง) ทำซ้ำในข้อ ข) และ ค) จนกระทั่งตำแหน่งของจุดศูนย์กลางกลุ่มไม่มีการเปลี่ยนแปลง

3. อัลกอริทึมในการพยากรณ์โรคมะเร็งเต้านม

ในงานวิจัยนี้ผู้วิจัยได้นำเสนออัลกอริทึมที่จะใช้ในการพยากรณ์โรคมะเร็งเต้านม 2 อัลกอริทึม ดังที่จะได้อธิบายในรายละเอียดต่อไป

3.1 อัลกอริทึมการจำแนกประเภทแบบเคมีน (K-mean classification algorithm)

อัลกอริทึมการจำแนกประเภทแบบเคมีนนี้เกิดจากการที่ผู้วิจัยได้ทำการปรับเปลี่ยนอัลกอริทึมการแบ่งกลุ่มแบบเคมีน (หัวข้อ 2.4) เล็กน้อยในส่วนของการสุ่มจุดศูนย์กลางเริ่มต้นให้กับแต่ละกลุ่มข้อมูลเพื่อให้อัลกอริทึมรับรู้ถึงประเภทของข้อมูลที่จะต้องทำการจำแนก โดยขั้นตอนการทำงานของอัลกอริทึมการจำแนกประเภทแบบเคมีนเป็นดังนี้

- ก) สุ่มข้อมูลขึ้นมา K ข้อมูลจากส่วนของข้อมูลที่ใช้ในการเรียนรู้ เพื่อใช้เป็นจุดศูนย์กลางของกลุ่มแต่ละกลุ่ม ในกรณีของอัลกอริทึมการจำแนกประเภทแบบเคมีนนี้ ค่า K จะไม่ได้ถูกกำหนดโดยผู้ใช้งานเหมือนกับกรณีของอัลกอริทึมการแบ่งกลุ่มแบบเคมีน แต่จะมีค่าเท่ากับจำนวนประเภทของข้อมูลที่จะต้องทำการจำแนก นอกจากนี้แล้ววิธีการในการสุ่มเลือกก็แตกต่างกันด้วย นั่นคือการสุ่มจุดศูนย์กลางเริ่มต้นให้กับกลุ่มแต่ละกลุ่มในอัลกอริทึมการจำแนกประเภทแบบเคมีนนี้นั้น จะสุ่มจากเฉพาะข้อมูลที่อยู่ในประเภทเดียวกับกลุ่มนั้นเท่านั้น ตัวอย่างเช่น จุดศูนย์กลาง

เริ่มต้นของกลุ่มที่ 1 ก็จะสุ่มมาจากข้อมูลที่อยู่ในประเภท 1 จุดศูนย์กลางเริ่มต้นของกลุ่มที่ 2 ก็จะสุ่มมาจากข้อมูลที่อยู่ในประเภท 2 เป็นต้น

- ข) จัดข้อมูลทั้งหมดเข้าในกลุ่ม โดยหาค่าระยะห่างระหว่างข้อมูลและจุดศูนย์กลางของกลุ่มตามสมการที่ (1) ถ้าค่าระยะห่างระหว่างข้อมูลและจุดศูนย์กลางกลุ่มใดน้อยที่สุด ให้จัดข้อมูลใส่ในกลุ่มนั้น

$$d_{ij} = \|X_i - C_j\| = \sqrt{\sum_{k=1}^N a_{jk} (x_{ik} - c_{jk})^2} \quad (1)$$

เมื่อ d_{ij} คือระยะห่างระหว่างข้อมูลตัวที่ i (X_i) กับจุดศูนย์กลางกลุ่ม j (C_j)

x_{ik} และ c_{jk} คือค่าคุณลักษณะที่ k ของข้อมูลตัวที่ i และค่าคุณลักษณะที่ k ของจุดศูนย์กลางกลุ่ม j ตามลำดับ

a_{jk} คือค่าถ่วงน้ำหนักของคุณลักษณะที่ k ของจุดศูนย์กลางกลุ่ม j โดยอัลกอริทึมการจำแนกประเภทแบบเคมีนจะกำหนดให้ $a_{jk} = 1$

- ค) หาค่าเฉลี่ยของข้อมูลภายในแต่ละกลุ่ม เพื่อนำมาเป็นจุดศูนย์กลางใหม่ของกลุ่ม
ง) ทำวิธีการในข้อ ข) และ ค) ซ้ำ จนกระทั่งค่าของจุดศูนย์กลางกลุ่มไม่มีการเปลี่ยนแปลง

3.2 อัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง (K-mean classification algorithm with self-adaptive weight)

อัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเองเป็นอัลกอริทึมที่ถูกพัฒนาต่อยอดมาจากอัลกอริทึมการจำแนกประเภทแบบเคมีนที่ผู้วิจัยได้นำเสนอไว้ในหัวข้อ 3.1 โดยมีจุดประสงค์เพื่อให้ผลการพยากรณ์มีความแม่นยำมากขึ้น ในการพัฒนาอัลกอริทึมใหม่นี้อาศัยแนวคิดที่ว่าประเภทแต่ละประเภทจะมีคุณลักษณะเด่นที่แตกต่างกัน ดังนั้นสมการในการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของประเภทแต่ละประเภทก็ควรจะให้ความสำคัญกับคุณลักษณะเด่นของประเภทตัวเองเป็นพิเศษ โดยการเพิ่มค่าถ่วงน้ำหนักให้กับคุณลักษณะเด่นนั้นให้มากกว่าค่าถ่วงน้ำหนักของคุณลักษณะอื่น ตัวอย่างเช่น ถ้าเราต้องการจำแนกโรค 2 ประเภท โรคประเภทที่ 1 มีจุดเด่นที่คุณลักษณะที่ 4, 6 และ 8

โรคประเภทที่ 2 มีจุดเด่นที่คุณลักษณะที่ 1, 3 และ 5 สมการที่ใช้หาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของโรคประเภท 1 ก็จะทำให้การเพิ่มให้ค่าถ่วงน้ำหนักให้กับคุณลักษณะเด่นที่ 4, 6 และ 8 มากกว่าค่าถ่วงน้ำหนักของคุณลักษณะอื่น ส่วนสมการที่ใช้หาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของโรคประเภท 2 ก็จะทำการเพิ่มให้ค่าถ่วงน้ำหนักให้กับคุณลักษณะเด่นที่ 1, 3 และ 5 มากกว่าค่าถ่วงน้ำหนักของคุณลักษณะอื่น ขั้นตอนการทำงานของอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเองเป็นดังนี้

- ก) ทำการหาตำแหน่งของจุดศูนย์กลางกลุ่มแต่ละกลุ่มด้วยขั้นตอนของอัลกอริทึมการจำแนกประเภทแบบเคมีนตามที่ได้อธิบายไว้ในหัวข้อ 3.1

- ข) ทำการปรับค่าถ่วงน้ำหนักของคุณลักษณะ (a_{jk}) ในสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของแต่ละประเภท โดยวิธีการจะเริ่มต้นจากการนำข้อมูลจากส่วนของข้อมูลที่ใช้ในการเรียนรู้ที่ละข้อมูลมาทำการจำแนกประเภทจนครบทุกข้อมูล ซึ่งผลการจำแนกจะมีได้ 2 กรณีคือ

กรณี 1 ข้อมูลที่ได้ผลของการจำแนกประเภทถูกต้อง แสดงว่าสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของประเภทของข้อมูลนั้นมีความเหมาะสมแล้ว จึงไม่ต้องทำการปรับเปลี่ยนใดๆ

กรณี 2 ข้อมูลที่ได้ผลของการจำแนกประเภทไม่ถูกต้อง เราจะทำกรปรับค่าถ่วงน้ำหนักของคุณลักษณะในสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของประเภทที่ถูกต้องของข้อมูลนั้น ในการปรับค่าถ่วงน้ำหนักเราจะพิจารณาไปที่ละคุณลักษณะ โดยพิจารณาส่วนต่างที่คุณลักษณะนั้นของข้อมูลกับกลุ่มที่ไม่ถูกต้อง เทียบกับส่วนต่างที่คุณลักษณะนั้นของข้อมูลกับกลุ่มที่ถูกต้อง ถ้าคุณลักษณะใดที่มีส่วนต่างที่คุณลักษณะนั้นของข้อมูลกับกลุ่มที่ไม่ถูกต้องน้อยกว่าส่วนต่างที่คุณลักษณะนั้นของข้อมูลกับกลุ่มที่ถูกต้อง ให้ทำการลดค่าถ่วงน้ำหนักของคุณลักษณะนั้นตามสมการที่ (2) และคุณลักษณะใดที่มีส่วนต่างที่คุณลักษณะนั้นของข้อมูลกับกลุ่มที่ไม่ถูกต้องมากกว่าส่วนต่างที่คุณลักษณะนั้นของข้อมูลกับกลุ่มที่ถูกต้อง ให้ทำการเพิ่มค่าถ่วงน้ำหนักของคุณลักษณะนั้นตามสมการที่ (3)

$$a_{jk} = a_{jk} - \varepsilon \quad (2)$$

$$a_{jk} = a_{jk} + \varepsilon \quad (3)$$

เมื่อ ε เป็นจำนวนจริงที่มีค่าน้อยๆ ที่ใช้เป็นสเกลในการลดหรือเพิ่มค่าถ่วงน้ำหนัก

ก) ทำซ้ำในขั้นตอน ข) จนครบตามจำนวนรอบที่ผู้ใช้กำหนด

4. ผลการทดลองและการอภิปราย

งานวิจัยนี้ทำการทดสอบประสิทธิภาพของอัลกอริทึมและโปรแกรมการพยากรณ์โรคมะเร็งเต้านมที่นำเสนอ โดยใช้ชุดข้อมูล Breast Cancer Coimbra [13] ซึ่งเป็นชุดข้อมูลมาตรฐานที่เปิดเป็นสาธารณะในปี ค.ศ. 2018 เพื่อให้ใช้ในการทำวิจัยตัวอย่างของชุดข้อมูลได้ถูกแสดงไว้ในตารางที่ 1 ชุดข้อมูลนี้บุคลากรของคณะแพทยศาสตร์ มหาวิทยาลัย Coimbra และ University Hospital Centre of Coimbra ได้รวบรวมจากกลุ่มตัวอย่างจำนวน 116 ราย แบ่งเป็นบุคคลทั่วไปที่ไม่ได้ป่วยเป็นโรคมะเร็งเต้านม 52 ราย และผู้ป่วยโรคมะเร็งเต้านม 64 ราย

โดยได้ทำการเก็บไว้ใน UCI Machine Learning Repository [14] ชุดข้อมูลนี้มี 10 แอททริบิวต์ โดย 9 แอททริบิวต์แรกเป็นคุณลักษณะของกลุ่มตัวอย่างที่ได้จากการสอบถามและตรวจวิเคราะห์เลือด โดยมีรายละเอียดดังแสดงไว้ในตารางที่ 2 ส่วนแอททริบิวต์สุดท้ายจะบอกถึงประเภทของกลุ่มตัวอย่าง (1: บุคคลทั่วไปที่ไม่ได้ป่วยเป็นโรคมะเร็งเต้านม และ 2: ผู้ป่วยโรคมะเร็งเต้านม)

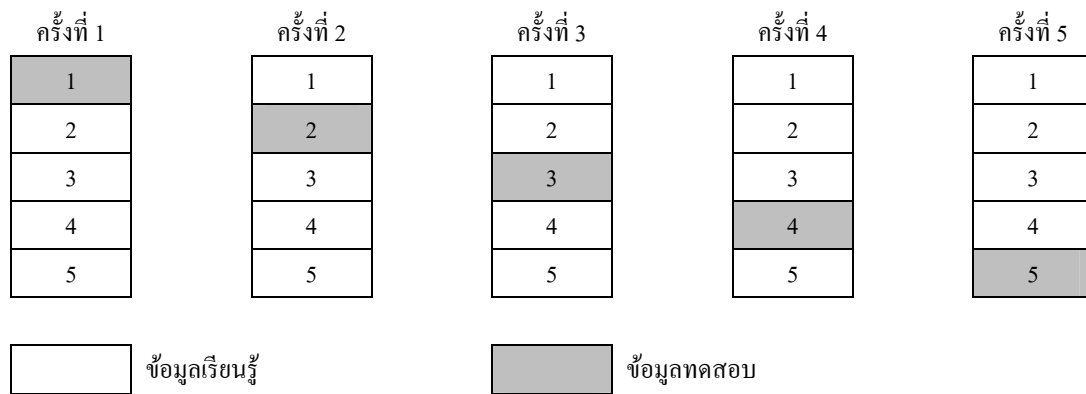
การทดสอบประสิทธิภาพของอัลกอริทึมในการจำแนกประเภทนั้นจำเป็นที่จะต้องแบ่งชุดข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มข้อมูลที่ใช้ในการเรียนรู้และกลุ่มข้อมูลที่ใช้ในการทดสอบ เพื่อให้การทดสอบมีความน่าเชื่อถือผู้วิจัยจึงเลือกใช้วิธี 5-fold Cross Validation ในการแบ่งชุดข้อมูล ซึ่งวิธีการของ 5-fold Cross Validation นั้นจะเริ่มจากการแบ่งข้อมูลทั้งหมดในชุดข้อมูลออกเป็น 5 ส่วน ข้อมูลในแต่ละส่วนจะผลัดกันไปเป็นข้อมูลทดสอบ ในขณะที่อีก 4 ส่วนที่เหลือถูกใช้เป็นข้อมูลที่ใช้ในการเรียนรู้ ดังนั้นการทดสอบจะต้องทำทั้งหมด 5 ครั้ง โดยแต่ละครั้งจะมีข้อมูลที่ใช้ในการเรียนรู้และข้อมูลที่ใช้ในการทดสอบที่แตกต่างกันไป ดังแสดงในรูปที่ 1

ตารางที่ 1. ตารางแสดงตัวอย่างของชุดข้อมูล Breast Cancer Coimbra

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP-1	Class
48	23.5	70	2.707	0.467408667	8.8071	9.7024	7.99585	417.114	1
83	20.69049454	92	3.115	0.706897333	8.8438	5.429285	4.06405	468.786	1
82	23.12467037	91	4.498	1.009651067	17.9393	22.43204	9.27715	554.697	1
...	...	...	...	...	...	...	...	...	...
45	21.30394858	102	13.852	3.4851632	7.6476	21.056625	23.03408	552.444	2
45	20.82999519	74	4.56	0.832352	7.7529	8.237405	28.0323	382.955	2
49	20.9566075	94	12.305	2.853119333	11.2406	8.412175	23.1177	573.63	2
...	...	...	...	...	...	...	...	...	...

ตารางที่ 2. ตารางแสดงแอททริบิวต์ที่เป็นคุณลักษณะของกลุ่มตัวอย่างในชุดข้อมูล Breast Cancer Coimbra

คุณลักษณะ	คำอธิบาย
Age (years)	อายุ
BMI (kg/m ²)	ค่าดัชนีมวลกาย หาได้จากน้ำหนักตัวเป็นกิโลกรัม หารด้วยส่วนสูงเป็นเมตรยกกำลังสอง
Glucose (mg/dL)	ความเข้มข้นของน้ำตาลกลูโคสในเลือด
Insulin (μU/mL)	ความเข้มข้นของอินซูลินในเลือด
HOMA	ความคัดค้านของอินซูลินของร่างกาย
Leptin (ng/mL)	ความเข้มข้นของฮอร์โมนเลปตินในเลือด
Adiponectin (μg/mL)	ความเข้มข้นของฮอร์โมนอดิโปเนคตินในเลือด
Resistin (ng/mL)	ความเข้มข้นของฮอร์โมนรีซิสตินในเลือด
MCP-1 (pg/dL)	ความเข้มข้นของอินโมโนไซต์เคมีแอตแทรกแตนท์โปรตีน-1



รูปที่ 1. วิธี 5-fold cross validation

ตารางที่ 3 แสดงผลการทดลองเปรียบเทียบระหว่างอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง อัลกอริทึมการจำแนกประเภทแบบเคมีน และโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Backpropagation Neural Network) ซึ่งเป็นอัลกอริทึมทางปัญญาประดิษฐ์ที่ได้รับการยอมรับในวงกว้างว่ามีประสิทธิภาพสูง จากผลการทดลองพบว่า เมื่อเราได้เพิ่มคุณสมบัติในการปรับค่าถ่วงน้ำหนักด้วยตัวเองให้กับอัลกอริทึมการจำแนกประเภทแบบเคมีนเปอร์เซ็นต์ความถูกต้องของการพยากรณ์จะสูงขึ้นมากเมื่อเทียบกับกรณีที่ไม่มีการปรับค่าถ่วงน้ำหนักด้วยตัวเองในทุกๆ ครั้งของการทดสอบ เมื่อสังเกตจากค่าเฉลี่ยของการทดสอบทั้ง 5 ครั้งพบว่าค่าความถูกต้องเฉลี่ยของอัลกอริทึมการจำแนกประเภทแบบเคมีนเป็น 56.37% ส่วนค่าความถูกต้องเฉลี่ยของอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเองเป็น 70.00% นอกจากนี้แล้วอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง

ก็ยังมีค่าความถูกต้องเฉลี่ยสูงกว่าโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับอยู่ถึง 7.27%

ตารางที่ 4 เป็นตารางเปรียบเทียบค่าความถูกต้องเฉลี่ยในการพยากรณ์โรคมะเร็งเต้านมของอัลกอริทึมทางปัญญาประดิษฐ์ 7 อัลกอริทึม โดยผู้วิจัยได้ทำการเขียนโปรแกรมและทำการทดลองเองสำหรับ 3 อัลกอริทึมแรก ส่วนอีก 4 อัลกอริทึมที่เหลือ ได้แก่ k-nearest neighbors (k-NN), Nonlinear support vector machine (Nonlinear SVM), Decision tree และ Naïve Bayes ผู้วิจัยได้นำค่าความถูกต้องเฉลี่ยมาจกบทความ [12] จากค่าความถูกต้องเฉลี่ยในตารางที่ 4 จะเห็นว่าอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเองได้ค่าความถูกต้องเฉลี่ยสูงสุด ตามมาด้วย Decision tree ที่ได้ค่าความถูกต้องเฉลี่ยน้อยกว่าเล็กน้อย อันดับ 3, 4, 5, 6 และ 7 จะเป็นโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ, Naive Bayes, Nonlinear SVM, k-NN และอัลกอริทึมการจำแนกประเภทแบบเคมีน ตามลำดับ

ตารางที่ 3. ความถูกต้องในการพยากรณ์โรคมะเร็งเต้านมของอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเองเปรียบเทียบกับอัลกอริทึมการจำแนกประเภทแบบเคมีน และ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

	ความถูกต้อง (%)		
	อัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง	อัลกอริทึมการจำแนกประเภทแบบเคมีน	โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ
ครั้งที่ 1	77.27	54.55	68.18
ครั้งที่ 2	63.64	59.09	50.00
ครั้งที่ 3	68.18	50.00	59.09
ครั้งที่ 4	72.73	54.55	77.27
ครั้งที่ 5	68.18	63.64	59.09
Mean	70.00	56.37	62.73
S.D.	5.18	5.18	10.36

ตารางที่ 4. ค่าความถูกต้องเฉลี่ยในการพยากรณ์โรคมะเร็งเต้านมของอัลกอริทึมทางปัญญาประดิษฐ์ 7 อัลกอริทึม

อัลกอริทึม	ความถูกต้องเฉลี่ย (%)
อัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง	70.00
อัลกอริทึมการจำแนกประเภทแบบเคมีน	56.37
โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ	62.73
k-NN	58.14
Nonlinear SVM	60.38
Decision tree	69.28
Naive Bayes	62.38

5. สรุปผลการทดลอง

บทความนี้ผู้วิจัยได้นำเสนออัลกอริทึมทางปัญญาประดิษฐ์ 2 อัลกอริทึมเพื่อใช้ในการพยากรณ์โรคมะเร็งเต้านม ได้แก่ อัลกอริทึมการจำแนกประเภทแบบเคมีน และอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง ในการทดสอบประสิทธิภาพของอัลกอริทึมทั้งสองผู้วิจัยได้ใช้ชุดข้อมูล Breast Cancer Coimbra ซึ่งดาวน์โหลดมาจาก UCI Machine Learning Repository ผลการทดสอบเปรียบเทียบกับอัลกอริทึมทางปัญญาประดิษฐ์อื่น 5 อัลกอริทึมแสดงให้เห็นว่าการที่อัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเองมีความสามารถปรับเพิ่มหรือลดค่าถ่วงน้ำหนักของคุณลักษณะเด่นในสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของประเภทแต่ละประเภทให้เหมาะสมได้ด้วยตัวเองในระหว่างการเรียนรู้ชุดข้อมูล ทำให้อัลกอริทึมนี้สามารถที่จะพยากรณ์โรคมะเร็งเต้านมได้ความถูกต้องสูงที่สุด ส่วนการวิจัยในอนาคตผู้วิจัยได้ตั้งเป้าหมายไว้ว่าจะสามารถนำข้อมูลที่มีขนาดใหญ่จากฐานข้อมูลผู้ป่วยมะเร็งเต้านมในประเทศไทยมาใช้ในการสร้างโมเดลการพยากรณ์ และจะนำโมเดลการพยากรณ์นี้ขึ้นใช้งานผ่านเว็บไซต์เพื่อให้เป็นประโยชน์ต่อประชาชนไทยในวงกว้าง

เอกสารอ้างอิง

- [1] National Cancer Institute, Department of Medical Services, Ministry of Public Health, Thailand, *Hospital-based cancer registry*. Bangkok: National Cancer Institute, 2013.
- [2] National Cancer Institute, Department of Medical Services, Ministry of Public Health, Thailand, “*National Cancer Institute Strategic Plan, 2019 to 2022*,” Bangkok: National Cancer Institute, 2020.
- [3] A. Ratanawichitrasin, “What is the difference between breast ultrasound and mammography,” *Siriraj E-public Library*, July, 2020. [Online]. Available: <https://www.si.mahidol.ac.th/sidoctor/epl/article/detail.asp?id=307>. [Accessed July. 10, 2020].
- [4] Khonkaen Ram Hospital, “Breast cancer screening using digital mammography,” *Khonkaen Ram Hospital*, July, 2020. [Online]. Available: <http://www.khonkaenram.com/th/services/health-information/healtharticles/mammogram>. [Accessed July. 10, 2020].
- [5] R. Siegel, J. Ma, Z. Zou, and A. Jemal, “Cancer statistics,” *CA Cancer J Clin*, Vol. 64, No. 1, pp. 9-29, 2014.
- [6] Siriraj Piyamaharajkarun Hospital, “Breast cancer treatment,” *Siriraj Piyamaharajkarun Hospital*, July, 2020. [Online]. Available: <https://www.siphhospital.com/th/news/article/share/1002/Breast-Cancer-Treatment>. [Accessed July. 10, 2020].
- [7] National Cancer Institute, Department of Medical Services, Ministry of Public Health, Thailand, *Hospital-*

- based cancer registry*. Bangkok: National Cancer Institute, 2010.
- [8] S. Senawong, "Types of tumor markers," *Siriraj E-public Library*, July, 2020. [Online]. Available: <https://www.si.mahidol.ac.th/sidoctor/e-pl/articledetail.asp?id=618>. [Accessed July. 10, 2020].
 - [9] S. Senawong, "Can a blood test detect cancer?," *Siriraj E-public Library*, [Online]. Available: <https://www.si.mahidol.ac.th/sidoctor/e-pl/articledetail.asp?id=619>. [Accessed July. 10, 2020].
 - [10] F. Sardouk, A. D. Duru, and O. Bayat, "Classification of breast cancer using data mining," *American Scientific Research Journal for Engineering, Technology, and Sciences*, Vol. 51, No. 1, pp. 38-46, 2019.
 - [11] R. Ray, A. A. Abdullah, D. K. Mallick, and S. R. Dash, "Classification of benign and malignant breast cancer using supervised machine learning algorithms based on image and numeric datasets," *Journal of Physics: Conference Series*, Vol. 1372, 2019, 012062, DOI:10.1088/1742-6596/1372/1/012062, 2019.
 - [12] Y. D. Austria, J. P. Lalata, L. B. Sta. Maria, Jr., J. E. E. Goh, M. L. I. Goh, H. N. Vicente, "Comparison of machine learning algorithms in breast cancer prediction using the coimbra dataset," *International Journal of Simulation: Systems*, DOI: 10.5013/IJSSST.a.20.S2.23, 2019.
 - [13] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seica, and F. Caramelo, "Using resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, Vol. 18, No. 1, 2018.
 - [14] D. Dua and C. Graff, *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science, 2019.