JIST Journal of Information Science and Technology
Volume 11, NO 2 | JUL – DEC 2021 | 72-83
ISSN: 2651-1053 (Online)
ISSN: 1906-9553 (Print)

การสร้างตัวแบบจำแนกเอพิโทปของเซลล์มะเร็ง

Cancer Epitope Classification

มานนท์ บุญบางยาง และ ศรายุทธ นนท์ศิริ*

Manon Boonbangyang and Sarayut Nonsiri*

ห้องปฏิบัติการวิจัยปัญญาประดิษฐ์และอินเทอร์เน็ตของสรรพสิ่ง

หลักสูตรเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ, สถาบันเทคโนโลยีไทย-ญี่ปุ่น

แขวงสวนหลวง เขตสวนหลวง กรุงเทพมหานคร ประเทศไทย 10250

Information of Technology, Faculty of Information of Technology, Thai-Nichi Institute of Technology,

Suan Luang, Bangkok Thailand 10250

Received: November 25, 2021; Revised: December 17, 2021; Accepted: December 23, 2021; Published: December 28, 2021

ABSTRACT – Cancer is a leading cause of death in the world. In 2020 World Health Organization (WHO) reported that approximately 10 million deaths caused by cancer and will increase for the coming years. This research paper aims to study the prediction of cancer epitope using machine learning for classifying between cancer cell surface and epitope on healthy cell surface. The comparison between the different machine learning algorithms is presented. This work can help to training T-cell for recognizing cancer cell and release enzyme to kill cancer cell (Targeted Therapy). The experiment results shown that imbalance data the model from Support Vector Machine (SVM) calculated based on Dipeptide Composition (DPC) feature achieved the best accuracy of 79% Sensitivity 16% and Specificity 100% on test dataset. While balance data with SMOTE Random Forest (RF) calculated based on Dipeptide Composition (DPC) feature achieved the best accuracy of 80% Sensitivity 28% and Specificity 96% on the same test dataset. In conclusion, Support Vector Machine (SVM) and Random Forest (RF) calculated based on Dipeptide Composition (DPC) feature can employ these models for predicting the cancer epitope in imbalance dataset and balanced dataset.

KEYWORDS: Epitope, Machine Learning, Precision Medicine, Cancer vaccine, SMOTE

บทคัดย่อ -- มะเร็งเป็นสาเหตุการเสียชีวิตลำดับต้น ๆ ของโลก โดยองค์การอนามัยโลก (WHO)[1,2] รายงานในปี 2020 มีผู้เสียชีวิตจากมะเร็ง ประมาณ 10 ล้านรายและคาดการว่าจะเพิ่มมากขึ้นในปีถัดไปโดยผู้ป่วยส่วนใหญ่อยู่ในกลุ่มประเทศที่มีรายได้ปานกลางไปจนถึงรายได้ต่ำเมื่อตรวจพบมักเป็นระยะที่รุนแรง มะเร็งบางชนิดหากตรวจพบได้เร็วก็มีโอกาสรักษาให้หายขาดได้การรักษาในปัจจุบันทำได้โดยการฉายแสง การผ่าตัด การใช้เคมีบำบัด ไปจนถึงการรักษาด้วย วัคซีนที่มีความจำเพาะสูง [2]งานวิจัยชิ้นนี้ได้นำเสนอตัวแบบทำนายเอพิโทป (Epitope) ของเอนติเจน (Antigen) บนผิวเซลล์ของผู้ป่วยมะเร็งเทียบกับเอพิโทปบนผิวเซลล์ของผู้ที่มีสุขภาพดีข้อดีของการทำนายเอพิโทปเซลล์มะเร็งได้จะช่วยให้สามารถนำไปฝึกเซลล์ภูมิคุ้มกัน (T-cell) เพื่อเป็นวัคซีนรักษามะเร็งแบบจำเพาะเจาะจง (Precision Medicine) ในงานวิจัยนี้ทดลองใช้คุณลักษณะ (Feature) จำนวน 5 คุณลักษณะวิธีจำแนกข้อมูลแบบ binary classes จำนวน 7 ขั้นตอนวิธี ผลการทดสอบ

*Corresponding Author: *bo.manon_st@tni.ac.th*

ประสิทธิภาพพบว่า ข้อมูลที่ยังไม่ได้ปรับสมดุล ตัวแบบที่ได้จาก คุณลักษณะองค์ประกอบกรดอะมิโนคู่ (DPC) ที่ใช้วิธีจำแนก ซัพพอร์ทเวกเตอร์แมชชีน (SVM) สามารถทำนายข้อมูลทดสอบมีค่าความแม่นยำสูงสุด 79% ค่าความไว 16% และค่าความจำเพาะ 100% ในข้อมูลทดสอบขณะที่ ข้อมูลที่ปรับสมดุลด้วย เทคนิค SMOTE สมดุล ตัวแบบที่ได้จาก คุณลักษณะองค์ประกอบกรดอะมิโนคู่ (DPC) ที่ใช้วิธีจำแนกป่าสุ่ม (RF) มีค่าความแม่นยำสูงสุดที่ 80% ค่าความไว 28% และค่าความจำเพาะ 96% ในข้อมูลทดสอบ จากผลการทดสอบข้างต้นแสดงให้เห็นว่าคุณลักษณะองค์ประกอบกรดอะมิโนคู่ (DPC) เมื่อใช้ร่วมกับ วิธีจำแนก ซัพพอร์ทเวกเตอร์แมชชีน(SVM) หรือ วิธีจำแนกป่าสุ่ม (RF) สามารถนำมาใช้ทำนายเอพิโทปของเซลล์มะเร็งได้ทั้งในข้อมูลที่ยังไม่ได้ปรับสมดุลและปรับสมดุลแล้วตามลำดับ

คำสำคัญ: เอพิโทป, การเรียนรู้ของเครื่อง, การรักาแบบแม่นยำและจำเพาะ, วัคซีนมะเร็ง, SMOTE

1. บทนำ

มะเร็งเป็นสาเหตุการเสียชีวิตลำดับต้น ๆ ของโลกโดยองค์การอนามัยโลก (WHO) รายงาน ในปี 2020 มีผู้เสียชีวิตจากมะเร็งประมาณ 10 ล้านรายคาดการว่าจะเพิ่มมากขึ้นในปีถัดไปโดยผู้ป่วยส่วนใหญ่อยู่ในกลุ่มประเทศที่มีรายได้ปานกลางไปจนถึงรายได้ต่ำ [1]

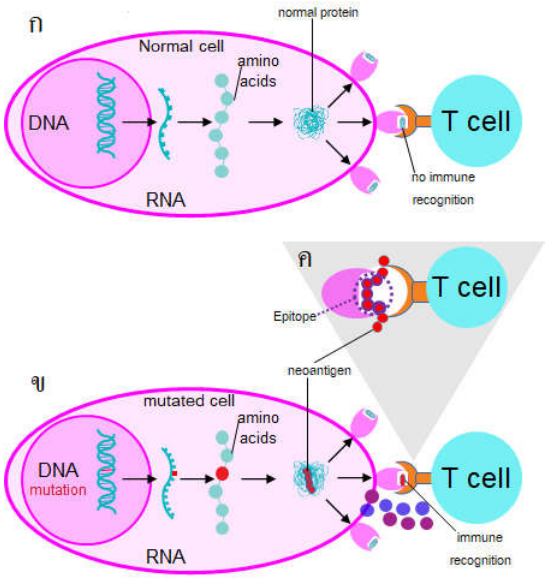
มะเร็งมักจะถูกตรวจพบเมื่อเป็นระยะสุดท้ายแต่มะเร็งหลายชนิดสามารถรักษาให้หายได้หากตรวจพบในระยะเริ่มแรกการรักษาในปัจจุบันสามารถทำได้โดยการฉายแสง การผ่าตัด การใช้เคมีบำบัด ไปจนถึงการรักษาด้วยวัคซีนที่มีความจำเพาะสูง ปรกติเซลล์ภูมิคุ้มกันของร่างกายสามารถกำจัดเซลล์มะเร็งบางส่วนได้แต่เซลล์มะเร็งเกิดจากการกลายพันธุ์ของเซลล์ปรกติซึ่งมีแอนติเจนบนผิวเซลล์ส่วนใหญ่คล้ายเซลล์ปรกติทำให้เซลล์ภูมิคุ้มกันเข้าใจผิดว่าเป็นเซลล์ปรกติจึงไม่ปล่อยสารเพื่อกำจัดเซลล์มะเร็งได้

หากสามารถทำนายได้ว่าเอพิโทปบนแอนติเจนของเซลล์มะเร็งมีลำดับ กรดอะมิโนอย่างไรจะเป็นประโยชน์ในการพัฒนาและกระตุ้นเซลล์ภูมิคุ้มกันให้สามารถตรวจจับและกำจัดเซลล์มะเร็งได้อย่างแม่นยำและมีประสิทธิภาพ

งานวิจัยนี้มุ่งเน้นหาคุณลักษณะและตัวแบบที่เหมาะสมสำหรับทำนาย เอพิโทปเพื่อนำมาใช้สำหรับพัฒนาวัคซีนรักษามะเร็ง

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

แอนติเจน หมายถึง สิ่งแปลกปลอมหรือสารที่ถูกสร้างโดยร่างกายหรือรับมาจากภายนอกร่างกาย ประกอบขึ้นจากกรดอะมิโนหลายชนิดเป็นสายมีทั้งแบบสั้นที่เรียกว่า เปปไทด์จนถึงสายยาวเป็นโปรตีนซึ่งมีคุณสมบัติหลากหลายขึ้นอยู่กับชนิดของกรดอะมิโนดังแสดงในตารางที่ 1 โดยปรกติทีเซลล์ทำหน้าที่เป็นภูมิคุ้มกันจะ ไม่มีปฏิกิริยากับแอนติเจนที่ร่างการผลิต แอนติเจนบนผิวของเซลล์ปรกติแต่บนผิวของเซลล์มะเร็งจะพบ แอนติเจน ที่เกิดการกลายพันธุ์ที่เรียกว่านีโอแอนติเจนหากทีเซลล์ตรวจพบจะทำปฏิกิริยากับนีโอแอนติเจนและปล่อยเอนไซม์เพื่อทำลายเซลล์มะเร็งเนื่องจากนีโอแอนติเจนมีอัตราการพบน้อยมากบนผิวของเซลล์มะเร็งทำให้ทีเซลล์ตรวจพบแต่แอนติเจนที่ร่างการผลิตจึงไม่ทำ



รูปที่ 1. แสดง (ก) เซลล์ปรกติ (ข) เซลล์กลายพันธ์ (ค) เอพิโทปบนนีโอแอนติเจน

ปฏิกิริยาเป็นสาเหตุให้เซลล์มะเร็งไม่ถูกทำลายและแพร่ กระจายอย่างรวดเร็วโดยบางส่วนของแอนติเจน ที่ทำปฏิกิริยาแบบจำเพาะกับแอนติบอดี้จนเกิดการตอบสนองของ ภูมิคุ้มกันของร่างกายเรียกว่า Antigenic determinant หรือ เอพิโทป ตำแหน่งที่แอนติบอดี้เข้าจับกับเอพิโทป เรียกว่า พาราโทป

2.1 การเรียนรู้ของเครื่อง (Machine Learning)

การทำนายแอนติเจนของเซลล์มะเร็งเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) แบบ 2 กลุ่มคือใช่แอนติเจนหรือไม่ใช่แอนติเจน (binary classes) วิธีจำแนก (Classification Algorithm) ที่นำมาใช้ในงานวิจัยนี้มีจำนวน 7 ขั้นตอนวิธีดังนี้

2.1.1 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจเป็นเทคนิคที่สร้างโมเดลการจำแนกข้อมูลในรูปแบบของต้นไม้ประกอบด้วยโหนดราก (Root node) กิ่ง (Branch) และโหนดใบ (Leaf node) ซึ่งจะพิจารณาจากข้อมูลคุณลักษณะและค่าต่าง ๆ ที่กำหนดให้ โดยผลลัพธ์ของโครงสร้างต้นไม้จะแสดงขั้นตอนวิธีการตัดสินใจที่มนุษย์สามารถเข้าใจได้ สำหรับขั้นตอนการคำนวณสามารถพิจารณาได้จากค่า Gini impurity หรือค่าเอนโทรปี (Entropy) ในการเลือกคุณลักษณะและค่าที่เหมาะสมในการสร้างโมเดลโดยสามารถแสดงได้จากสมการที่ (1) และ (2)

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} p_i^2 \quad (1)$$

$$Entropy = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (2)$$

เมื่อกำหนดให้ p_i คือค่าความน่าจะเป็นของกลุ่ม i

ในงานวิจัยนี้ใช้เทคนิคต้นไม้ตัดสินใจประเภท Classification and Regression Trees (CART) [4] เนื่องจากเหมาะสำหรับการจำแนกแบบไบนารี (Binary classification) ซึ่งจะใช้ Gini impurity ในการพิจารณาเลือกคุณลักษณะและค่าที่เหมาะสมเพื่อนำมาสร้างกฎ

2.1.2 เกรเดียนท์บูทติ้ง (Gradient Boosting)

เกรเดียนท์บูทติ้งเป็นเทคนิคที่พัฒนาต่อยอดมาจากเทคนิคต้นไม้ตัดสินใจเพื่อนำมาใช้แก้ปัญหาการวิเคราะห์การถดถอยและการจำแนกข้อมูลได้หลักการคือจะทำการการสร้างต้นไม้ตัดสินใจตามจำนวนรอบที่กำหนดโดยในแต่ละรอบจะมีประสิทธิภาพที่สูงขึ้นเนื่องจากมีการเรียนรู้ข้อผิดพลาดจากการทำนายรอบก่อนหน้า [5]

2.1.3 เคเนียเรสเนเบอร์ (K-Nearest Neighbors)

เคเนียเรสเนเบอร์เป็นเทคนิคการจำแนกข้อมูลประเภทหนึ่ง (Classification) โดยที่ไม่มีข้อมูลชุดที่ใช้ในการเรียนรู้ (Training data) แต่จะใช้วิธีการสุ่มเปรียบเทียบข้อมูลที่สนใจกับข้อมูลอื่น ๆ รอบข้างว่ามีความคล้ายคลึงมากน้อยเพียงใดตามจำนวนข้อมูลที่ล้อมรอบ (K) จำนวนหากพบว่ามีความคล้ายคลึงกันมากข้อมูลนั้นจะถูกอยู่จัดให้อยู่ในกลุ่มเดียวกันโดยการวัดความแตกต่างของข้อมูลจะใช้การวัดระยะทางแบบยุคลิด (Euclidean distance) สามารถแสดงได้ดังสมการที่ (3)

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

เมื่อ $d(x,y)$ คือระยะทางระหว่างข้อมูล x และข้อมูล y ในปริภูมิ n มิติ

2.1.4 การถดถอยโลจิสติค (Logistic Regression)

การถดถอยโลจิสติคเป็นเทคนิคการวิเคราะห์สถิติเชิงคุณภาพโดยมีตัวแปรตามเป็นตัวแปรเชิงคุณภาพสามารถแบ่งการวิเคราะห์เป็นสองแบบ แบบแรกการวิเคราะห์การถดถอยโลจิสติคทวิใช้กับตัวแปรที่แบ่งเป็น 2 กลุ่มย่อยมีค่าเป็น 0 กับ 1 และการวิเคราะห์การถดถอยโลจิสติคพหุกลุ่มใช้กับตัวแปรที่มีมากกว่า 2 กลุ่มเพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจซึ่งตัวแปรที่นำมาใช้กับทั้งสองแบบต้องมีความสัมพันธ์กันต่ำตามเกณฑ์ของ Burns and Grove (1993) ค่า r ไม่เกิน 0.65 หรือตามเกณฑ์ของ Stevens (1996) ค่า r ไม่เกิน 0.80

สำหรับการคำนวณความน่าจะเป็นของการเกิดเหตุการณ์ y สามารถพิจารณาได้จากฟังก์ชั่นของตัวแปรทำนายดังสมการที่ (4)

$$p(y) = \frac{1}{1 + e^{-f(x)}} \quad (4)$$

เมื่อกำหนดให้

$p(y)$ คือความน่าจะเป็นของการเกิดเหตุการณ์ y

e คือ exponential function โดยให้ e = 2.71828

f(x) คือฟังก์ชั่นของตัวแปรทำนาย

2.1.5 เนอีฟเบย์ (Naïve Bayes)

เนอีฟเบย์เป็นเป็นเทคนิคการแบ่งกลุ่มข้อมูลโดยใช้หลักความน่าจะเป็น (Probability) ตามทฤษฎีบทของเบย์ (Naïve Bayes Theorem) โดยกำหนดค่าความน่าจะเป็น (Probability) ให้อยู่ระหว่าง 0 ถึง 1 ถ้าค่าความน่าจะเป็นเข้าใกล้หนึ่งหมายถึงโอกาสที่จะเกิดเหตุการณ์ที่

ทำนายมีมากในทางกลับกันหากมีค่าความน่าจะเป็นเข้าใกล้ 0 แสดงว่าโอกาสที่จะเกิดเหตุการณ์ที่ทำนายน้อยมากแต่ค่าความน่าจะเป็นเท่ากับ 0 หมายความว่าเหตุการณ์ที่ทำนายจะไม่มีโอกาสเกิดขึ้นได้สำหรับวิธีการคำนวณสามารถแสดงได้จากสมการที่ (5) ดังนี้

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} \qquad (5)$$

เมื่อกำหนดให้

$P(C \mid A)$ คือความน่าจะเป็นของข้อมูลที่มีคุณลักษณะเป็น A จะอยู่ในคลาส C

$P(A \mid C)$ คือความน่าจะเป็นของข้อมูลที่อยู่ในคลาส C และมีคุณลักษณะเป็น A

$P(C)$ คือความน่าจะเป็นของคลาส C

$P(A)$ คือจำนวนคุณลักษณะทั้งหมด

2.1.6 ป่าสุ่ม (Random Forest) [25]

ป่าสุ่มเป็นเทคนิคการจำแนกที่ได้รับความนิยมในปัจจุบันพัฒนาต่อยอดมาจากวิธีการจำแนกต้นไม้ตัดสินใจ (Decision Tree) หลาย ๆ ต้นเพื่อเพิ่มประสิทธิภาพการทำนายให้แม่นยำมากขึ้น โดยเอาค่าการตัดสินใจของแต่ละตัวแบบมาลงคะแนนเพื่อเลือกกลุ่ม (Class) ไหนถูกเลือกมากที่สุด

2.1.7 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) [26]

ซัพพอร์ทเวกเตอร์แมชชีนเป็นเทคนิคการจำแนกข้อมูลโดยการวิเคราะห์การถดถอย (Regression) และการจัดกลุ่มข้อมูล (Classification) โดยการหาค่าสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งกลุ่มข้อมูลที่ดีที่สุดใช้ได้ทั้งสมการเชิงเส้นตรง (Linear) และสมการไม่เชิงเส้นตรง (Non Linear)

2.2 เทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่ม

Synthetic Minority Oversampling Technique (SMOTE) คือเทคนิคที่ช่วยแก้ปัญหาข้อมูลที่มีจำนวนแตกต่างกันมากใน แต่ละคลาสส่งผลให้ผลลัพธ์ของการจำแนกอาจจะอยู่ใน ข้อมูลกลุ่มที่มีจำนวนมากกว่า เทคนิค SMOTE นี้ทำการ เพิ่มจำนวนของข้อมูลในกลุ่มที่น้อยกว่าให้ เท่ากับกลุ่มที่มีข้อมูลมากกว่า โดยอาศัยหลักการ เคเนียเรสเนเบอร์ (K-Nearest Neighbors) วิธีทำเริ่มจากสุ่มข้อมูลของกลุ่มที่มีจำนวนน้อย มาจำนวน 1 ข้อมูลหลังจากนั้นคำนวณหาระยะทางด้วยวิธี Euclidean distance เพื่อหาข้อมูลที่ใกล้เคียงจำนวน k ข้อมูล สร้างข้อมูลเทียม ที่อยู่ระหว่าง ข้อมูลที่สุ่มกับค่า k และสุ่มเลือกข้อมูลอีกหลายครั้งเพื่อสร้างข้อมูลเทียม จนครบตามจำนวนที่ต้องการ

2.3 ทบทวนงานวิจัยที่เกี่ยวข้อง

ปัจจุบันมีงานวิจัยหลายชิ้นพยายามสร้างตัวแบบสำหรับทำนายเปปไทด์ปะเภทต่าง ๆ เช่นเปปไทด์ต้านมะเร็ง, เปปไทดล์ต้านจุลชีพ และ โปรตีนที่เกี่ยวข้องกับกลไกลต่างๆของร่างกายข้อมูลที่นำมาใช้สำหรับฝึกฝนและทดสอบตัวแบบมาจากหลายแหล่งดังนี้

ในปี 2019 Md. Mehedi Hasan [16] ได้นำเสนอ DiscriminationHLPpred-Fuse: improved and robustprediction of hemolytic peptide and its activity by fusing multiple feature representation Discrimination เพื่อสร้างตัวแบบสำหรับทำนาย hemolytic peptide และ กิจกรรมของเปปไทด์นี้ ซึ่งใช้วิธีจำแนกจำนวน 5 ขั้นตอนวิธีดังนี้ SVM, RF, GB, KNN และ AdaBoost ใช้คุณลักษณะจำนวน 9 คุณลักษณะหลักดังนี้องค์ประกอบกรดอะมิโน Amino Acid Composition (AAC), องค์ประกอบกรดอะมิโนคู่ Dipeptide Composition (DPC) , Amino acid index (AAI), Binary profile (BPF), Composition transition-distribution (CTD), Conjoint triad (CTF), Quasi-sequence order (QSO) , Grouped dipeptide composition (GDPC) และ Grouped tripeptide composition (GTPC) ข้อมูลแบ่งออกเป็น 2 ส่วนคือ Training และ Independent หลังจากนั้นแบ่งอีกครั้งเป็น 2 ชั้น และแบ่งข้อมูลฝึกฝนและทดสอบแบบ 10 fold cross validation

ประเมินประสิทธิภาพด้วยค่าความแม่นยำปรับสมดุล (balanced accuracy (BACC)), ความไว (Sensitivity), ความจำเพาะ (Specificity) , สัมประสิทธิ์ของ Matthews (MCC), พื้นที่ใต้กราฟเส้นโค้ง The area under curve (AUC) และ p-value ผลที่ดีที่สุดได้จากวิธีจำแนก PTPD มีค่าความไวสูงกว่า ตัวแบบที่เคยตีพิมพ์แล้วจำนวน 1 วิธี ผลสรุปดังตารางที่ 1

ตารางที่ 1. แสดงผลการทดสองเปรียบเทียบตัวแบบ HLPred-Fuse กับ HemoPl

Methods	MCC	BACC	Sn	Sp	AUC	P-value
HLPred --Fuse	0.585	0.792	0.821	0.762	0.869	-
HemoPl	0.577	0.780	0.887	0.673	0.842	0.404

ในปี 2019 Sayamon Hongjaisee, Chanin Nantasenamat, Tanawan Samleerat Carraway และ Watshara Shoombuatong [27] ได้นำเสนอ “HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage” เพื่อสร้างตัวแบบสำหรับทำนาย HIV-1 CRF01_AE coreceptor ซึ่งใช้วิธีจำแนกป่าสุ่ม (Random Forest) และซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) โดยใช้คุณลักษณะ องค์ประกอบกรดอะมิโน (Amino Acid Composition (AAC)), องค์ประกอบกรดอะมิโนเทียม (Pseudo Amino AcidComposition (PseAAC)) และ Relative Synonymous Codon Usage (RSCU) แบ่งข้อมูลฝึกฝนและทดสอบแบบ 10 fold cross validation ข้อมูล benchmark ชนิด R5 Internal set จำนวน 30 และ External set จำนวน120 และ ข้อมูล benchmark ชนิด X4 Internal set จำนวน 30 และ External set จำนวน 10 ประเมินประสิทธิภาพด้วยค่าความแม่นยำ (Accuracy), ความไว (Sensitivity), ความจำเพาะ (Specificity) , สัมประสิทธิ์ของ Matthews (MCC) และ พื้นที่ใต้กราฟเส้นโค้ง The area under curve (AUC) ผลที่ดีที่สุดได้จากวิธีจำแนก PTPD มีค่าความไวสูงกว่า ตัวแบบที่เคยตีพิมพ์แล้วจำานวน 5 วิธี ดังแสดงในตารางที่ 2

ตารางที่ 2. เปรียบเทียบประสิทธิภาพของ 6 ตัวแบบ

Validation	Model	AC (%)	SN (%)	Sp (%)	MCC	AUC
10-fold CV	SVM/ LMT	88.43	90.32	88.11	0.65	0.96
	HIVCoR	95.29	94.38	100.00	0.86	1.00
Internal test	SVM/ LMT	73.13	72.89	73.83	0.46	0.85
	HIVCoR	93.67	93.23	94.30	0.87	0.99
External test	SVM/ LMT	74.24	75.95	73.99	0.36	0.86
	HIVCoR	93.80	93.90	93.79	0.71	0.99

3. วิธีดำเนินการวิจัย

3.1 การเก็บรวบรวมและเตรียมข้อมูล

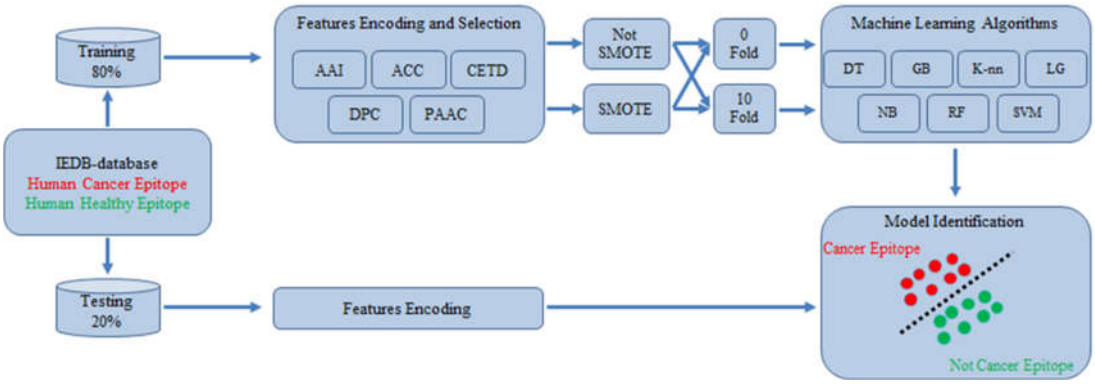
ข้อมูลที่นำมาใช้ในงานวิจัยนี้รวบรวมจาก IEDB (https://www.iedb.org/) ข้อมูลเอพิโทปจากเซลล์ของผู้ป่วยมะเร็งและจากเซลล์ของผู้ที่สุขภาพดีหลังจากนั้นแยกข้อมูลที่ซ้ำซ้อนออกจะได้ข้อมูลที่มีรายละเอียดตามตารางที่ 3. แปลงข้อมูลจาก csv format เปลี่ยนเป็น fasta format เพื่อสกัดคุณลักษณะ

ตารางที่ 3 แสดงจำนวนข้อมูลที่ได้จาก IEDB ตัดข้อมูลที่ซ้ำออกและเพิ่มข้อมูลให้เท่ากันด้วยวิธี SMOTE

แหล่งที่มา (iedb)	จำนวน (เอพิโทป)	จำนวนไม่ซ้ำ (เอพิโทป)	SMOTE
เซลล์ของผู้ป่วยมะเร็ง	947	804	2,421
เซลล์ของผู้ที่สุขภาพดี	2,549	2,421	2,421

3.2 การสกัดคุณลักษณะ

ในงานวิจัยนี้ได้ทำการสกัดคุณลักษณะโดยใช้เครื่องมือ Pfeature [13] จำนวน 5 คุณลักษณะ แต่ละคุณลักษณะประกอบด้วยคุณลักษณะย่อยมีรายละเอียงดังตารางที่ 4 และมีขั้นตอนการดำเนินงานตามรูปที่ 2



รูปที่ 2 แสดงขั้นตอนการเลือกตัวแบบที่เหมาะสมสำหรับทำนายเอพิโทปของเซลล์มะเร็ง

ตารางที่ 4 แสดงกรดอะมิโน 20 ชนิด

Abbreviation Name			Sidechain
Alanine	Ala	A	hydrophobic
Cysteine	Cys	C	hydrophobic
Aspartate	Asp	D	negative
Glutamate	Glu	E	negative
Phenylalanine	Phe	F	hydrophobic
Glycine	Gly	G	hydrophobic
Histidine	His	H	positive
Isoleucine	Ile	I	hydrophobic
Lysine	Lys	K	positive
Leucine	Leu	L	hydrophobic
Methionine	Met	M	hydrophobic
Asparagine	Asn	N	polar
Proline	Pro	P	hydrophobic
Glutamine	Gln	Q	polar
Arginine	Arg	R	positive
Serine	Ser	S	polar
Threonine	Thr	T	polar
Valine	Val	V	hydrophobic
Tryptophan	Trp	W	hydrophobic
Tyrosine	Tyr	Y	polar

3.2.1 องค์ประกอบกรดอะมิโน (Amino Acid Composition (AAC))

องค์ประกอบกรดอะมิโน คือการคำนวนหาความถี่ของกรดอะมิโนทั้ง 20 ชนิดในสายเปปไทด์ โดยสามารถคำนวนได้จากสมการที่ 6

$$\boldsymbol{f(x) = \frac{B(x)}{L}} \quad \boldsymbol{x \in \{A, C, D, \ldots, Y\}} \qquad (6)$$

3.2.2 องค์ประกอบกรดอะมิโนคู่ (Dipeptide Composition (DPC))

องค์ประกอบกรดอะมิโนคู่ คือการเรียงสับเปลี่ยนกรดอะมิโน 20 ชนิดทีละคู่สามารถซ้ำกันได้ ทำให้เกิดคุณลักษณะใหม่จำนวน 400 คุณลักษณะ สามารถคำนวนได้จากสมการที่ 7

$$\boldsymbol{f(x, y) = \frac{B(x, y)}{L - 1}} \quad \boldsymbol{x \in \{A, C, D \ldots Y\}} \qquad (7)$$

3.2.3 ดัชนีกรดอะมิโน (Amino acid index (AAI))

ดัชนีกรดอะมิโน คือชุดค่าตัวเลข 20 ค่าของคุณสมบัติทางเคมีกายภาพและชีวภาพที่แตกต่างกันของกรดอะมิโน AAindex1 ทำการรวบรวมดัชนีที่เผยแพร่พร้อมกับผลลัพท์ของคลัสเตอร์มาทำการวิเคราะห์โดยใช้สัมประสิทธิ์สหสัมพันธ์เป็นระยะห่างระหว่าง 2 ดัชนีไว้ทั้งหมด 556 ดัชนี เช่น Hydrophobicity index (Argos et al., 1982), alpha-CH chemical shifts (Andersen et al., 1992), Signal equence helical potential (Argos et al., 1982), Residue volume (Bigelow, 1967)

3.2.4 Composition enhanced Transition and Distribution (CETD)

Composition enhanced Transition and Distribution (CETD) แสดงคุณสมบัติของเปปไทด์ ไม่ชอบน้ำ (Hydrophobicity), ปริมาตรของ Waals (Norm Waals Volume), ขั้ว (Polarity),

ความสามารถในการโพลาไรซ์ (Polarizability), ประจุ (Charge), โครงสร้างรอง (Secondary Structure) และการเข้าถึงตัวทำละลาย (Solvent Access) ของการเข้ารหัสสามประเภท ดังนี้

3.2.4.1 องค์ประกอบ (C) คือจำนวนกรดอะมิโนของคุณสมบัติ เฉพาะเช่น ความไม่ชอบน้ำ หารด้วยจำนวนกรดอะมิโนทั้งหมด

3.2.4.2 การเปลี่ยนภาพ (T) คือลักษณะความถี่ร้อยละที่กรดอะมิโนของคุณสมบัติเฉพาะตามด้วยกรดอะมิโนที่มีคุณสมบัติต่างกัน

3.2.4.3 การกระจาย (D) คือวัดความยาวโซ่ที่กรดอะมิโนตำแหน่งแรก 25, 50, 75 และ 100 ที่คุณสมบัติเฉพาะตั้งอยู่ตามลำดับ

3.2.5 องค์ประกอบกรดอะมิโนเทียม (**Pseudo Amino Acid Composition (PAAC))**

องค์ประกอบกรดอะมิโนเทียมประกอบด้วยคุณลักษณะย่อยจำนวน 20 คุณลักษณะในปี 2544 Kuo-Chen Chou ได้เสนอตัวอย่างโปรตีนสำหรับปรับปรุงการทำนายการโลคัลไลเซชั่นของโปรตีนและทำนายประเภทโปรตีนเมมเบรนเช่นเดียวกับวิธีองค์ประกอบกรดอะมิโน (Amino acid composition (AAC)) วิธีนี้จะใช้เมทริกซ์ความถี่ของกรดอะมิโนเป็นหลักเพื่อกำหนดลักษณะของโปรตีนซึ่งช่วยให้โปรตีนไม่มีความคล้ายคลึงกันของโปรตีนอื่น ๆ เมื่อเปรียบเทียบกับ วิธีองค์ประกอบกรดอะมิโน (ACC) แล้วข้อมูลเพิ่มเติมจะรวมอยู่ในเมทริกซ์เพื่อแสดงคุณลักษณะในท้องถิ่นบางอย่าง เช่น ความสัมพันธ์ระหว่างสิ่งตกค้างในระยะทางที่กำหนด เมื่อจัดการกับกรณีขององค์ประกอบกรดอะมิโนเทียม มักใช้ทฤษฎีบทค่าคงที่ของ Chou

จำนวนคุณลักษณะหลักและคุณลักษณะย่อยที่นำมาใช้ในงานวิจัยนี้แสดงในตารางที่ 5

ตารางที่ 5 แสดงคุณลักษณะหลักและจำนวนคุณลักษระย่อย

คุณลักษณะ	คุณลักษณะย่อย
Amino Acid Composition (AAC)	20
Amino Acid Index (AAI)	553
Composition enhanced Transition and Distribution (CETD)	189
Dipeptide Composition (DPC)	400
Pseudo Amino Acid Composition (PAAC)	20

3.3 การแบ่งข้อมูล

หลังจากนั้นเป็นการทำ cross validation โดยแบ่งข้อมูลออกเป็น 10 ส่วน ซึ่งข้อมูลจำนวน 9 ส่วนจะนำมาใช้สอนเครื่องให้เรียนรู้ (train) ส่วนข้อมูที่เหลืออีก 1 ส่วนจะนำมาใช้สำหรับเป็นข้อมูลทดสอบ (test) โดยสลับข้อมูลทดสอบ จำนวน 10 ครั้ง วิธีนี้สามารถป้องกันความลำเอียง (bias) ที่เกิดจากข้อมูลที่มีการกระจายตัวน้อย ซึ่งส่งผลให้ตัวแบบทำนายข้อมูลที่ใช้ทำนายจริงมีความแม่นยำน้อย (overfitting)

3.4 การประเมินประสิทธิภาพของ ตัวแบบ

ในงานวิจัยนี้ วิธีจำแนกทั้ง 7 วิธี ใช้ค่าเริ่มต้นของแต่ละวิธี โดยไม่มีการปรับค่าใด ๆ การประเมินประสิทธิภาพของตัวแบบแต่ละตัวที่เป็นแบบ สองกลุ่ม (Binary Classes) ใช้ค่าสำหรับการประเมินดังนี้ [16-20]

ความแม่นยำ (Accuracy) คืออัตราของผลรวมถูกทั้งหมดกับผลรวมทั้งหมด ดังสมการที่ 8

ความไว (Sensitivity) คืออัตราผลบวกจริงของการทดสอบหรือความน่าจะเป็นของผลบวกจริง ดังสมการที่ 9

ความจำเพาะ (Specificity) คืออัตราผลลบจริงของการทดสอบหรือ ความน่าจะเป็นของผลลบจริง ดังสมการที่ 10

สัมประสิทธิ์ของMatthews (MCC) เหมาะสำหรับประเมินข้อมูลที่ไม่สมดุลโดยจะให้ความสำคัญกับค่าความสับสนทั้ง 4 ค่าคือผลบวกจริง (TP), ผลลบจริง (TN), ผลบวกลวง (FP) และผลลบลวง (FN) ดังสมการที่ 11 ซึ่งจะต่างจากค่า F1 ที่ไม่สนใจค่าผลลบจริง

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (9)$$

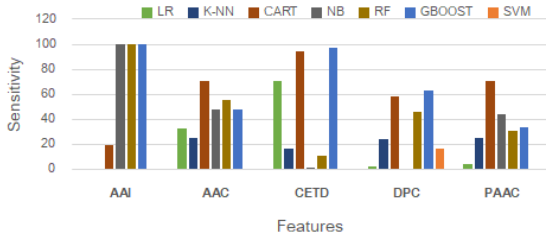
$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (11)$$

หลังจากนั้นนำค่าอัตราผลบวกจริงและอัตราผลลบจริงแสดงในกราฟเส้นโค้ง Receiver Operating Characteristic (ROC) [23] เพื่อหาจุดตัดการแบ่งกลุ่มที่ดีที่สุดและคำนวนหา

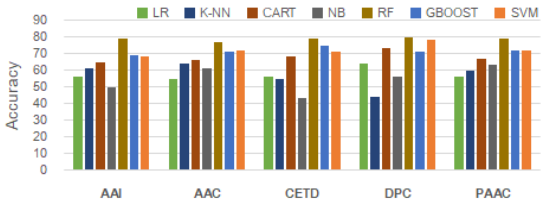
พื้นที่ใต้กราฟ The area under curve (AUC) เพื่อประเมินภาพรวมตัวแบบโดยพื้นที่ใต้กราฟ ยิ่งเข้าใกล้ 1 แสดงว่าตัวแบบนั้นทำนายได้ดีมาก

4. ผลการทดลองและการอภิปราย



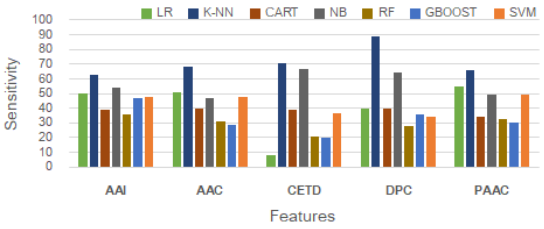
รูปที่ 3. แสดงกราฟเปรียบเทียบค่าค่าความแม่นยำของคุณลักษณะและวิธีจำแนกของข้อมูลที่ไม่ได้ปรับสมดุล

จากรูปที่ 3 แสดงข้อมูลที่ยังไม่ได้ปรับสมดุลวิธีจำแนก SVM มีค่าความแม่นยำสูงที่สุดเมื่อใช้คุณลักษณะ AAI, AAC, CETD และ DPC ในขณะที่คุณลักษณะ PAAC มีค่าความแม่นยำสูงสุดเมื่อถูกใช้โดยวิธีจำแนก KNN



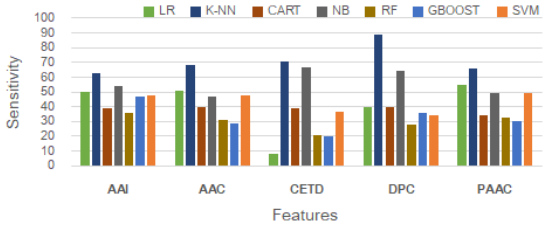
รูปที่ 4. แสดงกราฟเปรียบเทียบค่าค่าความแม่นยำของคุณลักษณะ และวิธีจำแนกของข้อมูลที่ปรับสมดุลด้วย SMOTE

จากรูปที่ 4 แสดงให้เห็นว่าในข้อมูลที่ปรับสมดุลด้วย SMOTE วิธีจำแนก RF มีค่าความแม่นยำสูงที่สุดในทุกคุณลักษณะ



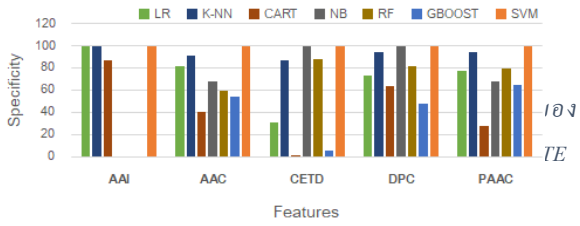
รูปที่ 5. แสดงกราฟเปรียบเทียบค่าค่าความไวของคุณลักษณะและวิธีจำแนกของข้อมูลที่ไม่ได้ปรับสมดุล

จากรูปที่ 5 แสดงให้เห็นว่าในข้อมูลที่ยังไม่ได้ปรับสมดุลวิธีจำแนก GBOOST มีค่าความไวสูงที่สุดเมื่อใช้คุณลักษณะ AAI, CETD และ DPC ในขณะที่คุณลักษณะ AAI มีค่าความไว สูงสุดเมื่อถูกใช้โดยวิธีจำแนก NB, RF และ GBOOST



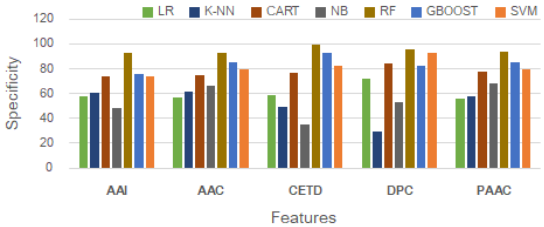
รูปที่ 6. แสดงกราฟเปรียบเทียบค่าค่าความไวของคุณลักษณะและวิธีจำแนกของข้อมูลที่ปรับสมดุลด้วย SMOTE

จากรูปที่ 6 แสดงให้เห็นว่าในข้อมูลที่ปรับสมดุลด้วย SMOTE วิธีจำแนก KNN มีค่าความไวสูงที่สุดในทุกคุณลักษณะ



รูปที่ 7 แสดงกราฟเปรียบเทียบค่าค่าความจำเพาะของคุณลักษณะและวิธีจำแนกของข้อมูลที่ไม่ได้ปรับสมดุล

จากรูปที่ 7 แสดงให้เห็นว่าในข้อมูลที่ยังไม่ได้ปรับสมดุลวิธีจำแนก SVM มีค่าความจำเพาะสูงที่สุดในทุกคุณลักษณะ



รูปที่ 8. แสดงกราฟเปรียบเทียบค่าค่าความจำเพาะของคุณลักษณะและวิธีจำแนกของข้อมูลที่ปรับสมดุลด้วย SMOTE

จากรูปที่ 8 แสดงให้เห็นว่าในข้อมูลที่ปรับสมดุลด้วย SMOTE วิธีจำแนก RF มีค่าความจำเพาะสูงที่สุดในทุกคุณลักษณะ

คุณลักษณะ AAI ข้อมูลไม่ได้ปรับสมดุลวิธีจำแนกการถดถอยโลจิสติค (LR), เคเนียเรสเนเบอร์ (K-NN) และซัพพอร์ตเวกเตอร์แมชชีน (SVM) มีค่า ความแม่นยำสูงสุดที่ 75% ค่าความจำเพาะที่ 100% และมีพื้นที่ใต้กราฟเส้นโค้ง 0.56, 0.51 และ 0.50 ตามลำดับส่วนข้อมูลปรับสมดุลวิธีจำแนกป่าสุ่ม (RF) ค่าความแม่นยำสูงสุดที่ 79% ค่าความจำเพาะที่ 93% และมีพื้นที่ใต้กราฟเส้นโค้ง 0.71

คุณลักษณะ AAC ข้อมูลไม่ได้ปรับสมดุลวิธีจำแนกเคเนียเรสเนเบอร์ (K-NN) และซัพพอร์ตเวกเตอร์แมชชีน (SVM) มีค่าความแม่นยำสูงสุดที่ 75% ค่าความจำเพาะที่ 91% และ 100% ตามลำดับและมีพื้นที่ใต้กราฟเส้นโค้ง 0.66 และ 0.50 ตามลำดับ ส่วนข้อมูลปรับสมดุลวิธีจำแนกป่าสุ่ม (RF) ค่าความแม่นยำสูงสุดที่ 77% ค่าความจำเพาะที่ 93% และมีพื้นที่ใต้กราฟเส้นโค้ง 0.73

ตารางที่ 6*. ผลประเมินประสิทธิภาพคุณลักษณะและตัวแบบ ที่ไม่ได้ปรับสมดุลข้อมูล*

Feature	Classifier	Ac (%)			Sn (%)			Sp (%)			MCC			AUC		
		0 fold	10 fold	test	0 fold	10 fold	test	0 fold	10 fold	test	0 fold	10 fold	test	0 fold	10 fold	test
AAI	LR	75	75	75	3	3	0	99	99	100	0.07	0.07	N/A	0.59	0.59	0.56
	K-NN	73	73	75	24	24	0	90	90	100	0.17	0.17	N/A	0.65	0.65	0.51
	CART	64	64	70	32	32	19	75	75	87	0.07	0.07	0.06	0.55	0.55	0.53
	NB	55	55	25	44	44	100	59	59	0	0.03	0.03	N/A	0.55	0.55	0.50
	RF	77	77	25	12	12	100	98	98	0	0.21	0.21	N/A	0.70	0.70	0.47
	GBOOST	71	71	25	19	19	100	88	88	0	0.10	0.10	N/A	0.66	0.66	0.52
	SVM	77	77	75	11	11	0	98	98	100	0.20	0.20	N/A	0.69	0.69	0.50
AAC	LR	75	75	69	3	3	33	99	99	81	0.08	0.08	0.14	0.60	0.60	0.59
	K-NN	75	75	75	29	29	25	91	91	91	0.25	0.25	0.20	0.67	0.67	0.66
	CART	69	69	48	42	42	71	79	79	40	0.21	0.21	0.10	0.59	0.59	0.56
	NB	73	73	63	22	22	48	90	90	68	0.16	0.16	0.14	0.59	0.59	0.59
	RF	78	78	58	21	21	55	98	98	59	0.32	0.32	0.12	0.72	0.72	0.61
	GBOOST	75	75	52	17	17	48	95	95	54	0.19	0.19	0.01	0.63	0.63	0.51
	SVM	77	77	75	15	15	0	98	98	100	0.25	0.25	N/A	0.65	0.65	0.50
CETD	LR	74	74	41	9	9	71	96	96	31	0.09	0.09	0.01	0.59	0.59	0.49
	K-NN	74	74	69	25	25	16	90	90	87	0.19	0.19	0.04	0.62	0.62	0.55
	CART	66	66	24	37	37	94	76	76	1	0.12	0.12	-0.14	0.57	0.57	0.47
	NB	46	46	75	68	68	1	38	38	100	0.05	0.05	0.03	0.53	0.53	0.54
	RF	78	78	69	17	17	11	98	98	88	0.28	0.28	-0.02	0.66	0.66	0.45
	GBOOST	73	73	28	11	11	97	95	95	5	0.09	0.09	0.04	0.59	0.59	0.53
	SVM	76	76	75	7	7	0	99	99	100	0.15	0.15	N/A	0.71	0.71	0.50
DPC	LR	73	73	65	37	37	42	84	84	73	0.22	0.22	0.13	0.61	0.61	0.60
	K-NN	75	75	77	21	21	24	93	93	94	0.20	0.20	0.26	0.66	0.66	0.68
	CART	73	73	61	37	37	58	84	84	63	0.22	0.22	0.18	0.60	0.60	0.60
	NB	40	40	75	83	83	0	26	26	100	0.09	0.09	N/A	0.58	0.58	0.50
	RF	80	80	72	25	25	46	98	98	81	0.38	0.38	0.26	0.72	0.72	0.71
	GBOOST	74	74	52	27	27	63	90	90	48	0.21	0.21	0.10	0.63	0.63	0.57
	SVM	79	79	79	22	22	16	98	98	100	0.34	0.34	0.34	0.69	0.69	0.73
PAAC	LR	75	75	66	2	2	34	99	99	77	0.04	0.04	0.11	0.59	0.59	0.59
	K-NN	76	76	77	32	32	25	90	90	94	0.28	0.28	0.27	0.61	0.61	0.67
	CART	68	68	39	41	41	71	77	77	28	0.17	0.17	0.00	0.59	0.59	0.50
	NB	73	73	62	23	23	44	90	90	68	0.17	0.17	0.11	0.60	0.60	0.58
	RF	79	79	67	21	21	31	98	98	79	0.32	0.32	0.10	0.71	0.71	0.61
	GBOOST	76	76	57	17	17	34	95	95	65	0.20	0.20	-0.01	0.64	0.64	0.49
	SVM	77	77	75	17	17	0	97	97	100	0.26	0.26	N/A	0.62	0.62	0.50

***ตารางที่** 7. ผลประเมินประสิทธิภาพคุณลักษณะและตัวแบบ ที่ปรับสมดุลข้อมูลด้วยวิธี SMOTE*

Feature	Classifier	Ac (%)			Sn (%)			Sp (%)			MCC			AUC		
		0 fold	10 fold	test	0 fold	10 fold	test	0 fold	10 fold	test	0 fold	10 fold	test	0 fold	10 fold	test
AAI	LR	60	58	56	50	56	50	63	60	58	0.11	0.16	0.07	0.60	0.60	0.60
	K-NN	60	74	61	63	93	63	59	55	61	0.19	0.53	0.21	0.65	0.65	0.67
	CART	66	74	65	40	79	39	74	69	74	0.13	0.48	0.12	0.57	0.57	0.56
	NB	52	55	50	59	62	54	50	47	48	0.08	0.09	0.02	0.56	0.56	0.55
	RF	77	90	79	30	90	36	91	91	93	0.25	0.80	0.36	0.69	0.69	0.71
	GBOOST	67	72	69	34	73	47	77	71	76	0.10	0.44	0.22	0.59	0.59	0.63
	SVM	66	72	68	42	72	48	73	73	74	0.14	0.45	0.20	0.64	0.64	0.68
AAC	LR	56	58	55	54	59	51	56	58	57	0.09	0.17	0.07	0.59	0.59	0.58
	K-NN	63	76	64	68	93	68	61	60	62	0.25	0.56	0.26	0.70	0.70	0.69
	CART	68	76	66	41	77	40	76	74	75	0.17	0.51	0.15	0.59	0.59	0.58
	NB	62	61	61	52	58	47	65	64	66	0.16	0.22	0.11	0.61	0.61	0.60
	RF	77	87	77	30	80	31	92	93	93	0.28	0.74	0.30	0.70	0.70	0.73
	GBOOST	71	79	71	31	73	29	84	84	85	0.16	0.57	0.16	0.57	0.57	0.64
	SVM	69	78	72	38	77	48	79	79	80	0.17	0.56	0.27	0.68	0.68	0.70
CETD	LR	56	62	56	51	63	48	58	60	59	0.07	0.24	0.06	0.55	0.55	0.56
	K-NN	56	71	55	73	96	71	51	47	49	0.20	0.49	0.17	0.67	0.67	0.63
	CART	68	77	68	43	78	39	75	77	77	0.17	0.55	0.16	0.59	0.59	0.58
	NB	44	55	43	68	75	67	37	35	35	0.04	0.10	0.02	0.52	0.52	0.53
	RF	78	89	79	14	79	21	99	98	99	0.28	0.78	0.35	0.73	0.73	0.72
	GBOOST	70	81	75	20	73	20	87	88	93	0.08	0.62	0.18	0.57	0.57	0.58
	SVM	70	76	71	40	74	37	80	78	82	0.19	0.53	0.20	0.65	0.65	0.64
DPC	LR	68	74	64	48	76	40	74	72	72	0.20	0.48	0.11	0.68	0.68	0.62
	K-NN	44	62	44	88	99	89	30	26	29	0.17	0.36	0.19	0.67	0.67	0.66
	CART	72	80	73	44	78	40	80	82	84	0.24	0.60	0.25	0.62	0.62	0.62
	NB	55	67	56	64	85	64	52	50	53	0.14	0.37	0.15	0.61	0.61	0.60
	RF	82	89	80	31	81	28	98	97	96	0.42	0.79	0.36	0.77	0.77	0.74
	GBOOST	72	79	71	39	76	36	82	83	82	0.21	0.58	0.19	0.65	0.65	0.60
	SVM	78	88	78	25	82	34	95	93	93	0.29	0.75	0.34	0.70	0.70	0.67
PAAC	LR	61	56	56	62	55	55	60	57	56	0.20	0.12	0.10	0.63	0.63	0.60
	K-NN	60	78	60	63	95	66	59	60	58	0.20	0.59	0.21	0.64	0.64	0.68
	CART	65	77	67	35	78	34	77	77	78	0.11	0.54	0.12	0.56	0.56	0.56
	NB	63	59	63	56	55	49	66	64	68	0.20	0.19	0.15	0.63	0.63	0.63
	RF	75	88	79	26	82	33	94	93	94	0.28	0.76	0.35	0.67	0.67	0.73
	GBOOST	70	78	72	34	73	30	84	83	85	0.20	0.56	0.17	0.61	0.61	0.64
	SVM	69	79	72	47	78	49	78	80	80	0.25	0.57	0.28	0.68	0.68	0.70

AAI: Amino Acid Index, AAC: amino acid composition, CETD: Composition enhanced Transition and Distribution, DPC: dipeptide composition, PAAC: pseudo amino acid composition
LR: การถดถอยโลจิสติค, K-NN: เคเนียเรสเนเบอร์, CART: ต้นไม้ตัดสินใจ, NB: เนอีฟเบย์, RF: ป่าสุ่ม, GBOOST: เกรเดียนท์บูทติ้ง, SVM: ซัพพอร์ทเวกเตอร์แมชชีน
Ac: ความแม่นยำ, Sn: ความไว, Sp: ความจำเพาะ, MCC: สัมประสิทธิ์ ของ Matthews, AUC: พื้นที่ใต้กราฟเส้นโค้ง

คุณลักษณะ CETD ข้อมูลไม่ได้ปรับสมดุลวิธีจำแนกเนอีฟเบย์ (NB) และซัพพอร์ตเวกเตอร์แมชชีน (SVM) มีค่าความแม่นยำสูงสุดที่ 75% ค่าความจำเพาะที่ 100% และมีพื้นที่ใต้กราฟเส้นโค้ง 0.54 และ 0.50 ตามลำดับส่วนข้อมูลปรับสมดุลวิธีจำแนกป่าสุ่ม (RF) ค่าความแม่นยำสูงสุดที่ 79% ค่าความจำเพาะที่ 99% และมีพื้นที่ใต้กราฟเส้นโค้ง 0.72

คุณลักษณะ DPC ข้อมูลไม่ได้ปรับสมดุลวิธีจำแนกซัพพอร์ตเวกเตอร์แมชชีน (SVM) มีค่าความแม่นยำสูงสุดที่ 79% ค่าความจำเพาะที่ 91% และ 100% ตามลำดับและมีพื้นที่ใต้กราฟเส้นโค้ง 0.66 และ 0.50 ตามลำดับส่วนข้อมูลปรับสมดุลวิธีจำแนกป่าสุ่ม (RF) ค่าความแม่นยำสูงสุดที่ 77% ค่าความจำเพาะที่ 93% และมีพื้นที่ใต้กราฟเส้นโค้ง 0.73

คุณลักษณะ PACC ข้อมูลไม่ได้ปรับสมดุลวิธีจำแนกเคเนียเรสเนเบอร์ (K-NN) มีค่าความแม่นยำสูงสุดที่ 77% ค่าความจำเพาะที่ 94% และมีพื้นที่ใต้กราฟเส้นโค้ง 0.67 ส่วนข้อมูลปรับสมดุลวิธีจำแนกป่าสุ่ม (RF) ค่าความแม่นยำสูงสุดที่ 79% ค่าความจำเพาะที่ 94% และมีพื้นที่ใต้กราฟเส้นโค้ง 0.73

5. ข้อเสนอแนะ

เนื่องจากงานวิจัยชิ้นนี้มุ่งเน้นเพื่อหาตัวแบบสำหรับการทำนายเอพิโทป จากเซลล์ผู้ที่มีสุขภาพดีและเซลล์ผู้ป่วยมะเร็ง ผ่านคุณลักษณะและวิธีการจำแนกข้างต้น ซึ่งยังไม่ครอบคลุมทุกคุณลักษณะ และวิธีจำแนกในอนาคตควรมีการศึกษาคุณลักษณะอื่นรวมถึงคุณลักษณะผสมเพิ่มเติม และศึกษาวิธีการจำแนกการเรียนรู้เชิงลึก (Deep Learning) เพื่อนำมาประยุกต์ใช้สำหรับเพิ่มความแม่นยำในการทำนาย

เอกสารอ้างอิง

[1] World Health Organization, “Cancer,” 21 September 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed Sep. 20, 2021].

[2] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A., “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,’ CA Cancer J Clin, 68(6), pp. 394–424, 2018.

[3] Shahid Akbar, Ateeq Ur Rahman, Maqsood Hayat, Mohammad Sohail, “cACP: Classifying anticancer peptides using discriminative intelligent model via Chou’s 5-step rules and general pseudo components,” Chemometrics and Intelligent Laboratory Systems, Volume 196, 103912, ISSN 0169-7439, 2020

[4] L. Breiman, J. H. Friedman, R. Olshen and C. J. Stone, “Classification and Regression Trees,” Wadsworth International Group, Belmont, California, 1984.

[5] Friedman, J. H., “Greedy Function Approximation: A Gradient Boosting Machine,” Annals of Statistics, 29, pp. 1189-1232, 2000.

[6] Quinlan, J. R. (1986). Induction of decision trees. Machinelearning, 1(1), 81-106, 1986.

[7] Quinlan, J. R. (1993). C4. 5: “programs for machine learning,” (Vol. 1). Morgan kaufmann, 1993.

[8] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B., “The Immune Epitope Database (IEDB): 2018 update,” Nucleic Acids Res. 2018 Oct 24. doi: 10.1093/nar/gky1006. [Epub ahead of print] PubMed PMID: 30357391, 2018.

[9] UniProt, “The universal protein knowledgebase,” Nucleic Acids Res. 45, D158–D169, 2016.

[10] Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, et al: “Human protein reference database as a discovery resource for proteomics.” Nucleic Acids Res, 32 Database: D497-501, 2004.

[11] Yu Wan1, Zhuo Wang and Tzong-Yi Lee1, “Incorporating support vector machine with sequential minimal optimization to identify anticancer peptides: Wan et al. BMC Bioinformatics,” 22:286, 2021.

[12] Wang, L.; Niu, D.; Wang, X.; Khan, J.; Shen, Q.; Xue, Y., “A Novel Machine Learning Strategy for the

Prediction of Antihypertensive Peptides Derived from Food with High Efficiency." Foods, 10, 550, 2021.

[13] Akshara Pande, Sumeet Patiyal, Anjali Lathwal, Chakit Arora, Dilraj Kaur, Anjali Dhall, Gaurav Mishra, Harpreet Kaur, Neelam Sharma, Shipra Jain, Salman Sadullah Usmani, Piyush Agrawal, Rajesh Kumar, Vinod Kumar, Gajendra P.S.Raghava: "Computing wide range of protein/peptide features from their sequence and structure : biorxiv," April 04 2019.

[14] Onkar Singh, Wen-Lian Hsu and Emily Chia-Yu Su., "Co-AMPpred for in silico-aided predictions of antimicrobial peptides by integrating composition-based features," Singh et al. BMC Bioinformatics, 22:389, 2021.

[15] Lei J, Sun L, Huang S, Zhu C, Li P, He J, Mackey V, Coy DH, He Q., "The antimicrobial peptides and their potential clinical applications," Am J Transl Res. 11(7):3919–31, 2019.

[16] Muthuirulan Pushpanathan, Paramasamy Gunasekaran and Jeyaprakash Rajendhran, }Antimicrobial Peptides: Versatile Biological Properties," Hindawi Publishing Corporation International Journal of Peptides, Volume 2013, Article ID 675391, 15 pages, http://dx.doi.org/10.1155/2013/675391, 2013.

[17] Usmani SS, Bhalla S, Raghava GPS., "Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features," Front Pharmacol, 9:954, 2018.

[18] Kao HJ, Nguyen VN, Huang KY, Chang WC, Lee TY., "SuccSite: incorporating amino acid composition and informative k-spaced amino acid pairs to identify protein succinylation sites," Genomics Proteomics Bioinform, 18(2):208–19, 2020.

[19] Huang CH, Su MG, Kao HJ, Jhong JH, Weng SL, Lee TY., "UbiSite: incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines," BMC Syst Biol, 10(Suppl 1):6, 2016.

[20] Chen SA, Lee TY, Ou YY., "Incorporating significant amino acid pairs to identify O-linked glycosylation sites on trans-membrane proteins and non-transmembrane proteins," BMC Bioinform, 11:536, 2010.

[21] Chou KC., "Prediction of protein cellular attributes using pseudo-amino acid composition," Proteins Struct Funct Bioinform. 43(3):246–55, 2001.

[22] Chou K-C., "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," Bioinformatics, 21(1):10–9, 2005.

[23] Hanley JA, McNeil BJ., "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiol-ogy, 143(1), pp. 29–36, 1982.

[24] Epitope human publications, 2021. [Online]. Available: https://github.com/manon089/Epitope_human_papers.git. [Accessed Sep. 20, 2021].

[25] Breiman, L., "Random Forests," Machine Learning 45, 5–32, 2001.

[26] Cortes, Corinna; Vapnik, Vladimir N., "Support-vector networks," Machine Learning. 20 (3), pp. 273–297, 1995.

[27] Sayamon Hongjaisee, Chanin Nantasenamat, Tanawan Samleerat Carraway, Watshara Shoombuatong, "HIVCoR: A sequence-based tool for predicting) HIV-1 CRF01_AE coreceptor usage," Computational Biology and Chemistry, Volume 80, Pages 419-432, ISSN 1476-9271, 2019.