รายงานการวิจัยฉบับสมบูรณ์

การเลือกกลุ่มของสนิปที่มีความสัมพันธ์กับลักษณะพิเศษทางพันธุกรรม

An Approach to Select a Group of Associated SNP with Genetic Traits

รศ. ดร. กิตติ์สุชาต พสุภา

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้คณะประจำปีงบประมาณ พ.ศ. 2561

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

รายงานการวิจัยฉบับสมบูรณ์

การเลือกกลุ่มของสนิปที่มีความสัมพันธ์กับลักษณะพิเศษทางพันธุกรรม

An Approach to Select a Group of Associated SNP with Genetic Traits

รศ. ดร. กิตติ์สุชาต พสุภา

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้คณะประจำปีงบประมาณ พ.ศ. 2561

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อโครงการ	การเลือกกลุ่มของสนิปที่มีความสัมพันธ์กับลักษณะพิเศษทางพันธุกรรม
แหล่งเงิน	เงินรายได้คณะ
ประจำปีงบประมาณ	2561
จำนวนเงินสนับสนุน	50,000 บาท
ระยะเวลาทำการวิจัย	1.5 ปี ตั้งแต่ 1 ตุลาคม พ.ศ. 2560 ถึง 31 มีนาคม พ.ศ. 2562
หัวหน้าโครงการ	รศ. ดร. กิติ์สุชาต พสุภา
คณะ	เทคโนโลยีสารสนเทศ
มหาวิทยาลัย	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

Single-nucleotide polymorphisms (SNPs) เป็นตัวแปรทางพันธุกรรมที่มีความสำคัญ และได้รับความนิยมในการทำวิจัยในด้าน Genome-wide Association Study (GWAS) เป็นอย่างมาก และมักจะถูกนำมาศึกษาเกี่ยวกับอาการของโรคที่มีความเกี่ยวข้องกับพันธุกรรม โดยมีลักษณะเด่นคือ มีตัวแปรจำนวนมาก เนื่องจากเป็นตัวแปรที่ได้มาจากตำแหน่งบนรหัสพันธุกรรม แต่กลุ่มตัวอย่างที่ถูกนำมาใช้ในการวิจัยนั้นมีจำนวนน้อยกว่าจำนวน SNPs ทำให้เกิดปัญหาในการสร้างแบบจำลองที่ใช้ในการจำแนกประเภท นั่นคือ ปัญหาการเข้ากับข้อมูลมากเกินไป (Over-fitting Problem) ดังนั้นการคัดเลือก SNPs เพื่อหาตัวที่มีความสัมพันธ์กับอาการของโรคเป็นสิ่งที่มีความสำคัญอย่างยิ่ง ในงานวิจัยนี้ได้พิจารณาชุดข้อมูล โรคธาลัสซีเมีย ซึ่งเป็นโรคทางพันธุกรรมที่พบเป็นจำนวนมากในประชากรไทย โดยแบ่งประชากรออกเป็นกลุ่มที่เป็นโรครุนแรง และเป็นโรคไม่รุนแรง การทดลองได้ทำการทดสอบวิธีการกรองและเรียงลำดับ SNPs ที่มีความเกี่ยวข้องกับความรุนแรงของโรค ด้วยวิธี χ^2, Information Gain และ Gradient Boosting (GB) และนำ SNPs ที่คัดเลือกแล้วมาสร้างแบบจำลองเพื่อจำแนกความรุนแรงของโรค ด้วยวิธี Support Vector Machine (SVM), GB และ Naïve Bayes ผลการทดลองแสดงให้เห็นว่า วิธีที่ดีที่สุดในการคัดเลือก SNPs และวิธีที่สร้างแบบจำลองที่ดีที่สุดเพื่อจำแนกความรุนแรงของโรคคือ χ^2-SVM และ χ^2-GB ซึ่งมีประสิทธิภาพที่ใกล้เคียงกันและใช้จำนวน SNPs ในแบบจำลองเพียงแค่ 10 ตัว

คำสำคัญ: SNP, GWAS, การเลือกคุณลักษณะ, การลดมิติข้อมูล, การเรียนรู้ของเครื่องจักร

Title	An Approach to Select a Group of Associated SNP with Genetic Traits
Principal Investigator	Assoc. Prof. Dr. Kitsuchart Pasupa
Faculty	Information Technology
University	King Mongkut's Institute of Technology Ladkrabang

Abstract

Single-nucleotide polymorphisms (SNPs) are important genetic variables that are very popular in Genome-wide association study at the present time. They are often used in studies related to genetic disorders. A distinctive trait of SNPs is that there are a lot of them since they are variables originated from various positions in a DNA sequence. Unfortunately, the number of samples investigated are usually far fewer than the number of SNPs and so an over-fitting often occurs when one wants to construct a predictive model for classifying a sample into a case or a control. This study investigated a dataset on beta-thalassemia, a common genetic disorder widely found in Thai population. The data in the set are divided into two groups: severe and mild groups. The aims of the study were to develop and evaluate methods for screening and ranking SNPs related to this disorder. The screening methods tested were Chi-squared test (χ^2), Information Gain, and Gradient Boosting (GB). The SNPs that were screened in and selected were then used to construct a predictive model for classifying a sample to be either a severe or mild case. The model construction methods tested were Support Vector Machine (SVM), GB, and Naïve Bayes. Several combinations of a screening method and a model construction method were evaluated, and the evaluation results show that the best combination was χ^2-SVM which used the number of selected SNPs of 10.

Keywords: SNP, GWAS, Feature Selection, High-Dimensional Reduction, Machine Learning

กิตติกรรมประกาศ

การวิจัยครั้งนี้ได้รับทุนสนับสนุนการวิจัยจากสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง จากแหล่งทุนเงินรายได้คณะ ประจำปีงบประมาณ พ.ศ. 2561

รศ. ดร. กิติ์สุชาต พสุภา

สารบัญ

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในงานวิจัยทางด้านชีวสารสนเทศศาสตร์ (Bioinformatics) ตัวแปรที่นิยมใช้ในการศึกษาปัญหาพันธุกรรม คือ Single Nucleotide Polymorphisms (SNPs) หรือ สนิปส์ ซึ่งเป็นการแปรผันของลำดับดีเอ็นเอที่ทำให้มนุษย์นั้นมีความแตกต่างกัน ซึ่งนำไปสู่ปัญหาต่าง ๆ เช่น โรคทางพันธุกรรม การตอบสนองกับยาชนิดต่าง ๆ ของมนุษย์ เป็นต้น [1] ดังนั้น ถ้าเราสามารถระบุได้ว่าสนิปตัวใดที่ทำให้เกิดปัญหาเหล่านี้ เราก็จะสามารถป้องกันหรือรักษาโรคต่าง ๆ ได้ งานวิจัยส่วนใหญ่ทางด้านชีวสารสนเทศศาสตร์ ต้องการที่จะจำแนกกลุ่มตัวอย่างออกเป็น 2 กลุ่ม นั่นคือ กรณีที่เป็นโรค (Case) และกรณีที่ไม่เป็นโรค (Control)

เนื่องจากสนิปนั้นมีจำนวนตัวแปรที่มีขนาดใหญ่มาก แต่มีจำนวนกลุ่มตัวอย่างที่น้อยมาก เช่น การศึกษาหาสนิปที่มีความสัมพันธ์กับโรค Rheumatoid Arthritis มีจำนวน 545,080 สนิปส์ (868 Cases และ 1,194 Controls) [2] การศึกษาโรค Multiple Sclerosis มีจำนวน 325,807 สนิปส์ (931 Cases และ 2,431 Controls) [3] และการศึกษาโรค Parkinson มีจำนวน 408,803 สนิปส์ (271 Cases และ 270 Controls) [4] ซึ่งเราจะประสบปัญหาที่เรียกว่า High Dimension, Low Sample Size (HDLSS) Problem ซึ่งจะเป็นปัญหาที่มักจะพบบ่อยในงานทางด้านเคมีสนเทศศาสตร์ (Chemoinformatics) และ Microarray Analysis ซึ่งจะมีกลุ่มตัวอย่างเพียงไม่กี่ร้อยตัวอย่าง [5] ในงานทางด้านวิทยาการคอมพิวเตอร์ การสร้างแบบจำลองโดยใช้ข้อมูลที่เป็น HDLSS นั้น มีโอกาสสูงมากที่แบบจำลองจะเกิดปัญหาการเข้ากันกับข้อมูลมากเกินไป (Over-fitting Problem) นั่นคือ โมเดลจะมีประสิทธิภาพที่ดีกับข้อมูลที่ใช้ในการสอน แต่จะมีประสิทธิภาพที่แย่ในข้อมูลชุดทดสอบ ซึ่งแบบจำลองที่ดีส่วนใหญ่นั้นจะถูกสร้างจากข้อมูลที่มีจำนวนตัวแปรที่มีขนาดเล็กกว่าจำนวนของข้อมูลในชุดข้อมูล [6] ดังนั้นการลดขนาดของข้อมูล หรือลดจำนวนตัวแปร (ในที่นี้คือการลดจำนวนสนิป) ที่มีความสัมพันธ์กับโรคนั้นจึงมีความสำคัญเป็นอย่างมาก เพราะว่ามีผลกับประสิทธิภาพของแบบจำลอง และยังใช้เวลาในการประมวลผลที่เร็วอีกด้วย นอกจากนี้มีหลักฐานว่าสนิปส่วนมากนั้น ไม่ได้มีความสัมพันธ์กับโรคที่ทำการศึกษา ดังนั้นการเลือกสนิปจึงเป็นงานที่มีความสำคัญที่สุดในการทำ GWAS (Genome-wide Association Data) [7]

ในการลดขนาดของข้อมูลนั้น สามารถแบ่งได้เป็น 2 ประเภทหลัก คือ

- Feature selection เป็นการคัดเลือกตัวแปรที่มีความสัมพันธ์กับผลลัพธ์ของกลุ่มประชากร ซึ่งง่ายต่อการทำการวิเคราะห์โดยตรงและสามารถอธิบายความสัมพันธ์ระหว่างตัวแปรและผลลัพธ์ได้อย่างง่าย
- Dimension Reduction เป็นการลดตัวแปรด้วยการฉายข้อมูลทั้งหมดไปยังอีกระนาบหนึ่งที่มีมิติที่เล็กลง ซึ่งยังเก็บสาระต่าง ๆ ไว้

อย่างไรก็ตามวิธี Dimension Reduction นั้นไม่สามารถที่จะหาความสัมพันธ์ระหว่างตัวแปรกับผลลัพธ์ได้โดยตรงดังนั้นในการหาสนิปที่มีความสัมพันธ์กับโรคจึงเลือกที่จะใช้วิธี Feature Selection ซึ่งสามารถแบ่งได้เป็น 3 วิธี [8] ได้แก่

1. Filter Method ซึ่งเป็นการหาความสัมพันธ์ระหว่างคุณลักษณะแต่ละตัวกับผลลัพธ์ เพื่อนำมาใช้กรองตัวแปร และใช้เวลาในการคำนวณน้อย เช่น Chi-squared Test [9], Information Gain [10] และ ReliefF [11]

2. Wrapper Method เป็นการสุ่มเลือกชุดของตัวแปรหลาย ๆ ชุด เพื่อหาชุดที่มีความสัมพันธ์ระหว่างผลลัพธ์มากที่สุด ซึ่งส่วนใหญ่ผลลัพธ์ที่ได้จะมีผลที่ดีกว่า Filter Method อย่างไรก็ตามวิธีการนี้มีโอกาสสูงที่จะทำให้แบบจำลองเกิดปัญหา Over-fitting นอกจากนี้ยังใช้เวลาในการประมวลผลที่นานกว่าวิธีอื่น ๆ มาก เช่น การสุ่ม [12, 13], Genetic Algorithm [14], Particle Swarm Optimisation [4] ที่สามารถหาผลลัพธ์ได้ดีที่สุด

3. Embedded Method เป็นการสร้างแบบจำลองที่มีการเลือกคุณลักษณะอัตโนมัติ ซึ่งแบบจำลองจะมีค่าน้ำหนักของตัวแปรที่คำนวนออกมาสำหรับแต่ละตัวแปร ซึ่งบ่งบอกถึงค่าความสำคัญของตัวแปร ซึ่งใช้พลังงานในการคำนวณน้อยกว่าวิธี Wrapper และมีโอกาสที่จะเกิดปัญหาการเข้ากันกับข้อมูลมากเกินไปที่น้อยกว่า งานวิจัยในปัจจุบันนิยมใช้วิธีนี้ เนื่องจากมีประสิทธิภาพที่ดีกว่าวิธีอื่น เช่น Decision Tree [14] และ Random Forest [7, 13]

ในงานวิจัยนี้มีความสนใจเกี่ยวกับโรค β-thalassemia ซึ่งเป็นโรคที่สืบทอดทางพันธุกรรมที่เกิดจากความผิดปกติในการสังเคราะห์ Beta Chain ของ Haemoglobin ที่พบได้มากในประเทศไทย มีผู้ป่วยร้อยละ 1 ของประชากร หรือ 6 แสนคน และมีพาหะของธาลัสซีเมียร้อยละ 40 ของประชากรหรือ 24 ล้านคน [15] โรคนี้สามารถแบ่งระดับความรุนแรงของโรคออกเป็น 2 ระดับคือ รุนแรง (Severe) และ ไม่รุนแรง (Mild) ซึ่งสามารถวัดระดับได้จากคะแนนที่คิดจาก ตัวแปรต่าง ๆ ทั้ง 6 ชนิด นั่นคือ ระดับเฮโมโกลบิน อายุเมื่อเริ่มถ่ายเลือกครั้งแรก ความต้องการในการถ่ายเลือด ขนาดของม้าม อายุเมื่อธาลัสซีเมียแสดงอาการ และการเจริญเติบโตของร่างกาย [13]

ในงานวิจัยนี้เราได้พยายามที่จะหาสนิปที่มีผลต่อระดับความรุนแรงของโรค โดยได้ทำการศึกษาข้อมูลสนิปของของกลุ่มตัวอย่างในประชากรไทย [16] โดยใช้วิธีการเรียนรู้ของเครื่องจักร โดยมีจุดประสงค์ที่จะหาหาเทคนิคที่ใช้ในการคัดเลือกสนิปที่มีความสัมพันธ์กับระดับอาการของ β-thalassemia ที่ดีที่สุด โดยการคัดเลือกสนิปที่มีจำนวนจำกัด และสามารถนำสนิปเหล่านี้ไปสร้างโมเดลจำแนกระดับอาการของโรคที่มีประสิทธิภาพ โดยเปรียบเทียบกับวิธีการ Baseline ต่าง ๆ

1.2 วัตถุประสงค์ของการวิจัย

1. ศึกษาเทคนิคในการเลือกคุณลักษณะต่าง ๆ ที่เหมาะสมกับข้อมูลที่เป็น HDLSS และนำมาใช้ในการคัดเลือกสนิปจากชุดข้อมูล

2. ศึกษาเทคนิคในการสร้างแบบจำลองต่าง ๆ ที่เหมาะสมกับการนำมาใช้กับข้อมูลสนิป และนำมาใช้กับข้อมูลสนิป

3. คัดเลือกสนิปที่มีผลต่อการจำแนกระดับอาการของโรค β-thalassemia เพื่อนำมาใช้ในการศึกษาต่อไป

1.3 ขอบเขตของการวิจัย

1. ศึกษาเฉพาะชุดข้อมูลสนิปของคนไทยจำนวน 618 คน ที่ตรวจสอบว่าเป็นโรค β-thalassemia ซึ่งแบ่งออกเป็น 2 กลุ่ม ตามความรุนแรงของโรค นั่นคือ รุนแรง และ ไม่รุนแรง
2. ศึกษาเทคนิคเลือกคุณลักษณะเฉพาะ Filter Method และ Embedded Method เท่านั้น

1.4 วิธีดำเนินการวิจัย

1. ศึกษาชุดข้อมูลสนิปของคนไทย
2. ทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องกับความสัมพันธ์ระหว่างสนิปและการเกิดโรค β-thalassemia
3. ทดสอบใช้อัลกอริทึมการคัดเลือกคุณลักษณะและการเรียนรู้ของเครื่องจักรต่าง ๆ เพื่อมาใช้เลือกสนิปที่มีความเกี่ยวข้องกับระดับความรุนแรงของโรค β-thalassemia
4. สร้างแบบจำลองและประเมินผล

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

ในหัวข้อนี้จะกล่าวถึงทฤษฎีที่ใช้ในงานวิจัยชิ้นนี้ ซึ่งจะแบ่งออกเป็น 3 ส่วน ได้แก่ วิธีกรองคุณลักษณะที่ใช้ในกระบวนการคัดเลือกคุณลักษณะ อัลกอริทึมการเรียนรู้ของเครื่องจักรที่ใช้ในการจำแนก และ วิธีการวัดประสิทธิภาพในงานนี้

2.1.1 วิธีกรองคุณลักษณะ

2.1.1.1 Chi-squared Test (χ^2)

เป็นการหาความสัมพันธ์ของตัวแปร 2 ตัวว่าเป็นอิสระต่อกันหรือไม่ (Test for Independence) เป็นวิธีที่เหมาะกับตัวแปรที่เป็นตัวแปรแบบกลุ่ม (Categorical Data) [17] กำหนดให้ x คือ สนิปที่ถูกพิจารณาในทุก ๆ ตัวอย่าง โดยที่ $x = \{0, 1, 2\}$ และ y คือ ประเภทของตัวอย่าง โดยที่ $y = \{\text{Case, Control}\}$

χ^2 สามารถคำนวณได้จากการแจกแจงความถี่ (N_{ij}) และ ค่าคาดหวัง (Expected Value: E_{ij}) ดังสมการต่อไปนี้

$$\chi^2 = \sum_{i=1}^{C}\sum_{j=1}^{L}\frac{(N_{ij} - E_{ij})}{E_{ij}}, \tag{2.1}$$

โดยที่ C และ L คือ จำนวนชนิดของตัวแปรคุณลักษณะที่ถูกนำมาทดสอบทั้งสองตัว

2.1.1.2 Information Gain (IG)

เป็นวิธีกรองคุณลักษณะที่ได้รับความนิยมด้วยการหาค่าน้ำหนักของแต่ละคุณลักษณะ โดยหาความสัมพันธ์ระหว่างคุณลักษณะ x และประเภทของข้อมูล y ด้วยการวัดค่าเอนโทรปี (Entropy: $H(x)$) ถ้ามีความใกล้เคียงกันมากเอนโทรปีจะมีค่าต่ำ แต่ถ้ามีความแตกต่างกันมากเอนโทรปีก็จะมีค่าที่สูง ซึ่งสามารถคำนวณได้จาก

$$H(x) = -\sum_{i=1}^{n} P(x_i) \cdot \log_2 P(x_i) \tag{2.2}$$

โดยที่ n คือ จำนวนของประเภทของตัวแปร x, $P(x_i)$ คือความน่าจะเป็นของโอกาสที่จะเกิดเหตุการณ์ x จากนั้น เอนโทรปีของข้อมูลที่พิจารณาร่วมกับประเภท สามารถคำนวณได้จากสมการต่อไปนี้

$$H(x|y) = -\sum_{j=1} P(y_j) \sum_{i=1} P(x_i|y_j) \log_2(P(x_i|y_j)) \tag{2.3}$$

จากนั้น เราสามารถหาค่า IG ได้จาก

$$IG(X|Y) = H(x) - H(x|y) \tag{2.4}$$

2.1.2 การเรียนรู้ของเครื่องจักร

2.1.2.1 Gradiant Boosting (GB)

เป็นการเรียนรู้ของเครื่องจักรชนิดหนึ่งที่ใช้เทคนิคการรวมโมเดล (Ensemble Model) โดยใช้ต้นไม้การตัดสินใจ (Decision Tree) มาสร้าง Collection ของ Weak Predictor ด้วยวิธี Boosting เป็นลำดับต่อเนื่อง ซึ่งมีการปรับค่าพารามิเตอร์ของแบบจำลองโดยใช้ค่าความผิดพลาดที่เกิดขึ้นในแต่ละรอบโดยการวนซ้ำ มีจุดประสงค์เพื่อทำให้ค่าความผิดพลาดนั้นต่ำที่สุด GB สามารถนำมาใช้ในการจำแนกประเภทหรือในการเลือกคุณลักษณะที่มีความสัมพันธ์กับผลลัพธ์ได้ โดยการหาค่าความสำคัญของตัวแปร (Variable Importance) ที่พิจารณาจากความถี่ของคุณลักษณะที่ถูกแบ่งในกระบวนการสร้างต้นไม้ตัดสินใจ จากนั้นผลลัพธ์จะเป็นค่าเฉลี่ยของค่าความสำคัญของตัวแปรของต้นไม้ตัดสินใจทุกต้น [18]

2.1.2.2 Naïve Bayes (NB)

เป็นวิธีที่ใช้ทฤษฎีความน่าจะเป็นเพื่อจำแนกประเภท ถูกนิยมนำมาใช้ในงานด้านชีวสารสนเทศศาสตร์ เนื่องจากสามารถใช้กับข้อมูลที่มีขนาดใหญ่ได้ดีและมีความรวดเร็ว สามารถใช้กับชุดข้อมูลที่มีคุณลักษณะเป็นตัวแปรแบบกลุ่มได้ดี ผลลัพธ์ของการทำนายของ NB เรียกว่า Posterior Probability [19] สามารถคำนวณได้จาก

$$P(x_1, x_2, \ldots, x_n|y) = \prod_{i=1}^{n} P(x_i|y) \tag{2.5}$$

2.1.2.3 Support Vector Machine (SVM)

เป็นวิธีการจำแนกประเภทเพื่อแบ่งข้อมูล 2 ประเภท ด้วยการสร้างระนาบ (Hyperplane) เพื่อแบ่งให้ข้อมูลแยกออกจากกัน ซึ่งอัลกอริทึมสามารถฉายข้อมูลจากปริภูมิดั้งเดิมไปยังอีกปริภูมิหนึ่ง โดยผ่านฟังก์ชันเคอร์เนล (Kernel Function) ทำให้อัลกอริทึมสามารถแบ่งข้อมูลที่เป็นข้อมูลไม่เชิงเส้นได้ SVM จะพยายามหาเวกเตอร์สนับสนุน (Support Vector) เพื่อนำมากำหนดขอบของแต่ละประเภท และพยายามที่จะสร้างระนาบที่สามารถแยกขอบทั้งสองออกจากกันมากที่สุด หรือทำให้มี Margin สูงที่สุด [16]

2.1.3 การวัดประสิทธิภาพ

2.1.3.1 ความแม่นยำ (Accuracy)

เป็นวิธีการวัดผลที่นิยมในการจำแนกข้อมูล เป็นการหาอัตราส่วนของข้อมูลที่ทำนายถูกกับข้อมูลทั้งหมด มีผลลัพธ์ตั้งแต่ 0–1 โดยค่าที่ยิ่งมากยิ่งมีประสิทธิภาพสูง สามารถคำนวณได้จาก

$$Accuracy = \frac{\#CorrectClassification}{\#Samples} \tag{2.6}$$

2.1.3.2 F1-Score

เป็นการวัดค่าประสิทธิภาพของผลลัพธ์ของปัญหาการจำแนกประเภท โดยคำนวณจากค่า Precision และ Recall ทำให้สามารถวัดประสิทธิภาพของการจำแนกข้อมูลได้ชัดเจนยิ่งขึ้น โดยพิจารณาจากความสัมพันธ์ของผลลัพธ์ทั้งจำนวนข้อมูลที่ถูกต้องและข้อมูลที่ทายผิดในแต่ละประเภท ซึ่งคะแนนที่ได้จะมีค่าตั้งแต่ 0–1 โดยประสิทธิภาพที่ดีที่สุดคือ 1 สามารถพิจารณาได้จากค่าต่าง ๆ ใน Confusion Matrix ดังตารางที่ 2.1 โดยนำค่าต่าง ๆ มาคำนวณหาค่า Precision

$$Precision = \frac{TP}{TP + FP} \tag{2.7}$$

และ Recall

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

จากนั้น สามารถหาค่า F1-Score ได้จาก

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{2.9}$$

ตารางที่ 2.1: Confusion Matrix

		Target Values	
		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

2.2 งานวิจัยที่เกี่ยวข้อง

กระบวนการวิจัยและเทคนิคต่าง ๆ ที่ถูกนำมาใช้ในการคัดเลือกสนิปที่มีความสัมพันธ์กับโรค ได้มีการทบทวนและศึกษาจากงานวิจัยต่าง ๆ และสรุปมาในหัวข้อนี้

การคัดเลือกตัวแปรสามารถแบ่งออกเป็น 3 วิธีหลัก ๆ คือ Filter method, Wrapper method, และ Embedded method [8] ซึ่ง Filter method นั้นคือการหาคุณลักษณะที่มีความเกี่ยวข้องหรือมีความซ้ำซ้อนกันระหว่างตัวแปร สำหรับ Wrapper method คือวิธีที่ใช้แบบจำลองเข้ามาช่วยประเมินคุณภาพของของชุดตัวแปรที่ได้รับการคัดเลือก และ Embedded method คือ การสร้างแบบจำลองเพื่อหาค่าน้ำหนักของตัวแปรในแบบจำลองโดยอัตโนมัติ เช่น การทำต้นไม้ตัดสินใจ ซึ่งตัวแปรนั้นถูกเลือกเข้าไปในแต่ละโหนดของต้นไม้ตัดสินใจ ในการทดลองของ [8] ได้ทำการศึกษาการคัดเลือกตัวแปรในข้อมูลยีนที่มีตัวแปรจำนวนมาก โดยทำการทดสอบวิธีต่าง ๆ จำนวน 4 วิธี นั่นคือ Relaxed Linear Separability

(RLS), ReliefF, SVM Recursive Feature Elimination (SVM-RFE), และ Minimum Redundancy Maximum Relevance (MRMR) เพื่อวัดประสิทธิภาพ โดยใช้วิธี Double Cross-validation เพื่อแก้ปัญหาการโน้มเอียงของการคัดเลือกตัวแปร สำหรับ RLS นั้นมีพื้นฐานมาจากการจำแนกประเภทแบบเชิงเส้น โดยการลดทอนระดับในการแบ่งแยกประเภทในเชิงเส้นของตัวแปรที่ถูกเลือก เพื่อเปรียบเทียบกับ ReliefF ซึ่งเป็นการจัดลำดับของตัวแปร ด้วยการหาวัตถุที่ใกล้ที่สุดของแต่ละคลาส และทำการให้ค่าน้ำหนักของตัวแปรตามความแตกต่างของแต่ละวัตถุ สำหรับ SVM-RFE นั้น เป็นวิธีชนิดวนซ้ำ โดยใช้ SVM เพื่อช่วยในการหาค่าน้ำหนักของตัวแปร ส่วน MRMR เป็นวิธีที่นำเสนอโดย [8] ซึ่งใช้เงื่อนไขพิเศษในการจัดลำดับตำแหน่งของตัวแปร ด้วยค่าสหสมพันธ์กับผลลัพธ์ของแบบจำลอง และ ค่าความแตกต่างของแต่ละตัวแปรที่อยู่เหนือขึ้นไปตามลำดับ ซึ่งในการทดลองได้แสดงให้เห็นว่า ReliefF มีความแตกต่างระหว่างชุดตรวจสอบ (Validation Set) และชุดทดสอบ (Test Set) น้อยที่สุดที่ 2.6 % (65.93 % ที่ชุดตรวจสอบ และ 63.33% ที่ชุดทดสอบ) ส่วนความแตกต่างที่มากที่สุดคือ SVM-RFE ที่ 41.67 % (100 % ที่ชุดตรวจสอบ และ 58.33 % ที่ชุดทดสอบ ในชุดทดสอบมีประเภทที่มีจำนวนมากที่สุดถึง 65 %) งานวิจัยนี้ได้เสนอว่า การใช้ชุดข้อมูลเดียวกันสำหรับการทำการเลือกคุณลักษณะและสอนแบบจำลองนั้นไม่เหมาะสม แนะนำให้ใช้การทำการตรวจสอบไขว้ในการคัดเลือกตัวแปร เพื่อช่วยในการลดปัญหาการโน้มเอียง

ในปี 2004 Shah และ Kusiak ได้ใช้วิธีทางเหมืองข้อมูล และ Genetic Algorithm (GA) มาใช้ในการหาสนิปที่มีความสัมพันธ์กับโรค [14] จากการทดลองแสดงให้เห็นว่า ผลลัพธ์จาก GA นั้นดีกว่าการเลือกสนิปโดยใช้ Information Gain และ Standard Regression เป็นอย่างมาก และยังสามารถระบุยีนหรือสนิปที่ไม่สามารถระบุได้จากสองวิธีนี้อีกด้วย

ในปี ค.ศ. 2008 Eppstein และ Haake ได้นำอัลกอริทึม ReliefF มาใช้สำหรับการวิเคราะห์ข้อมูล GWAS ที่มีขนาดใหญ่ เนื่องจากการปฏิสัมพันธ์กันระหว่างตัวแปรทางพันธุศาสตร์กับสิ่งแวดล้อมในโรคทั่วไป (Common Disease) นั้นซับซ้อนและเป็นชนิดไม่เชิงเส้น ซึ่งจำเป็นต้องใช้การคำนวณที่สามารถหาความสัมพันธ์แบบไม่เป็นเชิงเส้นของสนิปที่เกี่ยวข้องกับโรคได้ [20] ReliefF นั้นใช้ค่าน้ำหนักที่ถูกเปรียบเทียบระหว่างค่าความคล้ายระหว่างชุดตัวอย่างในการศึกษาแบบ Case-Control อย่างไรก็ตามเมื่อเพิ่มจำนวนของสนิปที่พิจารณาให้มากขึ้น ค่าความแม่นยำที่ได้นั้นได้ลดลงตาม จึงได้เสนออัลกอริทึม Very Large Scale ReliefF (VLSReliefF) โดยการประยุกต์ใช้ ReleifF โดยสุ่มเลือกสนิปในชุดย่อยของแต่ละชุดประชากรตัวอย่างด้วยค่าความคาดหวัง (Expectation Value) อย่างน้อย 1 ชุดย่อย แล้วนำมาผสมกับผลลัพธ์ส่วนอื่น ๆ ในกลุ่มประชากรหลาย ๆ ชุดย่อย เพื่อทำการปรับปรุงค่าน้ำหนักของแต่ละตัวแปร ด้วยค่าน้ำหนักที่สูงสุดในแต่ละชุดย่อย ผลการทดลองได้แสดงให้เห็นว่า ReliefF มีความแม่นยำสูงเมื่อนำมาใช้กับข้อมูลที่มีจำนวนสนิปขนาดเล็ก แต่เมื่อนำมาใช้กับข้อมูลที่มีสนิปขนาดใหญ่แล้ว ประสิทธิภาพนั้นไม่ได้ดีกว่าการสุ่ม ส่วนการใช้ VLSReliefF นั้นประสบความสำเร็จเมื่อนำไปใช้กับข้อมูลขนาดใหญ่ที่มีขนาด 100,000 สนิป และมีประชากรตัวอย่างจำนวน 1600 ตัวอย่าง

ในปี 2009 Maldonado และ Weber ได้นำเสนอการใช้ Wrapper Method เพื่อคัดเลือกตัวแปรร่วมกับแบบจำลอง SVM โดยใช้เทคนิค Sequential Backward มาช่วยในการเลือก จากผลการทดลองสรุปได้ว่า วิธีที่นำเสนอมีประสิทธิภาพดีกว่าวิธี Filter และ Wrapper ชนิดอื่น ๆ อย่างไรก็ตาม ผลการเลือกตัวแปรจากวิธีที่นำเสนอนี้ อาจจะได้ตัวแปรที่ต่างกันในแต่ละครั้ง เนื่องจากชุดข้อมูลที่ใช้ในการพิจารณานั้นเกิดจากการสุ่มในแต่ละรอบ ดังนั้นการที่จะหลีกเลี่ยงปัญหานี้

ควรจะทำซ้ำ 3-4 รอบเพื่อเปรียบเทียบตัวแปร

จากงานวิจัยของ Wei และคณะในปี 2010 ได้มีการศึกษาเกี่ยวกับความสัมพันธ์ระหว่างโรคและการเปลี่ยนแปลงของ DNA ที่มีความสัมพันธ์ด้วยการใช้สนิป ซึ่งมีประมาณ 12 ล้านตำแหน่งที่มีการเปลี่ยนแปลงที่เสถียร[1] ในการทดลองได้ทดสอบบนข้อมูลโรค Crohn ซึ่งมีอัตราการเกิดโรค 150-200 คน ในประชากรหนึ่งแสนคนในกลุ่มประชากรประเทศตะวันตก โดยพยายามที่หาความสัมพันธ์ของสนิปกับโรคและทำการแก้ปัญหาการคัดเลือกตัวแปรด้วยการทำสับเซตในการสร้างแบบจำลอง โดยใช้เทคนิคที่พัฒนาขึ้น ที่เรียกว่า USVM ซึ่งมีพื้นฐานมาจาก Univariate Marginal Distribution และ SVM [1] จากการทดลองแสดงให้เห็นว่า ผลลัพธ์ที่ได้จาก USVM นั้นดีกว่าเทคนิคอื่น เช่น IG, ReliefF, Sequential Forward Selection อย่างชัดเจน

Goldstein และคณะ ได้นำเสนอการใช้ Random Forest (RF) กับข้อมูล GWAS ของโรค Multiple Sclerosis ที่มีขนาดของสนิปมากกว่า 300,000 ตำแหน่ง ในปี 2010 [3] โดยที่ในงานวิจัยได้ทำการปรับหาค่าพารามิเตอร์ที่เหมาะสมที่สุดดังนี้ (1) การหาค่าของตัวแปรที่จะค้นหาในแต่ละโหนด ในการค้นหาตัวแปรเพียงไม่กี่ตัวต่อโหนดนั้น จะส่งผลให้ความสัมพันธ์ของต้นไม้ตัดสินใจนั้นก็จะมีน้อย ทำให้ความแปรปรวนของการทำนายลดน้อยลง อีกทั้งค่าความแม่นยำของแต่ละต้นไม้ตัดสินใจก็จะลดลงด้วย ซึ่งค่านี้คือการกำหนดภาพรวมของความซับซ้อนของแบบจำลอง ยิ่งมีค่าน้อยก็ยิ่งทำให้แบบจำลองมีความซับซ้อนมากขึ้น (2) จำนวนของต้นไม้ตัดสินใจ โดยปัจจัยในการเลือกนั้นขึ้นอยู่กับชุดข้อมูล ถ้าชุดสอนนั้นเป็นข้อมูลที่ดีก็จะทำให้การสร้างแบบจำลองนั้นเร็วและมีความจำเป็นที่จะต้องสร้างต้นไม้จำนวนน้อย เมื่อจำนวนต้นไม้มีขนาดใหญ่จะส่งผลให้แบบจำลองต้องการความสามารถในการคำนวณมากขึ้น และ (3) ค่าน้ำหนัก ซึ่งเมื่อจำนวนของตัวอย่างในแต่ละประเภทมีความไม่สมดุลกันนั้น จะมีการโน้มเอียง หรือ ใช้ค่าไบอัส (Bias) กับคลาสที่เป็นส่วนมาก ดังนั้นจึงจำเป็นที่จะต้องทำให้ค่าน้ำหนักแต่ละประเภทมีความสมดุลกัน งานวิจัยนี้เป็นงานวิจัยแรกที่นำ RF มาใช้กับปัญหา GWAS และได้แสดงให้เห็นว่า RF นั้นมีความต้องการใช้ข้อมูลที่น้อยกว่า เป็นแนวทางที่ยืดหยุ่นกว่าในการวิเคราะห์ข้อมูล

ในปี 2012 Wu และคณะ ได้ใช้เทคนิค Stratified Sampling Random Forests เพื่อคัดเลือกสนิปเพื่อจำแนกโรคพาร์กินสันและโรคอัลไซเมอร์ [21] เนื่องจากการคัดเลือกตัวแปรด้วย RF นั้น มีพารามิเตอร์ที่ต้องปรับหลายค่า ทำให้ขั้นตอนในการปรับค่าพารามิเตอร์นั้นต้องใช้เวลามาก ดังนั้น คณะวิจัยจึงใช้ Stratified Sampling เข้ามาใช้สำหรับสร้างต้นไม้ตัดสินใจใน RF โดยการสุ่มเลือกสนิปในแต่ละกลุ่มและนำมารวมกันเพื่อสร้างสับเสปซของต้นไม้ตัดสินใจ ข้อดีของกระบวนการนี้คือ สามารถสร้างสับเสปซที่มีสนิปที่มีความเกี่ยวข้องได้ จากการทดลองแสดงให้เห็นว่าวิธีการนี้นั้นมีประสิทธิภาพ สามารถลด Generalization Error, ความผิดพลาดในชุดทดสอบได้, และลดเวลาที่ใช้ในการคำนวณอย่างเห็นได้ชัด

ในปี ค.ศ. 2012 Briones และ Dinu [22] ได้นำเสนอ การค้นหาความเกี่ยวข้องทางพันธุกรรมเป็นปัจจัยที่มีความสำคัญในการศึกษาความเข้าใจความเจ็บป่วยของมนุษย์ที่ได้รับการสืบทอดมาจากบรรพบุรุษ การหาความเปลี่ยนแปลงทางพันธุกรรมแบบที่มีปฏิสัมพันธ์หลากหลายกับโรค เป็นความท้าทายที่ควรศึกษาเกี่ยวกับสาเหตุของการเกิดโรคแบบ

[1]ตำแหน่งของสนิปที่มีการเปลี่ยนแปลงสูงและมีเสถียรภาพนั้นเหมาะที่จะใช้เป็น Biomarker อย่างมากในการศึกษาความสัมพันธ์กับโรค

ซับซ้อน ว่า SNP นั้นมีความเกี่ยวโยงกับการตอบสนองของภูมิคุ้มกัน คลอเรสโตรอล ไขมัน และการเผาผลาญ ซึ่งมีเกี่ยวข้องกับโรคอัลไซเมอร์ ซึ่งอัตราส่วนของการสืบทอดของโรคอัลไซเมอร์นั้นยังไม่สามารถอธิบายได้ ในงานวิจัยนี้จึงพยายามหาตัวแปรทางพันธุศาสตร์ที่อาจมีอิทธิพลกับความเสี่ยงของโรคด้วยกระบวนการเหมืองข้อมูล ทำการทดลองสองแนวทางเพื่อทำการคัดเลือกตัวแปร โดยใช้การทำ filter method ด้วยแบบจำลอง logistic regression เพื่อหาตัวแปรที่มีค่า p-value ที่มากกว่าค่า Bonferroni threshold นำผลลัพธ์ที่ได้มาสร้างแบบจำลองใหม่ด้วย Random forest เพื่อหาผลลัพธ์การจำแนกผลลัพธ์ กับอีกวิธีหนึ่งคือ ใช้ความรู้ทางด้านชีววิทยาเพื่อเลือกตัวแปรที่มีความสัมพันธ์ การทดลองทำการ stratification sampling และ LD (Linkage Disequilibrium) ในการคัดเลือก SNP ได้จำนวน 19 SNP ใช้ logistic regression วิเคราะห์ผลของการเลือกและนำ SNP ไปสร้างแบบจำลองเช่นเดียวกับวิธีการแรก ได้ผลลัพธ์ดังต่อไปนี้ โมเดลที่หนึ่งใช้ SNP จำนวน 200 SNP และมีความแม่นยำที่มีความผิดพลาดที่น้อยกว่าแบบที่สอง แต่อย่างไรก็ตามในแบบที่สองนั้นนั่นมีความง่ายกว่าในการทดสอบ ซึ่งตรงกันข้ามกับในโมเดลแบบที่ 1 ซึ่งมี SNP จำนวน 200 SNP ซึ่งมีความยากในการทดสอบ แต่ว่ามีตัวแปรทางพันธุศาสตร์ที่มีความเกี่ยวพันธ์กันมากกว่า ซึ่งสามารถนำมาพิจารณาในกรณีที่เป็นโรคแบบซับซ้อนได้

ในปี 2015 Hira และ Gillies ได้พยายามที่จะแก้ปัญหาของมิติข้อมูล (Curse of Dimensionality) ที่นำมาสู่ปัญหาการเข้ากันกับข้อมูลมากเกินไป ซึ่งทำให้เกิดความผันผวนในการสร้างแบบจำลอง ซึ่งหนึ่งในสาเหตุของปัญหานี้คือข้อมูลรบกวน (Noise Data) ที่ส่งผลกระทบต่อประสิทธิภาพโดยตรง ดังนั้นเราควรจะกรองข้อมูลรบกวนออกให้มากที่สุด นั่นคือการลดจำนวนตัวแปรที่ไม่เกี่ยวข้องกับปัญหาที่กำลังพิจารณา ซึ่งสามารถทำได้โดยการเลือกกลุ่มย่อยของคุณลักษณะ [19] ในงานวิจัดนี้ได้ศึกษาวิธีต่าง ๆ โดยแบ่งตามชนิดของวิธีนั่นคือ Filter Method, Wrapper Method, และ Embedded Method โดยชนิด Filter Method นั้นได้แบ่งออกเป็น 2 ประเภท นั่นคือ (1) Univariate Method ที่มีการพิจารณาคุณลักษณะเป็นรายตัวแยกกัน เช่น Unconditional Mixture Modelling, Information Gain, Markov Blanket ขณะที่ (2) Multivariate Method นั้นพิจารณาเป็นกลุ่ม เช่น Error-weighted Uncorrelated Shrunken Centroid, Correlation-based Feature Selection, ReliefF เป็นต้น สำหรับอัลกอริทึม ReliefF นั้น เป็นอัลกอริทึมที่นิยมกันเป็นอย่างมาก โดยเฉพาะข้อมูล Microarray สำหรับ Wrapper Method เป็นวิธีที่ต้องเลือกตัวแปรก่อน จากนั้นนำตัวแปรเหล่านั้นมาสร้างเป็นแบบจำลองและทดสอบ ซึ่งจะต้องใช้การคำนวณที่สูงมาก ซึ่ง Wrapper นั้น สามารถแบ่งออกได้เป็น 2 วิธี คือ (1) Deterministic Wrapper โดยใช้ Hill-climbing Search ในการคัดเลือก และ (2) Randomised Wrapper ซึ่งเกิดจากการสุ่ม เช่น GA ซึ่งใช้เวลาในการคำรนวณและหน่วยความจำที่สูง สำหรับวิธีประเภทสุดท้าย Embedded Method เป็นวิธีที่ใช้การคำนวณที่น้อยกว่าชนิด Wrapper โดยที่แบบจำลองจะทำการเลือกตัวแปรให้โดยอัตโนมัติ เช่น RF และ SVM

ในปี ค.ศ. 2015 งานวิจัยของ Nguyen และคณะ [7] ได้เสนอการคัดเลือกสนิปโดยใช้ RF แบบ Two-stage มาแก้ปัญหาสนิปที่มีขนาดใหญ่ ซึ่ง RF เป็นหนึ่งในอัลกอริทึมที่สามารถจัดการปัญหานี้ได้ดีที่สุดในการทดลอง GWAS ในการใช้ค่าความสำคัญของตัวแปรที่ได้จากโมเดล RF นั้นอาจจะมีการรบกวนในข้อมูลเกิดขึ้นเรียกว่า Shadows ซึ่ง Shadows สนิปนั้นมีผลทำให้การสร้างแบบจำลองเพื่อจำแนกข้อมูลไม่มีประสิทธิภาพ จึงได้ใช้การทดสอบ Wilconxon Rank-sum เพื่อสร้างค่าขีดแบ่ง สำหรับสนิปที่มีค่า *p*-value จาก Wilconxon Test ที่มีค่ามากกว่าค่าขีดแบ่ง จะถูกนำออก

จากชุดข้อมูล เนื่องจากเป็นสนิปที่ไม่มีความสัมพันธ์กับคลาส ผลการทดลองพบว่าวิธีนี้มีประสิทธิภาพในการเลือกกลุ่มย่อยของสนิป และสามารถหาสนิป ที่มีความสัมพันธ์กับโรคได้ซึ่งวิธีเดิมที่ใช้กันไม่สามารถทำได้ ซึ่งแก้ปัญหาจำนวนของกลุ่มประชากรมีน้อยกว่าจำนวนสนิปมาก ๆ ได้ ซึ่งเป็นปัญหาหลักของการวิจัย GWAS

บทที่ 3
วิธีดำเนินการวิจัย

3.1 ชุดข้อมูลที่ใช้ในการทำวิจัย

ข้อมูลสนิปที่ใช้เป็นกลุ่มตัวอย่างประชากรตัวอย่างที่เป็นโรค β-thalassemia จำนวน 618 คน ซึ่งแบ่งออกเป็น Case (Severe) จำนวน 383 คน และ Control (Mild) จำนวน 235 คน ตามระดับความรุนแรงของอาการ ซึ่งทั้งหมดเป็นเป็นประชากรของประเทศไทย ข้อมูลชุดนี้มีสนิปจำนวน 564,831 สนิป จากนั้นได้มีการกรองสนิปที่มีค่าสูญหายที่มากกว่าร้อยละ 1 ออกไปจากชุดข้อมูล ส่วนข้อมูลที่มีค่าสูญหายแต่ว่าน้อยกว่าร้อยละ 1 นั้น ได้ทำการแทนค่าสูญหายด้วยฐานนิยมของข้อมูลในสนิปนั้น ๆ เมื่อทำการแทนค่าสูญหายเสร็จสิ้นแล้ว จะได้ชุดข้อมูลตัวอย่างกลุ่มประชากรที่มีจำนวน 618 คนเท่ากับชุดข้อมูลเริ่มต้น แต่สนิปที่จะนำมาทดลองเหลือเพียงแค่จำนวน 451,856 สนิป

3.2 การออกแบบการทดลอง

การทดลองได้แบ่งออกเป็นสองขั้นตอนหลัก คือ (1) การคัดเลือกและจัดลำดับสนิปที่มีความสัมพันธ์กับระดับของความรุนแรงของโรค ซึ่งสนิปจะถูกจัดอันดับด้วยการเรียงตามค่าความสำคัญจากมากไปน้อย (2) การคัดเลือกสนิปเพื่อสร้างแบบจำลองทำนายความรุนแรงของโรค โดยที่สนิปนั้นถูกกรองมาจากขั้นตอนแรก

3.2.1 การจัดลำดับและการกรองสนิปที่มีความเกี่ยวข้องกับความรุนแรงของโรค

จำนวนของสนิปในข้อมูลชุดตัวอย่างมีขนาดใหญ่มาก ซึ่งขนาดที่ใหญ่กว่าจำนวนกลุ่มตัวอย่างมาก ๆ นั้น เมื่อนำมาสร้างแบบจำลอง จะมีโอกาสสูงที่จะเกิดปัญหาการเข้ากันของข้อมูลมากเกินไป ดังนั้น การคัดเลือกสนิปที่มีความเกี่ยวข้องกับโรค จำนวนน้อย อาจจะทำให้แบบจำลองมีประสิทธิภาพมากขึ้น ในการศึกษานี้ได้ใช้การคัดเลือกตัวแปรชนิด Filter Method และ Embedded Method เท่านั้น ไม่ได้พิจารณา Wrapper Method เนื่องจากจุดมุ่งหมายหลักนั้น คือ การคัดเลือกสนิปที่มีความเกี่ยวข้องกับความรุนแรงของโรคจากชุดข้อมูลขนาดใหญ่อย่างรวดเร็ว ซึ่งเทคนิคที่ถูกนำมาใช้ใน Filter Method คือ χ^2 และ IG ส่วน Embedded Method คือ GB ซึ่งเทคนิคเหล่านี้สามารถสร้างค่าความสำคัญของแต่ละตัวแปรได้ จำนวนสนิปที่ถูกจัดลำดับในแต่ละวิธีจะถูกคัดเลือกออกมา 250 สนิปตามลำดับค่าความสำคัญของแต่ละวิธี และจะถูกนำไปใช้ในขั้นตอนถัดไป

3.2.2 การคัดเลือกสนิปเพื่อสร้างแบบจำลองการทำนายความรุนแรงของโรค

นำสนิปที่มีค่าความสำคัญสูงสุดของแต่ละวิธีจำนวน 250 สนิปที่ถูกคัดเลือกในขั้นตอนที่ผ่านมา เพื่อมาสร้างแบบจำลองทำนายระดับความรุนแรงของโรค และหาแบบจำลองที่มีผลลัพธ์ดีที่สุด ด้วยการทดลองสร้างแบบจำลองโดยเลือกสนิปเพิ่มครั้งละ 1 ตัว เริ่มต้นจากใช้ 2 สนิป ที่มีค่าความสำคัญสูงที่สุดมาสร้างแบบจำลอง และเพิ่มสนิปเข้ามา

จนครบ 250 สนิปเรียงตามลำดับ โดยใช้อัลกอริทึมต่อไปนี้มาสร้างแบบจำลอง คือ SVM, NB และ GB โดยแต่ละแบบจำลองถูกนำมาหาค่าพารามิเตอร์ที่ดีที่สุด ด้วยการทำการตรวจสอบไขว้ m ชุด (m-fold Cross Validation) และ Grid Search บนชุดสอน โดยที่ $m = 4$ ในการทดลองนี้ สำหรับการสร้างแบบจำลองด้วย SVM จะใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้น ซึ่งมีการปรับค่าพารามิเตอร์ C เพื่อปรับขอบ (Margin) ของระนาบแบ่ง โดยมีการปรับค่า C ดังต่อไปนี้ $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}]$ ส่วนการสร้างแบบจำลองด้วย NB และ GB ไม่ได้มีการปรับค่าพารามิเตอร์ของแบบจำลอง การสร้างแบบจำลอง NB และ SVM นั้นสร้างด้วยภาษา Python และไลบรารี่ scikit-learn ส่วนแบบจำลอง GB ถูกสร้างด้วยไลบรารี่ XGBoost หลังจากที่ได้ค่าพารามิเตอร์ที่ดีที่สุด เราจึงสร้างแบบจำลองที่ดีที่สุดจากข้อมูลสอนทั้งหมด โดยใช้ค่าพารามิเตอร์นั้น ๆ และทดสอบด้วยชุดทดสอบ

การสร้างแบบจำลองเพื่อจำแนกระดับความรุนแรงของโรคด้วยการใช้เทคนิค SVM, NB และ GB ใช้ตัวแปรต้น คือ ลำดับของ สนิปที่มีลำดับความสำคัญจากขั้นตอนแรก ด้วยการใช้ χ^2 , IG และ GB กรองลำดับความสำคัญของสนิป และนำมาสร้างแบบจำลองด้วยเทคนิคดังต่อไปนี้ : (1) χ^2+GB, (2) χ^2+NB, (3) χ^2+SVM, (4) IG+GB, (5) IG+NB, (6) IG+SVM, (7) GB+SVM, (8) GB+NB ประสิทธิภาพของแบบจำลองจะถูกนำมาประเมินผล กับแบบจำลองที่ไม่ได้ทำการคัดเลือกสนิป ซึ่งจะถูกนำมาใช้เปรียบเทียบเป็นการทดลองฐาน

เทคนิคต่าง ๆ ที่ได้กล่าวไปนั้น ได้ถูกประเมินผลเหมือนกัน โดยใช้เทคนิคการตรวจสอบไขว้ 5 ชุด โดยชุดข้อมูลกลุ่มตัวอย่างสนิปจะถูกแบ่งออกเป็น 5 ชุดในขนาดที่ใกล้เคียงกัน ในแต่ละเซตมีจำนวนกลุ่มตัวอย่างประมาณ 123–124 ตัวอย่าง โดยใช้ข้อมูล 4 ชุดสำหรับสร้างแบบจำลอง และอีก 1 ชุดสำหรับใช้ทดสอบ ทดลองจำนวน 5 รอบสลับกันไปในแต่ละรอบ จากนั้นวัดประสิทธิภาพของแบบจำลองด้วย ค่าความแม่นยำและ F1-score ของแต่ละชุด และรายงานค่าเฉลี่ยทั้งหมด

บทที่ 4
ผลการวิจัย

จากการทดลองได้ทำการเก็บผลลัพธ์ และรายงานค่าความแม่นยำ, F1-score, Precision, และ Recall จากการทดสอบเทคนิคแบบต่าง ๆ ตามตารางที่ 4.1 ที่แสดงประสิทธิภาพของแต่ละวิธีจากข้อมูลชุดสอนและชุดทดสอบ จำนวนของสนิปที่ใช้ในแต่ละวิธีที่ทำให้ได้ผลลัพธ์ที่ดีที่สุด จากการทดลองพบว่า เมื่อนำข้อมูลสนิปทั้งหมดมาใช้ในการสร้างแบบจำลอง แล้วนำแบบจำลองที่ได้มาทดสอบด้วยข้อมูลชุดเดียวกัน ค่า F1-score และความแม่นยำจะมีค่าเท่ากับ 1 ซึ่งหมายความว่าแบบจำลองสามารถทำนายได้ถูกต้องทั้งหมด อย่างไรก็ตาม เมื่อแบบจำลองใช้ข้อมูลชุดทดสอบเพื่อทำนายผลลัพธ์ พบว่าประสิทธิภาพของแบบจำลองนั้นมีค่าที่ต่ำมาก ค่า F1-score มีค่าที่เข้าใกล้ 0 ซึ่งแสดงให้เห็นว่าแบบจำลองนั้นเกิดปัญหาการเข้ากับกับข้อมูลมากเกินไปในทุก ๆ แบบจำลองที่ใช้สนิปทั้งหมด ยกเว้นแบบจำลองที่ใช้เทคนิค GB ซึ่งให้ค่า F1-score ที่ดีที่สุดที่ 0.3637 เนื่องจาก GB นั้นสามารถให้ค่าความสำคัญของตัวแปรที่ใช้ในแบบจำลอง แต่ก็ยังมีปัญหาการเข้ากับกับข้อมูลมากเกินไปเช่นเดียวกันแบบจำลองอื่น ๆ แบบจำลองที่มีประสิทธิภาพที่ดีที่สุดคือ χ^2+SVM ซึ่งมีค่า F1-score เท่ากับ 0.4873 โดยการใช้สนิปเพียงแค่ 10 สนิป ตามมาด้วย χ^2+GB และ GB+SVM ซึ่งมีค่า F1-score 0.4797 และ 0.4220 และใช้จำนวนสนิปจำนวน 23 และ 10 สนิป ตามลำดับ นอกจากนี้ χ^2+SVM นั้นให้ค่า Recall ที่ดีที่สุดที่ 0.4992 ซึ่งเป็นวิธีที่มีค่าประสิทธิภาพที่ดีกว่าการทดลองฐาน และ GB ที่ใช้ข้อมูลสนิปทั้งหมด

ในกรณีของแบบจำลองที่ทำนายผลลัพธ์เพียงค่าเดียวเป็น Severe ค่าประสิทธิภาพของแบบจำลองจะมีค่าความแม่นยำและ F1-score เท่ากับ 0.6190 และ 0 ตามลำดับ ดังตารางที่ 4.1 มี 6 วิธีที่แบบจำลองให้ผลลัพธ์ดีกว่าการทดลองฐานแบบสุ่มดังนี้ χ^2+NB, GB, χ^2+GB, NB, χ^2+SVM และ IG+NB เรียงตามลำดับ χ^2+NB ไม่เป็นเพียงวิธีการที่ได้ค่าความแม่นยำที่ดีที่สุดแต่ยังเป็นวิธีเดียวที่มีค่าความแม่นยำและ F1-score ใกล้เคียงกันทั้งในข้อมูลชุดสอนและข้อมูลชุดทดสอบ และใช้สนิปเพียง 5 สนิป ซึ่ง χ^2+NB มีค่า Precision ดีเป็นอันดับที่ 2 เท่ากับ 0.6502 แต่ว่ามีค่า Recall ที่เพียงแค่ 0.2313

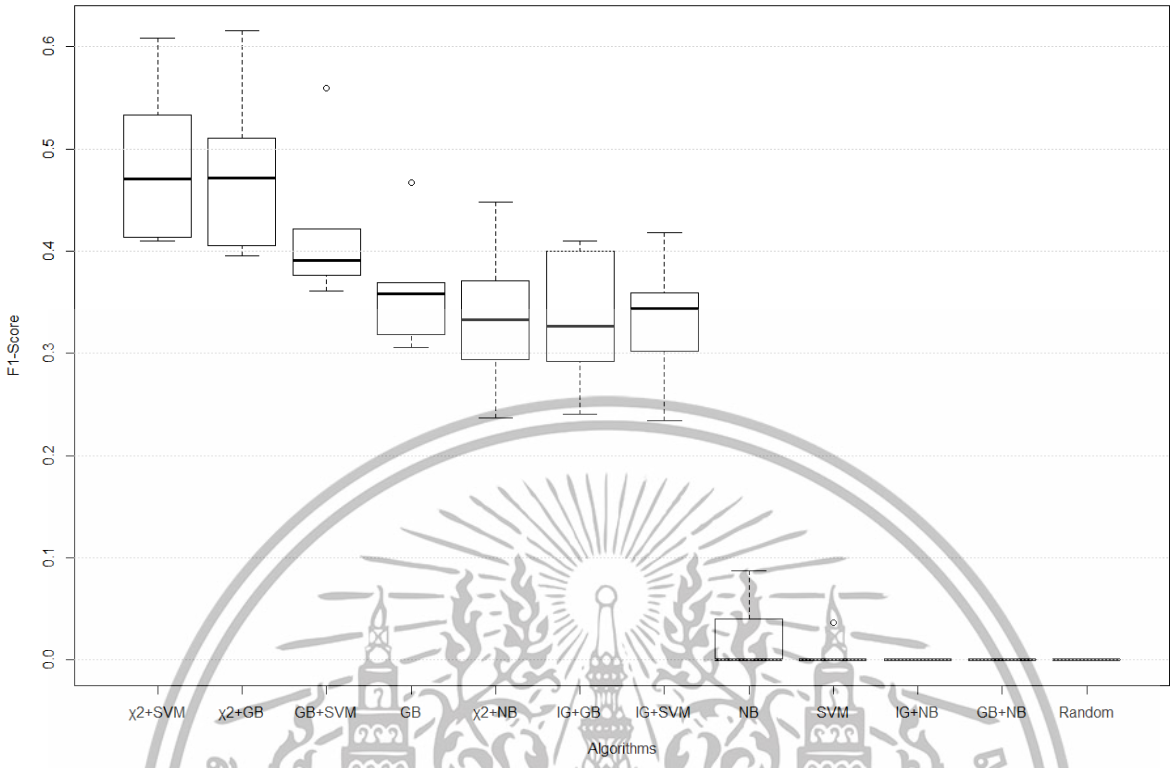
รูปที่ 4.1 แสดงค่า F1-score ของข้อมูลชุดทดสอบจากแบบจำลองที่เหมาะสมที่สุด โดยเรียงลำดับจากค่าเฉลี่ยจากการทดลอง 5 ครั้ง ในกรณีของ F1-score ผลลัพธ์ที่ได้จากการทดสอบเรียงลำดับดังต่อไปนี้ χ^2+SVM, χ^2+GB, GB+SVM, GB, χ^2+NB, IG+GB, IG+SVM, NB, SVM, IG+NB, GB+NB และ การทดลองแบบสุ่ม สำหรับค่าความแม่นยำเรียงลำดับดังต่อไปนี้ χ^2+NB, GB, χ^2+GB, NB, χ^2+SVM, IG+NB, การทดลองแบบสุ่ม, SVM, GB+NB, GB+SVM, IG+SVM และ IG+GB ดังที่แสดงในรูป 4.2 จะเห็นได้ว่าวิธีที่สุดในการคัดเลือกสนิป คือ χ^2 เนื่องจากแบบจำลองที่ถูกสร้างขึ้นมาที่มีประสิทธิภาพสูงสุดใน 5 อันดับแรกอย่างสม่ำเสมอ ทั้งในการวัดผลแบบค่าความแม่นยำและ F1-score

นอกจากนี้ตารางที่ 4.2 ซึ่งแสดงค่าเฉลี่ยของ F1-score ในทุก ๆ วิธี จะพบว่าค่า F1-score ของ χ^2 มีค่าเท่ากับ 0.4346 และ χ^2 สามารถใช้ในการคัดเลือก SNP ได้ดีกว่า ทั้ง IG และ GB เมื่อวัดจากค่า F1-score ซึ่งมีค่าเท่ากับ 0.2218 และ 0.2110 ตามลำดับ

ตารางที่ 4.1: แสดงผลเปรียบเทียบระหว่างแบบจำลองต่าง ๆ เมื่อทดสอบด้วยข้อมูลชุดสอนและข้อมูลชุดทดสอบ ในแบบจำลองที่มีการใช้ค่าพารามิเตอร์ที่เหมาะสมที่สุด และจำนวนตัวแปรที่ทำให้ค่า F1-score สูงที่สุด

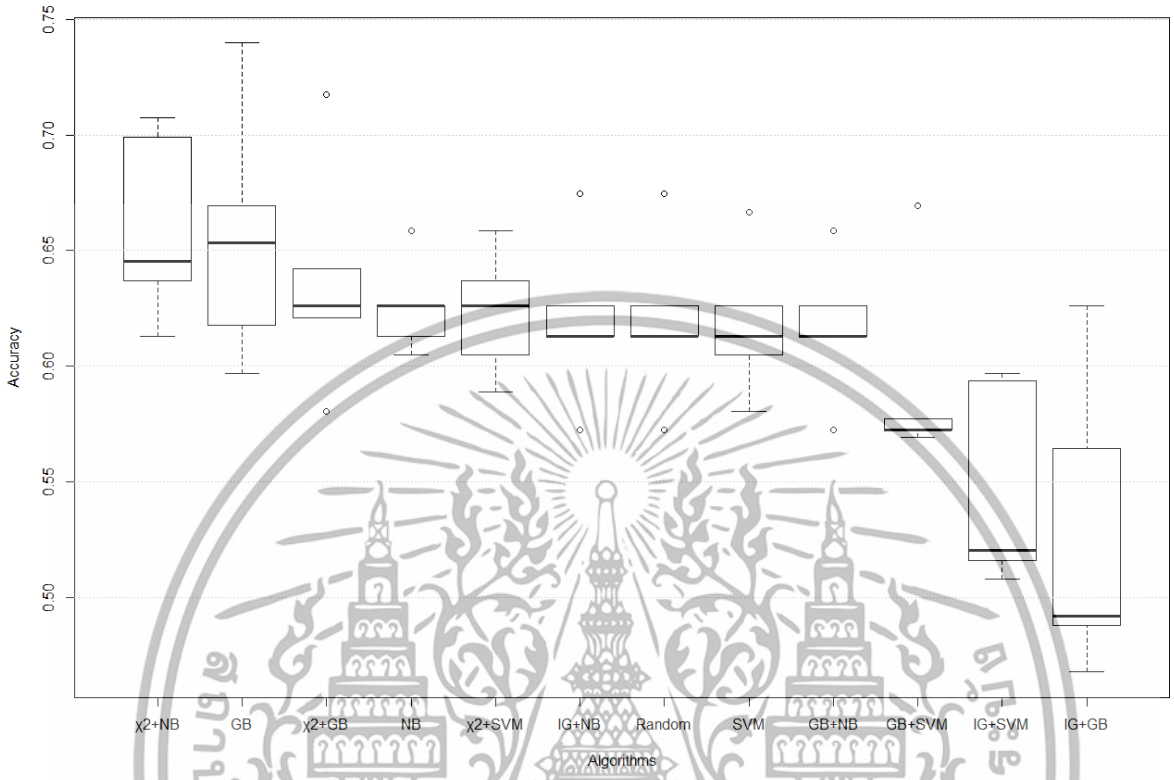
Algorithm	#Feature	Training Set				Test Set			
		F1-score	Accuracy	Precison	Recall	F1-score	Accuracy	Precison	Recall
Random	All	0	0.6190	0	0	0	0.6190	0	0
GB	All	1	1	1	1	0.3637	0.6554	**0.6738**	0.2900
NB	All	1	1	1	1	0.0254	0.6256	0.2000	0.0038
SVM	All	1	1	1	1	0.0074	0.6182	0.3667	0.0179
χ^2+GB	10	0.7047	0.8021	0.7918	0.6529	0.4797	0.6375	0.5182	0.4535
χ^2+NB	5	0.3329	0.6614	0.6648	0.2224	0.3368	**0.6603**	0.6502	0.2313
χ^2+SVM	10	0.7066	0.7989	0.8025	0.6357	**0.4873**	0.6230	0.5398	**0.4992**
IG+GB	9	0.8578	0.8981	0.8931	0.8075	0.3340	0.5276	0.363	0.3122
IG+NB	245	0.0325	0.6235	0.2000	0.0085	0	0.6198	0	0
IG+SVM	17	0.9735	0.9871	0.9931	0.9457	0.3316	0.5470	0.3804	0.2979
GB+NB	2	0.0325	0.6235	0.1389	0.0128	0	0.6166	0	0
GB+SVM	23	0.9973	0.9980	0.9921	0.9960	0.4220	0.5922	0.39662	0.3622

เมื่อเปรียบเทียบผลลัพธ์ของ F1-score และค่าความแม่นยำที่ดีที่สุด χ^2+SVM, χ^2+NB และ GB ด้วย Confusion Metric ที่แสดงในรูปที่ 4.3 พบว่า χ^2+NB และ GB ให้ผลลัพธ์การทำนายมีค่าเป็น 1 คือ severe เป็นจำนวนมาก เนื่องจากในกลุ่มชุดข้อมูลมีจำนวนผลลัพธ์ชนิดนี้มากกว่า ซึ่งการทดลองนี้เป็นแบบ case-control และมีจำนวนที่เป็น case มากกว่า ดังนั้น ถ้าแบบจำลองมีการทำนายผล 1 ทั้งหมดจะทำให้มีค่าความแม่นยำที่ดีโดยไม่ต้องใช้ชุดข้อมูลใดๆที่มีผลลัพธ์เป็น control เลย เนื่องจากจำนวนของ false-positive ที่ถูกทำนายออกมามีค่าสูงมากกว่าจำนวนของ true-negative และ false-negative ต่ำกว่าแบบจำลองที่มีค่า F1-score สูงอย่าง χ^2+SVM ดังนั้นการประเมินผลลัพธ์ของประสิทธิภาพของแบบจำลองในชุดข้อมูลนี้ ไม่ควรที่จะใช้ค่าความแม่นยำเพียงอย่างเดียว เนื่องจากจำนวนของ case และ control ในข้อมูลกลุ่มตัวอย่างมีจำนวนที่ต่างกันมาก ซึ่งเรียกว่าข้อมูลแบบ imbalance dataset ในกรณีของข้อมูลชนิดนี้ถ้าแบบจำลองทำนาย ผลออกมาเป็นคลาสที่มีขนาดใหญ่ที่สุดเพียงอย่างเดียว จะทำให้ค่าความแม่นยำนั้นมีค่าที่สูง แต่แบบจำลองไม่สามารถให้ผลลัพธ์การทำนายที่ถูกต้องได้เมื่อ ผลลัพธ์ที่แท้จริงเป็นอีกค่าหนึ่ง ในกรณีนี้ค่า F1-score เหมาะสำหรับการวัดผลข้อมูลที่มีลักษณะเช่นนี้ ซึ่งดีกว่าการวัดด้วยค่าความแม่นยำเพียงอย่างเดียว ในการทดลองนี้ 2 method ที่มีค่าความแม่นยำสูงคือ χ^2 และ GB นั้นเมื่อวัดด้วยค่าของ F1-score จะอยู่ที่อันดับ 5 และ 4 ตามลำดับ แต่ในกรณีของ χ^2+SVM และ χ^2+GB มีประสิทธิภาพที่ใกล้เคียงกันทั้งในด้านของค่าความแม่นยำและ F1-score และสูงกว่าสองวิธีแรกที่ได้ค่าความแม่นยำสูงมาก เมื่อเปรียบเทียบระหว่างสองวิธีที่ดีที่สุดคือ χ^2+SVM และ χ^2+GB ซึ่งทั้ง



รูปที่ 4.1: F1-score ของข้อมูลชุดทดสอบ จากแบบจำลองที่เหมาะสมที่สุด โดยเรียงลำดับจากค่าเฉลี่ยมากไปน้อย

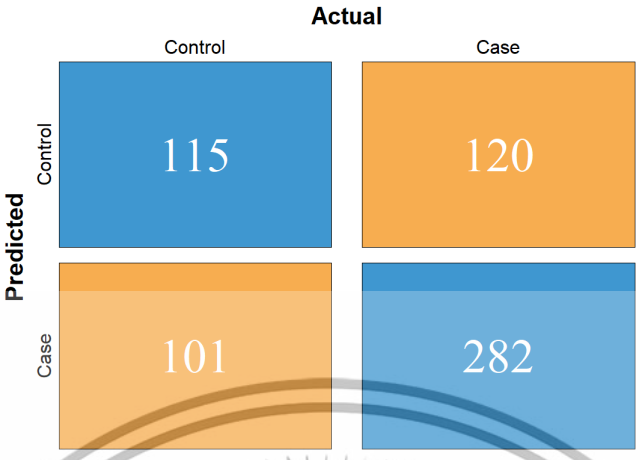
คู่ใช้จำนวนตัวแปร χ^2 จากกลุ่มตัวอย่าง χ^2+SVM นั้นมีประสิทธิภาพที่ดีกว่า χ^2+GB อย่างชัดเจน เนื่องจาก F1-score, precision และ recall ของ χ^2+SVM สูงกว่า χ^2+GB อย่างชัดเจน



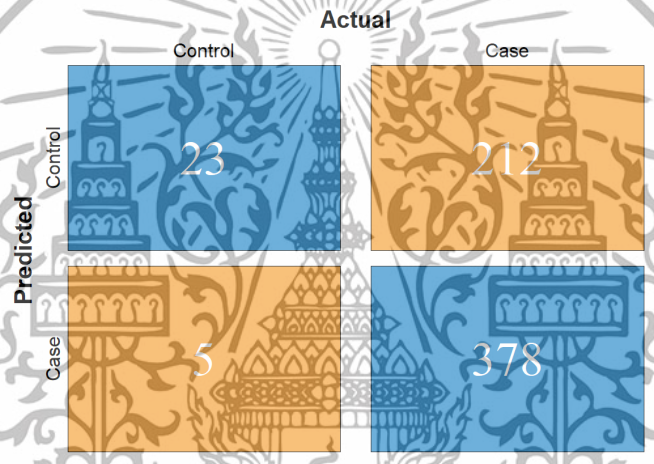
รูปที่ 4.2: ค่าความแม่นยำ ของข้อมูลชุดทดสอบ จากแบบจำลองที่เหมาะสมที่สุด โดยเรียงลำดับจากค่าเฉลี่ยมากไปน้อย

ตารางที่ 4.2: เปรียบเทียบค่าเฉลี่ย F1-score โดยเฉลี่ยทุก ๆ แบบจำลอง สำหรับแต่ละวิธีการกรองสนิป

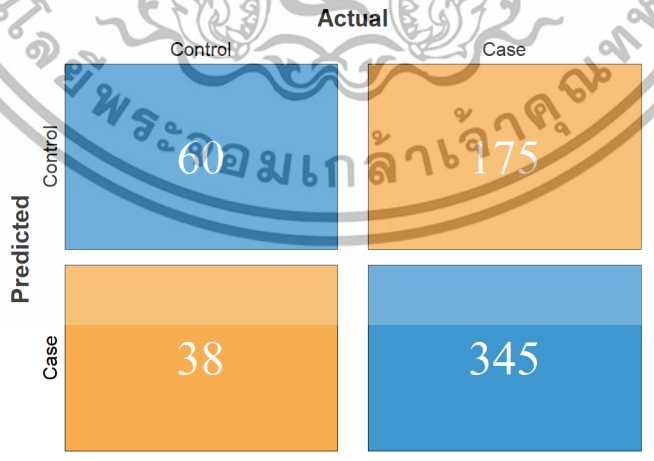
Feature Selection	Model			Average F1
	GB	NB	SVM	
χ^2	0.4797	0.3368	0.4873	**0.4346**
IG	0.3340	0	0.3316	0.2218
GB	-	0	0.4220	0.2110



(ก)



(ข)



(ค)

รูปที่ 4.3: Confusion Matrix เปรียบเทียบระหว่างวิธี (a) χ^2+SVM, (b) χ^2+NB, และ (c) GB

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยชิ้นนี้ได้นำเสนอการทดลองหาวิธีการกรองและการเลือกสนิปที่มีความเกี่ยวข้องกับความรุนแรงของโรค β-thalassaemia ที่เหมาะสมที่สุดจากข้อมูลสนิปที่มีขนาดใหญ่มาก ๆ และมีจำนวนตัวอย่างที่น้อย โดยพยายามคัดเลือกสนิปที่มีความจำเป็นมากที่สุด และมีประสิทธิภาพในการใช้สร้างแบบจำลองสูงสุด จากการทดลองได้แสดงให้เห็นว่าการใช้ χ^2 ซึ่งเป็นวิธีที่ได้รับความนิยมในการคัดเลือกตัวแปรที่เป็นลักษณะตัวแปรแบบกลุ่ม และเป็นวิธีพื้นฐานในการใช้เลือกสนิปที่มีประสิทธิภาพสูงสุด เมื่อเทียบกับวิธีพื้นฐานอย่าง IG เป็นต้น นอกจากนี้ การใช้ χ^2 ในชุดข้อมูลระดับความรุนแรงของโรคธาลัสซีเมียในกลุ่มประชากรไทย นั้นได้ผลลัพธ์ที่ดีกว่าวิธีที่ได้รับความนิยมนั่นคือ GB ซึ่งเป็นวิธีชนิด Embedded Method ที่มักจะได้ผลลัพธ์ที่ดีในการคัดเลือก SNP ที่ใช้จำแนกกลุ่มประชากร หลังจากที่กรองสนิปด้วย χ^2 แล้วนั้น จึงนำสนิปที่ถูกเลือกมาสร้างแบบจำลองด้วยอัลกอริทึมต่าง ๆ ซึ่งสามารถสรุปได้ว่า SVM และ GB นั้นให้ผลลัพธ์ประสิทธิภาพที่มีความใกล้เคียงกันมาก ในจำนวนสนิปที่มีจำนวนเท่ากัน ซึ่งใน Future Work จะนำสนิปที่ได้จากอัลกอริทึมต่าง ๆ มาวิเคราะห์และตรวจสอบกับงานวิจัยอื่น ๆ ว่ามีความสัมพันธ์กับโรคหรือไม่ต่อไป

บทที่ 6
สรุปผลผลิตงานวิจัย

ผลผลิตจากโครงการวิจัยนี้มีดังนี้

6.1 บทความวิจัย

1. Ek Thamwiwatthana, Kitsuchart Pasupa, and Sissades Tongsima. "Selection of SNP Subsets for Severity of Beta-thalassaemia Classification Problem." In Proceedings of the 9th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2018), 10-13 December 2018, Bangkok, Thailand, pp. 1-7, 2018. doi: 10.1145/3291757.3291770 (ภาคผนวก ก)

6.2 บทความวิชาการ

1. เอก ธรรมวิวัฒน และ กิตติ์สุชาต พสุภา, วิธีการคัดเลือกสนิปที่มีความสัมพันธ์กับคุณลักษณะพิเศษหรือโรค, วารสารเทคโนโลยีสารสนเทศลาดกระบัง (Submitted). (ภาคผนวก ข)

บรรณานุกรม

[1] B. Wei, Q. Peng, J. Li, X. Kang, and C. Li, "Usvm: Selection of snps in diseases association study using UMDA and SVM," in *Proceedings of the 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE'2010)*. Chengdu, China: IEEE, 2010, pp. 1–5.

[2] M. Zhang, Y. Lin, L. Wang, V. Pungpapong, J. C. Fleet, and D. Zhang, "Case-control genome-wide association study of rheumatoid arthritis from genetic analysis workshop 16 using penalized orthogonal-components regression-linear discriminant analysis," *BMC Proceedings*, vol. 3, no. 7, p. S17, 2009.

[3] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos, "An application of random forests to a genome-wide association dataset: methodological considerations & new findings," *BMC genetics*, vol. 11, no. 1, p. 49, 2010.

[4] C.-H. Yang, C.-H. Ho, and L.-Y. Chuang, "Improved tag snp selection using binary particle swarm optimization," in *Proceedings of the 2008 IEEE Congress on Evolutionary Computation (CEC'2008)*. Hong Kong, China: IEEE, 2008, pp. 854–860.

[5] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 3, pp. 427–444, 2005.

[6] J. Subramanian and R. Simon, "Overfitting in prediction models–is it a problem only in high dimensions?" *Contemporary clinical trials)*, vol. 36.2, pp. 636–641, 2013.

[7] T.-T. Nguyen, J. Z. Huang, Q. Wu, T. T. Nguyen, and M. J. Li, "Genome-wide association data classification and SNPs selection using two-stage quality-based random forests," *BMC Genomics*, vol. 16, no. 2, p. S5, 2015.

[8] J. Krawczuk and T. Lukaszuk, "The feature selection bias problem in relation to high-dimensional gene data," *Artificial intelligence in medicine*, vol. 66, pp. 63–71, 2016.

[9] M. B. Baldursdóttir, "Analysis of single nucleotide polymorphisms (snps) associated with classical hodgkin lymphoma in patients with infectious mononucleosis: Identification of a common genetic risk," Iceland, 2015.

[10] J. Li, D. Huang, M. Guo, X. Liu, C. Wang, Z. Teng, R. Zhang, Y. Jiang, H. Lv, and L. Wang, "A gene-based information gain method for detecting gene–gene interactions in case–control studies," *European Journal of Human Genetics*, vol. 23, no. 11, p. 1566, 2015.

[11] J. H. Moore and B. C. White, "Tuning relieff for genome-wide genetic analysis," in *Proceedings of the 5th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO'2007)*. Valencia, Spain: Springer, 2007, pp. 166–175.

[12] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Information Sciences*, vol. 179, no. 13, pp. 2208–2217, 2009.

[13] Y. Meng, Q. Yang, K. T. Cuenco, L. A. Cupples, A. L. DeStefano, and K. L. Lunetta, "Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and bayesian networks," *BMC proceedings*, vol. 1, no. 1, p. S56, 2007.

[14] S. C. Shah and A. Kusiak, "Data mining and genetic algorithm based gene/snp selection," *Artificial intelligence in medicine*, vol. 31, no. 3, pp. 183–196, 2004.

[15] T. T. Foundation. (n.d.) Diagnosis of thalassemia carrier. [Online]. Available: http://www.thalassemia.or.th/thaiversion/diag-carrier-th.htm

[16] W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang, "A support vector machine based naive bayes algorithm for spam filtering," in *Proceedings of the 35th International on Performance Computing and Communications Conference (IPCCC'2016)*. Las Vegas, NV, USA: IEEE, 2016, pp. 1–8.

[17] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles," in *Proceedings of the International Workshop on Data Mining for Biomedical Applications (BioDM'2006)*. Singapore: Springer, 2006, pp. 106–115.

[18] Á. Alonso Liso, "Feature selection with random forest and gradient boosting," Master's thesis, Department of Computer Engineering, Universidad Autónoma de Madrid, 2016.

[19] Z. M. Hira and D. F. Gillies, "Ga review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics*, 2015.

[20] M. J. Eppstein and P. Haake, "Very large scale relieff for genome-wide association analysis," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'08)*. IEEE, 2008, pp. 112–119.

[21] Q. Wu, Y. Ye, Y. Liu, and M. K. Ng, "Snp selection and classification of genome-wide SNP data using stratified sampling random forests," *IEEE Transactions on Nanobioscience*, vol. 11, no. 3, pp. 216–227, 2012.

[22] N. Briones and V. Dinu, “Data mining of high density genomic variant data for prediction of alzheimer’s disease risk,” *BMC medical genetics*, vol. 13, no. 1, pp. 370–375, 2012.

ภาคผนวก ก

บทความวิจัย

1. Ek Thamwiwatthana, **Kitsuchart Pasupa**, and Sissades Tongsima. "Selection of SNP Subsets for Severity of Beta-thalassaemia Classification Problem." In Proceedings of the 9th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2018), 10-13 December 2018, Bangkok, Thailand, pp. 1-7, 2018. doi: 10.1145/3291757.3291770

Selection of SNP Subsets for Severity of Beta-thalassaemia Classification Problem

Ek Thamwiwatthana
Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
nupippo@gmail.com

Kitsuchart Pasupa*
Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
kitsuchart@it.kmitl.ac.th

Sissades Tongsima
National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency
Pathum Thani, Thailand
sissades@biotec.co.th

ABSTRACT

Single-nucleotide polymorphisms (SNPs) are important genetic variables that are very popular in Genome-wide association study at the present time. They are often used in studies related to genetic disorders. A distinctive trait of SNPs is that there are a lot of them since they are variables originated from various positions in a DNA sequence. Unfortunately, the number of samples investigated are usually far fewer than the number of SNPs and so an over-fitting often occurs when one wants to construct a predictive model for classifying a sample into a case or a control. This study investigated a dataset on beta-thalassemia, a common genetic disorder widely found in Thai population. The data in the set are divided into two groups: severe and mild groups. The aims of the study were to develop and evaluate methods for screening and ranking SNPs related to this disorder. The screening methods tested were Chi-squared test (χ^2), Information Gain, and Gradient Boosting (GB). The SNPs that were screened in and selected were then used to construct a predictive model for classifying a sample to be either a severe or mild case. The model construction methods tested were Support Vector Machine (SVM), GB, and Naïve Bayes. Several combinations of a screening method and a model construction method were evaluated, and the evaluation results show that the best combination was χ^2-SVM which used the number of selected SNPs of 10.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; • **Computing methodologies** → *Feature selection*;

KEYWORDS

SNP Selection, Beta-thalassaemia

ACM Reference Format:
Ek Thamwiwatthana, Kitsuchart Pasupa, and Sissades Tongsima. 2018. Selection of SNP Subsets for Severity of Beta-thalassaemia Classification Problem. In *The 9th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2018), December 10–13, 2018, Bangkok, Thailand.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3291757.3291770

*Corresponding Author

CSBio 2018, December 10–13, 2018, Bangkok, Thailand

ACM ISBN 978-1-4503-6560-4/18/12…$15.00
https://doi.org/10.1145/3291757.3291770

1 INTRODUCTION

In the bioinformatics field, the most commonly used variables for genetic studies are single nucleotide polymorphisms (SNPs). A SNP is a variation of human DNA sequence that makes every human individual different from other individuals. If a SNP that is responsible for a certain disease or a response to a drug can be identified, it is possible to prevent that disease or adverse response. The basic goal is to be able to classify each unknown sample to be either a case or a control.

There are a very large number of SNPs but a much smaller number of sample groups. For example, in a study on a SNPs that are related to rheumatoid arthritis, the number of to-be-identified SNPs were 545,080 (868 Cases and 1,194 Controls) [23]. In a study on a SNPs that are related to multiple sclerosis, the number of investigated SNPs were 325,807 (931 Cases and 2,431 Controls) [5], while a study on a SNPs that are related to Parkinson's disease, the number of investigated SNPs were 408,803 (271 Cases and 270 Controls) [21]. For all of these studies, the problem of High Dimension, Low Sample Size (HDLSS) were encountered. This problem has been seen frequently in the fields of chemoinformatics and microarray analysis where only a few hundred samples are available [6, 14, 15]. In the field of computer science, a high chance of having an over-fitting problem occurs when a model is constructed from a group of data that is HDLSS. An over-fitting problem is a problem that a constructed model makes accurate predictions when operating on a training dataset but inaccurate predictions when operating on a test dataset. Most good predictive models have been constructed from a dataset that contained a smaller number of variables than the number of sample data in the set [18].

Therefore, for our purpose of identifying the best SNPs related to a disease, a reduction of the number of variables (SNPs) was most important not only in order for the model to make accurate predictions but also in order for it to give faster predictions (to use less computation time). Moreover, there are pieces of evidence that most intended-to-be-identified SNPs were not related to the investigated disease. For all of these reasons, screening for the most significant SNPs is the most important task for obtaining Genome-wide Association Study (GWAS) Data [12].

The reduction of the size of features or the number of variables can be done in two major ways: (1) feature selection which is a

selection of variables that are closely related to an issue among the population; good variables make direct analysis easy and the relationships between the manifestation of the issue and the variables can also be easily explained. and (2) Dimension Reduction which is a reduction of the number of variables by projecting all of the data onto another space with smaller dimensions but still retaining useful information; however, this method cannot be used to directly find the relationships between the manifestation of the issue and the variables. Therefore, to find the best SNPs that are closely related to a disease, we could use feature selection methods. Feature selection methods can be of three types: (1) Filter method which finds the relationship between each variable and the outcomes and uses the relationship to filter the variables; this method does not use much computational time; several filter methods are such as Chi-squared test [2], Information Gain [8], and ReliefF [11]; (2) Wrapper method which randomly selects several sets of variables and finds the best set of variables that are the most closely related to the outcomes; this method generally gives a better result than the filter method, but a model constructed by using the optimum set of variables found has a high chance of running into an over-fitting problem; moreover, it uses more computational time than any other methods; several of the best methods of this type are such as randomisation [9, 19], Genetic Algorithm [16], and Particle Swarm Optimisation [22]; and (3) Embedded method which is a model construction method that selects features automatically; the constructed model comes with the calculated weight for each variable that indicates the significance of that variable; this method uses less computational time than the wrapper method and also has less chance of getting an over-fitting problem. Most recent studies have used this method because it is more effective than the rest; methods of this type are such as Decision Tree [16] and Random Forest [10, 12].

The disease of interest in this study was beta-thalassemia. Widely found in Thailand, it is a genetic disorder that stems from a disorder of the synthesis of the beta chain of hemoglobin. One percent of Thailand's population (600,000 Thai people) suffer from this disease and 40 % of the population (24 million Thai people) carry this disease [4]. Severity of this disease is classified into two levels: severe and mild. A particular level can be predicted by the values of six kinds of variables: hemoglobin level, the age when the patient first made a blood transfusion, the need for blood transfusion, the size of the spleen, the age when the disease started to manifest its symptoms, and the level of body growth [17].

In this study, we attempted to find the SNPs that are related to the level of severity of this disease. We investigated the SNPs data of sample groups of Thai population [13] by using machine learning. Our main aims were to find the best technique for selection of the SNPs that are most closely related to the level of severity of beta-thalassemia, to select a limited number of SNPs and to use them to construct an effective model for classifying the level of severity of the disease; its classification performance was evaluated against several baseline methods.

2 METHODOLOGY

This section describes all algorithms and performance measures used in this work.

2.1 Algorithm

2.1.1 Chi-squared Test (χ^2). It is an approach to test the independence of two variables. It works well on categorical data [7]. Assuming that x is a considered SNP for all sample, $x = \{0, 1, 2\}$ and y is a target, y={Case, Control}. Hence, χ^2 can be calculated by the observed frequencies (N_{ij}) and expected frequencies (E_{ij}).

$$\chi^2 = \sum_{i=1}^{C} \sum_{j=1}^{L} \frac{(N_{ij} - E_{ij})}{E_{ij}}, \tag{1}$$

where C and L are the numbers of categories in both variables.

2.1.2 Information Gain (IG). It is a conventional filter technique that assigned weight to each variable. This can be done by finding a relationship between x and y with entropy. The higher the entropy, the higher the degree of association. Entropy, $H(x)$, can be calculated by

$$H(x) = -\sum_{i=1}^{n} P(x_i) \cdot \log_2 P(x_i), \tag{2}$$

where n is a number of category in x. Then entropy of the data that is considered with target class can be calculated by,

$$H(x|y) = -\sum_{j=1} P(y_j) \sum_{i=1} P(x_i|y_j) \log_2(P(x_i|y_j)). \tag{3}$$

Therefore, we can calculate an information gain by

$$IG(X|Y) = H(x) - H(x|y). \tag{4}$$

2.1.3 Gradient Boosting. Gradient boosting (GB) is a machine learning technique that uses an ensemble model, a decision tree, and a boosting method to construct a collection of weak predictors. GB works by sequentially and iteratively adjusting the parameter values of the model to minimise a loss function. It can be used to classify or select features related to the outcome. The importance of a few selected top-ranked variables is determined by the frequency of occurrences of those variables that are split in the process of decision tree construction. Every variable has a variable importance associated with it. The variable importance of each variable is the average value of the variable importances of all decision trees in all iterations [1].

2.1.4 Naïve Bayes (NB). NB is a technique that uses probability concepts in its classifying method. It has become popular in bioinformatics research because it can be used efficiently with data that has a large number of dimensions. Especially well with a dataset with features that are categorical. The outcome of the application of this technique is called a posterior probability [20], which can be calculated by the following equation,

$$P(x_1, x_2, \ldots, x_n|y) = \prod_{i=1}^{n} P(x_i|y), \tag{5}$$

where n is a number of feature.

2.1.5 Support Vector Machine (SVM). SVM is a classification technique that partitions data into two classes by a constructed hyperplane. This technique can project the data in the space of certain dimensions onto another space of a higher number of dimensions by using a kernel function, which enables it to classify non-linear data. SVM attempts to construct support vectors that determine the

Table 1: Confusion Matrix

		Target Values	
		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

boundary of each class and then a hyperplane that can separate the two boundaries farthest apart or to obtain the highest margin [3].

2.2 Performance Measure

2.2.1 Accuracy. In this study, accuracy is a metric for evaluation of the data classification results. It was defined as the ratio of the number of correct predictions to the total number of predictions as shown in (6), which can be any real number between 0 and 1 and the higher the ratio is, the higher the prediction performance is.

$$Accuracy = \frac{\#CorrectClassification}{\#Samples} \tag{6}$$

2.2.2 F1-score. It is another metric for evaluation of the data classification results. It is calculated from the values of 'precision' and 'recall' which makes it more effective in differentiating classification performances. Precision and recall are relations between the number of correct predictions and the number of incorrect predictions for each class. Each of them assumes a value in the confusion matrix shown in Table 1. F1-score values are between 0 and 1, the higher the score is, the higher the prediction performance is. The F1-score is calculated by the following equation,

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{7}$$

where

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

$$Recall = \frac{TP}{TP + FN}. \tag{9}$$

3 EXPERIMENTAL FRAMEWORK

3.1 Dataset

This research investigates on the SNP data of 618 Thai population who are carriers and diagnosed with beta-thalassemia. The dataset consists of two classes that are the severity of this disease–383 of case (severe) and 235 of control (mild). The number of SNPs used in this dataset is 564,831. Any SNPs that contain more than 1 % of missing values are removed. Those with less than 1 % of missing values are substituted by mode. After the processes, we still have 618 population samples but the number of SNP is 451,856.

3.2 Experimental Design

The experiments in this study consist of two main parts: (1) primary screening of the SNPs by ranking all of the SNPs according to their relations to the level of severity of the disease and screening in a number of the most significant SNPs according to the ranks and (2) using these selected SNPs to construct a model for predicting the level of severity of the disease.

3.2.1 Ranking and screening in the SNPs that were related to the level of severity of the disease. An issue with this dataset was that the number of SNPs included in it was very large, much larger than the number of samples, so the chance that a constructed model from these SNPs would encounter an over-fitting problem was very high. Therefore, it was necessary to screening in only a small number of the most significant variables for model construction. In this study, we used a filter method and an embedded method, but not a wrapper method, as our primary screening methods because they could screen a large number of SNPs fast. Therefore, the following techniques were used: Filter method–χ^2 and IG; Wrapper method–GB. These techniques produced a significance value for each variable. The SNPs were then ranked according to these significance values, and the first 250 SNPs of the highest ranks were taken for use in the subsequent part of this experiment.

3.2.2 Using selected SNPs to construct a model. Two hundred and fifty of the most significant SNPs for our purpose that were screened in the previous step was used to construct a classification model for indicating the level of severity of the disease. In order to achieve a highly accurate predictive model, the model was constructed with an individual SNP that was the most significant first then the second most significant SNP was added to the model then the third was added in the same way until all of the selected 250 SNPs were included. The algorithms that were used to construct the models were the following: (1) SVM, (2) NB, and (3) GB. For each model, the optimum values for all of the parameters in the model were determined by m-fold cross validation and grid search. Besides these two techniques, in particular for the SVM model construction, a linear kernel was used and only one necessary parameter, C, needed adjustment; C is a parameter for adjusting the size of the margin of the hyperplane. The trial values of C were $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}]$. On the other hand, NB and GB did not have any necessary parameters to be adjusted in particular like the one in SVM. NB and SVM are implemented in Python with scikit-learn machine learning library and GB is implemented with XGBoost library.

These classification model construction techniques–SVM, NB, and GB–need proper input, ranked input of SNPs that can be obtained from ranking algorithms. The ranking algorithms that we used were χ^2 and IG while GB also did as a ranking algorithm for feature selection. The finer details of all of these techniques and algorithms have been presented in section 2. The combinations of a ranking algorithm and a classification model construction technique that we tested are the following: (1) χ^2+GB, (2) χ^2+NB, (3) χ^2+SVM, (4) IG+GB, (5) IG+NB, (6) IG+SVM, (7) GB+SVM, (8) GB+NB. Their classification performances were evaluated against three conventional classification methods without feature selection techniques–that we considered as our performance baseline.

All of these algorithms, techniques, and methods were configured and evaluated in the same way, i.e., with a five-fold cross validation technique. The SNPs dataset was partitioned into five subsets of exact or nearly the same size: each subset contained around 123–124 samples. Five rounds of validation runs were conducted using five different combinations of four training subsets and one test subset

in a round robin manner. The metrics that we used for classification performance were accuracy and F1-score.

To complete the objective of finding optimum parameters for the best model, cross-validation runs were still conducted further, but this time it was a four-fold cross-validation according to the conventional procedural steps of a cross-validation approach. Hence, this time the entire dataset was partitioned into 4 smaller subsets. set, and every dataset was used as a training dataset and, in its turn, also used as a test dataset in the exact same way as the first round of validation. In other words, four-fold cross-validation were run using a different combination of training and test datasets and the outcomes were recorded. Then, the optimum combination of parameter values that yielded the best mean F1-score was used to construct the final model with all of the training datasets which was then tested with the corresponding test dataset.

4 RESULTS & DISCUSSION

Our experimental results, F1-scores, accuracy, precision and recall values achieved by various tested methods, are shown in Table 2. These values for runs on training and test datasets as well as the number of SNPs used in each method are also presented in the table. It can be seen that in the run that used all SNPs in the dataset to train the model, the prediction results were outstanding: F1-score and accuracy values were 1, meaning that all of the predictions were correct. However, when the model was used with the test dataset, its prediction performance was very low: F1-score was almost 0. This was because the model encountered an over-fitting problem. The models constructed by every method had this problem except the one constructed by GB which achieved a better F1-score of 0.3637. GB achieved the better score because it automatically incorporated weights to features, but it still had some over-fitting problem anyway although not as severe as all of the rest of the methods. The best performance was from the χ^2+SVM method that achieved an F1-score of 0.4873 with only 10 SNPs, followed by χ^2+GB and GB+SVM that achieved F1-scores of 0.4797 and 0.4220 while using 23 and 10 SNPs, respectively. Moreover, χ^2+SVM could achieve the best recall at 0.4992. Of note is that all of these methods performed better than the baseline method, GB, that used all SNPs in the dataset.

In the case that any models gave only one prediction value such as 'severe' for all of the samples, the accuracy value and F1-score would be 0.6190 and 0, respectively. It can be seen in the Table 2 that six methods yielded better results than the randomisation baseline method: χ^2+NB, GB, χ^2+GB, NB, χ^2+SVM, and IG+NB in this order. χ^2+NB not only yielded the best accuracy value but it was also the only method that produced a model that performed equally well with both the training dataset and test dataset in terms of both F1-score and accuracy value, by using only 5 SNPs. It is noted that χ^2+NB yielded the the second best precision at 0.6502 but very poor in recall at 0.2313.

Figure 1 shows a boxplot of mean F-1 scores achieved by various methods from five runs. In terms of F-1 score, the tested methods can be ranked as follows: χ^2+SVM, χ^2+GB, GB+SVM, GB, χ^2+NB, IG+GB, IG+SVM, NB, SVM, IG+NB, GB+NB, and Randomisation, whereas in terms of accuracy value, they are ranked as: χ^2a+NB, GB, χ^2+GB, NB, χ^2+SVM, IG+NB, Randomisation, SVM, GB+NB, GB+SVM, IG+SVM, and IG+GB, as shown in Figure 2. It can be seen that the best method for screening SNPs was χ^2 since the model constructed from it was invariably in the top five models in terms of both F1-score an accuracy value. Furthermore, the mean F1-scores achieved by every method, shown in Table 3, also indicate that, achieving an F1-score of 0.4346, χ^2 screened SNPs better IG and GB that achieved F1-scores of 0.2218 and 0.2110, respectively.

In addition, for comparing the models that yielded the best F1-scores and Accuracy values–χ^2+SVM, χ^2+NB, and GB–we show their confusion matrices in Figure 3. It can be seen that χ^2+NB and GB yielded a lot of '1' results because the samples' data had more 'cases' than 'controls', hence all '1s' predictions would yield a good accuracy without needing to predict any samples to be a 'control' at all. Therefore, the number of false-positive predictions were very high but the numbers of true-negative and false-negative predictions were much lower than those yielded by a method that achieved a very high F1-score such as χ^2-SVM. Therefore, for evaluation of the performances of the models on this dataset, it is not advisable to use only accuracy value because the numbers of 'cases' and 'controls' in the set were very different, i.e., the dataset was an imbalanced dataset. With this kind of dataset, any models that give a single prediction of a 'case' will give a high accuracy value even when it cannot correctly predict the other alternative at all. Consequently, in this case, using F1-score can help reflect the true nature of this kind of prediction better than accuracy value alone can. In this experiment, the top two methods that yielded the highest accuracy values were χ^2+NB and GB, but in terms of F1-score, they ranked the 5th and the 4th, respectively, while χ^2-SVM and χ^2-GB yielded comparable accuracy values but achieved the highest F1-scores, much higher than those achieved by the former two methods. Comparing between the two best methods for this task, χ^2-SVM and χ^2-GB, both of which select 10 SNPs from this dataset, χ^2-SVM is clearly better than χ^2-GB. This is because F1-score, precision and recall values of χ^2-SVM are higher than those obtained by χ^2-GB.

5 CONCLUSION

This paper proposes an approach to screening and selecting the best SNPs related to beta-thalassaemia, which is a necessary step for constructing a good predictive model for the level of severity of the disease because the SNP dataset for this disease is very large but the number of samples is very small. Our attempt was to find a small number of the most significant SNPs related to this disease that can be used to construct the most effective predictive model. Our experiments show that χ^2, which is a popular method for selection of variables that are categorical and a basic method for SNP selection, was the most effective method for SNP selection from this dataset of SNPs and beta-thalassaemia in Thai population. In particular, it was more effective than GB, an embedded method which is a very popular at the present time. The selected SNPs screened by χ^2 were used in the construction of predictive models by SVM, which was shown to perform the best. An expected future study should be to analyse these SNPs in terms of their biological properties in order to pinpoint their true relations to χ^2-thalassaemia.

Table 2: Prediction results of various tested models on the training and test data sets; the parameters of the models were already optimised and the numbers of SNPs used in the models were the ones that yielded the highest F1-score.

Algorithm	#Feature	Training Set				Test Set			
		F1-score	Accuracy	Precison	Recall	F1-score	Accuracy	Precison	Recall
Random	All	0	0.6190	0	0	0	0.6190	0	0
GB	All	1	1	1	1	0.3637	0.6554	**0.6738**	0.29
NB	All	1	1	1	1	0.0254	0.6256	0.2	0.0038
SVM	All	1	1	1	1	0.0074	0.6182	0.3667	0.0179
χ^2+GB	10	0.7047	0.8021	0.7918	0.6529	0.4797	0.6375	0.5182	0.4535
χ^2+NB	5	0.3329	0.6614	0.6648	0.2224	0.3368	**0.6603**	0.6502	0.2313
χ^2+SVM	10	0.7066	0.7989	0.8025	0.6357	**0.4873**	0.6230	0.5398	**0.4992**
IG+GB	9	0.8578	0.8981	0.8931	0.8075	0.3340	0.5276	0.363	0.3122
IG+NB	245	0.0325	0.6235	0.2	0.0085	0	0.6198	0	0
IG+SVM	17	0.9735	0.9871	0.9931	0.9457	0.3316	0.5470	0.3804	0.2979
GB+NB	2	0.0325	0.6235	0.1389	0.0128	0	0.6166	0	0
GB+SVM	23	0.9973	0.9980	0.9921	0.9960	0.4220	0.5922	0.39662	0.3622

Figure 1: F1-scores from every test dataset achieved by the best models in which the number of SNPs were optimised; the best ranks are from left to right.

Table 3: Comparison of average F1-score–averaged across all classifiers for each feature selection technique.

Feature Selection	Model			Average F1
	GB	NB	SVM	
χ^2	0.4797	0.3368	0.4873	**0.4346**
IG	0.3340	0	0.3316	0.2218
GB	-	0	0.4220	0.2110

ACKNOWLEDGMENTS

This work is supported by the Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang under Grant No.: 2561-02-06003.

REFERENCES

[1] Álvaro Alonso Liso. 2016. *Feature selection with Random Forest and Gradient*

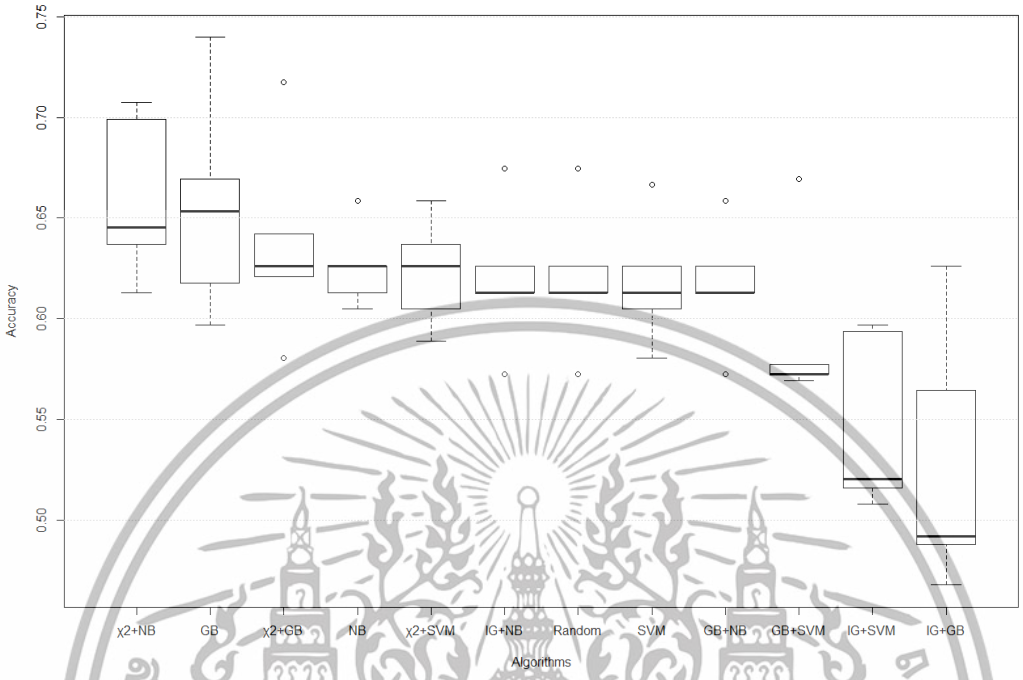
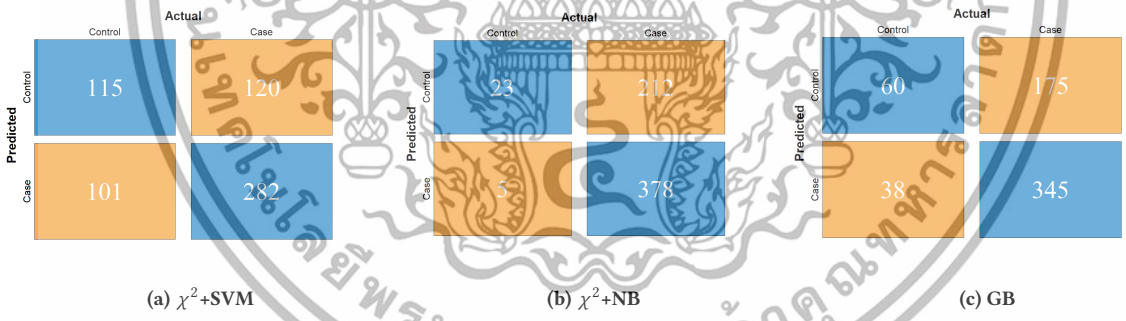


Figure 2: Accuracy from every test dataset achieved by the best models in which the number of SNPs were optimised; the best ranks are from left to right.



(a) χ^2+SVM **(b) χ^2+NB** **(c) GB**

Figure 3: Confusion matrix of the contenders that yield the best F1-scores and Accuracy values.

Boosting. Master's thesis. Department of Computer Engineering, Universidad Autónoma de Madrid.

[2] María Björk Baldursdóttir. 2015. *Analysis of single nucleotide polymorphisms (SNPs) associated with classical Hodgkin lymphoma in patients with infectious mononucleosis: Identification of a common genetic risk.* Bachelor's Thesis. University of Iceland, Iceland.

[3] Weimiao Feng, Jianguo Sun, Liguo Zhang, Cuiling Cao, and Qing Yang. 2016. A support vector machine based naive Bayes algorithm for spam filtering. In *Proceedings of the 35th International on Performance Computing and Communications Conference (IPCCC'2016)*. IEEE, Las Vegas, NV, USA, 1–8.

[4] Thai Thalassemia Foundation. n.d.. Diagnosis of thalassemia carrier. http://www.thalassemia.or.th/thaiversion/diag-carrier-th.htm

[5] Benjamin A Goldstein, Alan E Hubbard, Adele Cutler, and Lisa F Barcellos. 2010. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC genetics* 11, 1 (2010), 49.

[6] Peter Hall, James Stephen Marron, and Amnon Neeman. 2005. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 3 (2005), 427–444.

[7] Xin Jin, Anbang Xu, Rongfang Bie, and Ping Guo. 2006. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In *Proceedings of the International Workshop on Data Mining for Biomedical Applications (BioDM'2006)*. Springer, Singapore, 106–115.

[8] Jin Li, Dongli Huang, Maozu Guo, Xiaoyan Liu, Chunyu Wang, Zhixia Teng, Ruijie Zhang, Yongshuai Jiang, Hongchao Lv, and Limei Wang. 2015. A gene-based information gain method for detecting gene–gene interactions in case–control studies. *European Journal of Human Genetics* 23, 11 (2015), 1566.

[9] Sebastián Maldonado and Richard Weber. 2009. A wrapper method for feature selection using support vector machines. *Information Sciences* 179, 13 (2009), 2208–2217.

[10] Yan Meng, Qiong Yang, Karen T Cuenco, L Adrienne Cupples, Anita L DeStefano, and Kathryn L Lunetta. 2007. Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and Bayesian networks. *BMC proceedings* 1, 1 (2007), S56.

[11] Jason H Moore and Bill C White. 2007. Tuning ReliefF for genome-wide genetic analysis. In *Proceedings of the 5th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO'2007)*. Springer, Valencia, Spain, 166–175.

[12] Thanh-Tung Nguyen, Joshua Zhexue Huang, Qingyao Wu, Thuy Thi Nguyen, and Mark Junjie Li. 2015. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics* 16, 2 (2015), S5.

[13] Manit Nuinoon, Wattanan Makarasara, Taisei Mushiroda, Iswari Setianingsih, Pustika Amalia Wahidiyat, Orapan Sripichai, Natsuhiko Kumasaka, Atsushi Takahashi, Saovaros Svasti, Thongperm Munkongdee, et al. 2010. A genome-wide association identified the common genetic variants influence disease severity in β 0-thalassemia/hemoglobin E. *Human genetics* 127, 3 (2010), 303–314.

[14] Kitsuchart Pasupa. 2007. *Data Mining and Decision Support in Pharmaceutical Databases*. Ph.D. Dissertation. Department of Automatic Control & Systems Engineering, University of Sheffield.

[15] Kitsuchart Pasupa. 2013. A Comparison of Dimensionality Reduction Techniques in Virtual Screening. In *Proceeding of the 12th International Conference on Artificial Intelligence and Soft Computing (ICAISC' 2013) (Lecture Notes in Computer Science)*, Leszek Rutkowski, Marcin Korytkowski, Rafal Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada (Eds.), Vol. 7895. Springer-Verlag, Zakopane, Poland, 297–308.

[16] Shital C Shah and Andrew Kusiak. 2004. Data mining and genetic algorithm based gene/SNP selection. *Artificial intelligence in medicine* 31, 3 (2004), 183–196.

[17] Orapan Sripichai, Wattanan Makarasara, Thongperm Munkongdee, Chutima Kumkhaek, Issarang Nuchprayoon, Ampaiwan Chuansumrit, Suporn Chuncharunee, Nawarat Chantrakoon, Piathip Boonmongkol, Pranee Winichagoon, et al. 2008. A scoring system for the classification of β-thalassemia/Hb E disease severity. *American journal of hematology* 83, 6 (2008), 482–484.

[18] Jyothi Subramanian and Richard Simon. 2013. Overfitting in prediction models–is it a problem only in high dimensions? *Contemporary Clinical Trials* 36, 2 (2013), 636–641.

[19] Bin Wei, Qinke Peng, Jing Li, Xuejiao Kang, and Chenyao Li. 2010. USVM: Selection of SNPs in Diseases Association Study Using UMDA and SVM. In *Proceedings of the 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE'2010)*. IEEE, Chengdu, China, 1–5.

[20] Wei Wei, Shyam Visweswaran, and Gregory F Cooper. 2011. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association* 18, 4 (2011), 370–375.

[21] Qingyao Wu, Yunming Ye, Yang Liu, and Michael K Ng. 2012. SNP selection and classification of genome-wide SNP data using stratified sampling random forests. *IEEE Transactions on Nanobioscience* 11, 3 (2012), 216–227.

[22] Cheng-Hong Yang, Chang-Hsuan Ho, and Li-Yeh Chuang. 2008. Improved tag SNP selection using binary particle swarm optimization. In *Proceedings of the 2008 IEEE Congress on Evolutionary Computation (CEC'2008)*. IEEE, Hong Kong, China, 854–860.

[23] Min Zhang, Yanzhu Lin, Libo Wang, Vitara Pungpapong, James C Fleet, and Dabao Zhang. 2009. Case-control genome-wide association study of rheumatoid arthritis from Genetic Analysis Workshop 16 using penalized orthogonal-components regression-linear discriminant analysis. *BMC Proceedings* 3, 7 (2009), S17.

ภาคผนวก ข

บทความวิชาการ

1. เอก ธรรมวิวัฒน และ **กิติ์สุชาต พสุภา**, วิธีการคัดเลือกสนิปที่มีความสัมพันธ์กับคุณลักษณะพิเศษหรือโรค, วารสารเทคโนโลยีสารสนเทศลาดกระบัง (Submitted)

วิธีการคัดเลือก SNP ที่มีความสัมพันธ์กับคุณลักษณะพิเศษหรือโรค

เอก ธรรมวิวัฒน และ กิติ์สุชาต พสุภา

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Emails: e.thamwiwatthana@gmail.com, kitsuchart@it.kmitl.ac.th

บทคัดย่อ

GWAS (Genome-wide association study) เป็นเทคโนโลยีที่กำลังจะมีส่วนสำคัญในด้านการแพทย์ทั้งในเรื่องของการวินิจฉัยโรค การแพทย์แม่นยำ (Precision medicine) และการรักษาแบบจำเพาะเจาะจงบุคคล เนื่องด้วยลักษณะข้อมูลของ GWAS มักจะมีขนาดมิติที่ใหญ่ แต่ว่ามีขนาดของจำนวนตัวอย่างที่น้อย ทำให้ประสิทธิภาพในการคำนวณนั้นทำได้ไม่ดี และใช้กำลังในการคำนวณมาก อีกทั้งยังไม่สามารถหาความสัมพันธ์ของตัวแปรทางพันธุศาสตร์กับลักษณะพิเศษหรือโรค การคัดเลือกตัวแปรจึงเป็นเรื่องที่มีความสัมพันธ์ และยังมีประโยชน์ช่วยให้ทราบถึงความสัมพันธ์ของตำแหน่งของ SNP ที่มีต่อลักษณะพิเศษหรือโรค ทำให้สามารถวินิจฉัยโรคที่มีความเกี่ยวข้องกับพันธุกรรมได้ง่ายขึ้นอีกด้วย

คำสำคัญ – การคัดเลือกตัวแปร; GWAS; SNP;

1. บทนำ

สิ่งมีชีวิตมีคุณสมบัติถ่ายทอดลักษณะต่างๆ ทั้งเด่นและด้อย ไปสู่รุ่นต่อไป ทั้ง มนุษย์ สัตว์ รวมถึงพืช เรียกว่าการสืบทอดทางพันธุกรรม ทำให้สิ่งมีชีวิตรุ่นมีความแตกต่างจากรุ่นที่ผ่านมา แต่มีลักษณะที่คล้ายกันที่สืบทอดกันมา โดยมีการเริ่มศึกษาโดย เกรกอร์ เมนเดล (Gregor Mendel) บิดาแห่งพันธุศาสตร์ชาวออสเตรีย ได้ทำการศึกษาลักษณะการถ่ายทอดทางพันธุกรรมธ์ ในปี ค.ศ. 1865 ทดลองด้วยการปลูกต้นถั่ว จนสามารถสร้างกฎของเมนเดลขึ้นมาอธิบายลักษณะการถ่ายทอดลักษณะต่างๆของต้นถั่ว ซึ่งเป็นจุดเริ่มต้นของการศึกษาพันธุกรรม

โรคทางพันธุกรรมเป็นโรคที่พบได้บ่อยในวัยทารกตั้งแต่แรกเกิด เป็นสาเหตุสำคัญของการเสียชีวิตในวัยเด็ก มักเป็นโรคเรื้อรังและก่อให้เกิดความพิการ หลายโรคสามารถป้องกันและทำการรักษาได้ การคัดกรองโรคสามารถทำให้ตรวจพบโรคทางพันธุกรรมเพื่อป้องกัน และทำการรักษาได้อย่างทันกาล ในอดีตมีการบันทึกเกี่ยวกับโรคทางพันธุกรรมในตำนานและนิทานพื้นบ้าน เช่น โรคที่ที่ปุ่มขึ้นเต็มตัว คือ โรคท้าวแสนปม [1] หรือโรคหลังค่อมในนิทานเรื่อง คนค่อมแห่งนอร์ธเทอดัม (the Hunchback of Notre Dame) ซึ่งจะเห็นได้ว่าโรคทางพันธุกรรมปรากฏอยู่ในประวัติศาสตร์ทั้งในไทยและในต่างประเทศ ซึ่งเกิดจาก

ความผิดปกติของที่กลายพันธ์ของยีนหรือโครโมโซม ที่มาจากการถ่ายทอดพันธุกรรมจากพ่อและแม่

GWAS เป็นการศึกษาทางด้านพันธุศาสตร์แขนงหนึ่ง ว่าด้วยการศึกษาลักษณะทางพันธุกรรมในกลุ่มประชากร เพื่อหาความสัมพันธ์ของตัวแปรแฝง (Latent variables) ที่อยู่ในรหัสพันธุกรรมกับการเกิดโรค เพื่อพัฒนากระบวนการตรวจหา วิเคราะห์ รักษา หรือป้องกันโรคที่เกิดจากพันธุกรรม สามารถพัฒนาเพื่อหาวิธีการรักษาหรือป้องกันให้เหมาะสมจำเพาะกับบุคคล เพื่อเพิ่มประสิทธิภาพในการใช้ยาหรือกระบวนการรักษา [2] ซึ่งเป็นเครื่องมือสำคัญที่ใช้จัดการตัวแปรทางพันธุศาสตร์ ที่ใช้ระบุคุณลักษณะสำคัญของพันธุกรรมได้ และสามารถใช้ได้ในหลากหลายสาขา เช่น สุขภาพ เกษตรกรรม ปศุสัตว์ การวิวัฒนาการ เป็นต้น การศึกษา GWAS ส่วนใหญ่มุ่งเน้นหาความสัมพันธ์ระหว่าง SNP (Single nucleotide polymorphism) กับคุณลักษณะพิเศษหรือโรค ซึ่งอยู่บนลำดับทางพันธุกรรมหรือที่เรียกว่า DNA ซึ่งลำดับของ DNA ถูกแทนด้วยเบสนิวคลีโอไทด์ 4 ชนิดด้วยกันคือ อะดีนีน (A) กวานีน (G) ไซโทซีน (C) และไทมีน (T) ในสิ่งมีชีวิตเดียวกันมีการเรียงลำดับบน DNA เหมือนกันเกือบทั้งหมด แต่จะมีบางตำแหน่งที่ลำดับเบส แตกต่างกัน ในกลุ่มประชากรถ้าพบตำแหน่งเบส 1 เบสถูกแทนที่บนสาย DNA บริเวณเดียวกัน น้อยกว่าร้อยละ 1 และมักไม่ทำให้เกิดความผิดปกติในสิ่งมีชีวิตเรียกว่า SNP แต่ตำแหน่งการเกิด SNP มักมีความสัมพันธ์กับการเกิดโรคหรือตอบสนองต่อยา [3]

เพื่อที่จะเข้าใจคุณลักษณะแบบซับซ้อนโดยหาความสัมพันธ์ของโรค มักจะใช้ข้อมูล SNP ขนาดใหญ่เพื่อหาชุดของ SNP ที่มีความสัมพันธ์กับคุณลักษณะของโรคการหาตำแหน่งที่มีความสัมพันธ์ของลักษณะพิเศษ ซึ่งมักจะมีจำนวนตำแหน่งจำนวนหลายแสนตำแหน่งขึ้นไปในชุดข้อมูล ในหลายการวิจัยและการศึกษาที่ทำการใช้ SNP เช่น โรคพาร์กินสันมีจำนวน 408,803 SNP โรคอัลไซเมอร์ มีจำนวน 380,157 SNP [3] โรค multiple sclerosis ซึ่งประกอบไปด้วย 325,807 SNP [4] เป็นต้น ข้อมูลที่มีลักษณะจำนวนหลายแสนมิติ ทำให้เกิดปัญหาในการทำการจำแนกคลาสของผลลัพธ์เนื่องจากมีส่วนผมสของข้อมูลอื่น ที่ไม่มีความสัมพันธ์กับผลลัพธ์เป็นจำนวนมาก ใช้กำลังการคำนวณที่มาก ซึ่งถูกเรียกว่า Curses of Dimensionality

2. Curses of Dimensionality

ปัญหาที่มีมิติจำนวนมากเรียกว่า Curses of Dimensionality ส่งผลกระทบโดยตรงต่อการใช้กำลังในการคำนวณ และประสิทธิภาพของโมเดล มีการศึกษาแก้ปัญหานี้ในหลายสาขาด้วยกัน เช่น การทำนายการเลิกใช้งานในโทรศัพท์มือถือ [29] การพยากรณ์เชิงอนุกรมเวลา [30] และใช้ตรวจสอบชิ้นงานที่ผิดพลาดในอุตสาหกรรม [31] เป็นต้น ซึ่งการเพิ่มจำนวนของตัวแปรเข้าสู่โมเดลนำให้เกิด ความต้องการที่เพิ่มขึ้นอย่างเอ็กโปเน็นเชี่ยลของจำนวนข้อมูลกลุ่มตัวอย่าง [21] ต้องใช้ข้อมูลที่มากขึ้นเพื่อรักษาประสิทธิภาพของโมเดล บางอัลกอริทึ่มมีประสิทธิภาพที่ลดลงเมื่อนำมาใช้กับข้อมูลที่มีจำนวนมิติสูงเช่นอัลกอริทึ่มที่เป็นตระกูล Nearest neighbor เช่น KNN, R-tree และ KD-tree เป็นต้น ซึ่งมีหลายวิธีที่พยายามแก้ปัญหา Curses of dimensionality คือ การคัดเลือกตัวแปร และการลดจำนวนมิติ [22]

มีการทดลองเกี่ยวกับ Paralinguistic analysis โดยเน้นการทำการคัดเลือกตัวแปร ทั้งการคัดเลือกกลุ่มตัวแปร และการเรียงลำดับความสำคัญของตัวแปร ด้วยการใช้ k-nearest neighbor เป็นตัวจำแนกคำตอบ พบว่าเมื่อเลือกจำนวนตัวแปรที่สูงในแต่ละเซตแล้ว ประสิทธิภาพของการทำการเลือกตัวแปรนั้นส่งผลที่ไม่ดีในชุดข้อมูลที่ไม่เคยพบมาก่อน แต่ว่าการเลือกตัวแปรด้วยวิธี greedy hill-climbing นั้นมีความทดทานต่อการ overfitting คือการที่โมเดลมีความแม่นยำที่สูงในข้อมูลชุดสอน แต่มีความแม่นยำที่ต่ำในข้อมูลชุดทดลอง หรือข้อมูลที่ไม่เคยเห็นมาก่อน [32] และการทำการคัดเลือกตัวแปรแบบอัตโนมัติเพื่อลดจำนวนของตัวแปรทั้งหมด ให้ผลลัพธ์ที่ดีกว่าการใช้ Support vector machine (SVM) หรือ RF (Random Forest) ที่ใช้ข้อมูลทั้งหมดในการคำนวณ [23] มีการนำเสนออัลกอริทึม Fast clustering-based feature selection (FAST) ซึ่งมีขั้นตอนการทำงานด้วยการแบ่งตัวแปรลงในคลัสเตอร์โดยใช้ Graph-theoretic ซึ่งชุดตัวแปรที่มีความสัมพันธ์กับคลาสเป้าหมายจะถูกเลือกจากแต่ละคลัสเตอร์มาเป็นสับเซตของตัวแปร [24]

จำนวน SNP ที่มักจะมีใช้ในการวิจัยในแต่ละโรคนั้นมักจะมีจำนวนประมาณ หลักร้อยถึงหลักพัน แต่มีจำนวนของมิติหลักแสน ซึ่งลักษณะข้อมูลนี้เรียกว่า HDLSS (High-dimension low sample size) ซึ่งต้องการการคัดเลือกตัวแปรหรือการลดมิติก่อนที่จะทำการจำแนกคำตอบของข้อมูลก่อน ซึ่งข้อมูลที่มีซับซ้อนมากกว่าข้อมูลที่มีมิติ น้อย แต่มีจำนวนมิติ ที่มีความสัมพันธ์กับโรคเพียงจำนวนเล็กน้อย ทำความเข้าใจและแสดงผลได้ยาก มีปัญหาความซับซ้อนสูงต้องการการคำนวณที่มาก ทำให้สามารถประมวลผลได้ช้าหรือไม่สามารถประมวลผลได้ มีหลายงานวิจัยที่พยายามแก้ไขปัญหานี้โดยพัฒนาวิธีการ หาความสัมพันธ์ระหว่างโรคกับ SNP ที่มีข้อมูลลักษณะนี้ มีหลักวิธีอยู่ 3 แบบ คือ 1. Filter method 2. Wrapper method 3. Embedded method [5]

3. Filter Method

Filter Method เป็นวิธีการคัดเลือกตัวแปรด้วยการหาสหสัมพันธ์ (Correlation) ระกว่างชุดข้อมูลกับค่าของผลลัพธ์ แล้วทำการเรียงอันดับค่าคะแนนความสัมพันธ์ของแต่ละตัวแปร เป็นขั้นตอนในการทำงานช่วง pre-processing ที่ไม่ต้องเกี่ยวข้องกับอัลกอริทึมที่ใช้จำแนกผลลัพธ์ ตัวอย่างเช่น ระยะทางระหว่างผลลัพธ์ หรือ ความสัมพันธ์ทางสถิติ ซึ่งเป็นวิธีที่มีความเร็วกว่าวิธีการเลือกตัวแปรตัวแปรที่มีความสัมพันธ์วิธีอื่น แต่อย่างไรก็ตาม วิธีนี้มีแนวโน้มที่จะเลือกชุดของข้อมูลที่มีขนาดใหญ่ และต้องการกำหนดค่าขีดแบ่งก่อนที่จะเลือกข้อมูลกลุ่มย่อย [6]

3.1. Chi-square

Chi-square ใช้สำหรับจัดเรียงค่าตัวแปรของข้อมูล เป็นวิธีทดสอบพื้นฐาน โดยทั่วไปแล้ววิธีการนี้มักจะใช้ช่วยในการตัดสินใจว่าควรตอบรับหรือปฏิเสธสมมติฐาน [7] เปรียบเทียบความสัมพันธ์ระหว่าง 2 สิ่งที่เป็นอิสระต่อกันว่ามีความเกี่ยวข้องกันหรือไม่ แต่ในบางการศึกษาพบว่าวิธีนี้ไม่เหมาะสมกับข้อมูลแบบที่มีมิติสูง [8] ซึ่งบางครั้งวิธีนี้ถูกนำไปใช้เพื่อกรองตัวแปร โดย p-values ก่อนนำไปใช้ในการเลือกตำแหน่งของตัวแปร ที่มีความสัมพันธ์ที่เกี่ยวข้องต่อไปอีกทีหนึ่ง [9]

3.2. ReliefF

เป็นอัลกอริทึมที่หาค่าระยะทางของตัวละตัวแปร พัฒนามากจากอัลกอริทึม Relief ใช้ค่าน้ำหนักของแต่ละตัวแปรเพื่อทำการคัดเลือกตัวแปร ที่มีความสัมพันธ์กับค่าคำตอบ มีข้อดีคือสามารถทำงานกับข้อมูลที่ไม่สมบูรณ์และปัญหาแบบหลายคลาสได้ และมีความแตกต่างกับ อัลกอริทึม Relief ตรงที่ใช้ระยะทางแบบ Manhattan แทนแบบ Euclidean มีวิธีการทำงานด้วยการหา nearest neighbors ระหว่างคลาสคำตอบ หาผลต่างของจำนวน nearest neighbors เพื่อนำมาคำนวณกับค่าน้ำหนัก แล้วทำงานวนซ้ำจนครบทุกคุณลักษณะ จัดเรียงค่าน้ำหนักของแต่ละตัวแปร เพื่อคัดเลือกชุดตัวแปรที่มีความสัมพันธ์ที่เกี่ยวข้อง [10,11,12]

3.3. Feature Selection using Feature Similarity (FSFS)

เป็นวิธีการที่ใช้การคำนวณมากในการสร้างเซตย่อยของ SNP แต่ลดความซ้ำซ้อนของ SNP ด้วยวิธีการเลือกตัวแปรที่มีความสัมพันธ์ ซึ่งมีการทำงานโดยการใช้วิธีวัด Feature Similarity ด้วยการใช้ค่า k เพื่อหา nearest neighbor เพื่อทำการ tagging SNP ซึ่งเป็นตัวแปรที่มีคุณลักษณะ ในการหาความสัมพันธ์กับโรค ด้วยการหาระยะทางระหว่างตัวแปร ด้วยค่า k ในแต่ละรอบเพื่อการคำนวณ เพื่อหาเซตย่อยของ SNP และค่อยๆทำการลดค่า k แล้วทำการคำนวณใหม่ จนกว่าระยะทางไม่มากกว่าค่าขีดแบ่ง เพื่อนำ SNP ที่มีความต่างมากกว่าค่าขีดแบ่งออกไปจากเซ็ตย่อยของข้อมูล [13]

4. Wrapper Method

ใช้การเลือกชุดสับเซตของตัวแปรด้วย learning algorithm ที่ใช้ข้อมูลชุดสอนของตัวโมเดลเอง มักจะมีประสิทธิภาพที่ดีกว่าวิธีแบบ Filter method แต่มีข้อเสียคือใช้การคำนวณและเวลาที่มากกว่า ซึ่งวิธีที่นำเสนอต่อไปนี้เรียกว่า Randomized wrapper ตัวอย่างเช่น PSO (Particle Swarm Optimization) หรือ GA (Genetic algorithm) [14]

4.2. PSO

PSO ถูกนำมาหาสับเซตของตัวแปรที่ดีที่สุดในหลายงานวิจัย [33,34] เพื่อนำไปทำการจำแนกโรคหรือคุณลักษณะพิเศษ ซึ่งมีแนวคิดจากการจำลองพฤติกรรมทางสังคมในการแลกเปลี่ยนข่าวสารเพื่อเคลื่อนที่ไปแหล่งอาหารของนก ซึ่งมีหลักการทำงานโดย 1. สุ่มตำแหน่งเริ่มต้นให้กับประชากร 1 ตัวในกลุ่ม แล้วจึงกำหนดความเร็ว 2. หาค่าความเหมาะสม (Fitness) ถ้าค่าที่ได้มีค่ามากกว่าค่าเดิม ให้เลือกใช้ค่าที่มีค่ามากกว่า 3.เลือกค่าความเหมาะสมที่ดีที่สุด 4. วนซ้ำแล้วกำหนดความเร็วรอบใหม่ของประชากร ซึ่งมีความแตกต่างจาก GA คือ มีวิธีการบางอย่างที่มีความคล้ายคลึงกับ กัน แต่ว่า PSO นั้นไม่มีกระบวนการที่ GA มีเช่น Crossover และ Mutation ซึ่ง PSO ใช้การปรับปรุงค่า ความเร็วภายในล่าสุดแทน [35]

4.3. GA

ใช้สำหรับหาเซตของ feature ที่เหมาะสมที่สุด ที่ส่งผลกับความแม่นยำในการ classification [14] ซึ่งวิธีนี้เป็นวิธีที่ได้รับแนวทางมาจากกลไกทางชีววิทยา [17] มีวิธีการทำงานโดย 1. นำข้อมูลที่มีอยู่เข้ารหัส (Encode) ชุดข้อมูล 2. สุ่มชุดประชากรเริ่มต้น 3. พิจารณาหาความเหมาะสมของประชากรบน Fitness function ที่กำหนดไว้ 4. ทำการวนซ้ำการทำ Crossover, Mutation และคำนวณหาค่า Fitness จนกว่าจะได้กลุ่มประชากรที่ดีที่สุด ด้วยการนำมาทดลองกับโมเดล หาประสิทธิภาพ เพื่อหาสับเซตของตัวแปรที่ดีที่สุด ซึ่งเป็นตัวแปรที่ Fitness ที่สุด และ

ผลลัพธ์อยู่ในเกณฑ์ที่พอใจหรือถึงรุ่นสุดท้ายที่ได้กำหนดเอาไว้

5. Embedded Method

เป็นวิธีที่เกิดขึ้นจากการรวมเอาข้อดีของทั้งสองวิธีก่อนหน้านี้เข้าด้วยกัน โดยมีความสามารถในการคำนวณที่ดีกว่าแบบ wrapper method เทคนิคที่เป็นที่นิยมและมีประสิทธิภาพ มักจะถูกนำมาใช้งานคือ RF ซึ่งเป็นชุดของโมเดล classifier ที่สร้างเป็น ensemble tree เพื่อนำมาระบุคลาสของคำตอบ โดยแต่ละ tree ถูกสร้างมาจากการใช้ bootstrap ของข้อมูลกลุ่มตัวอย่าง ซึ่งถูกใช้โดยทั่วไปในงานศึกษาหาความสัมพันธ์ทางด้านพันธุศาสตร์ ซึ่งสามารถทำหน้าที่ได้ทั้งพยากรณ์คำตอบและหาค่าความสำคัญของตัวแปร (Variable importance) เป็นเพียงไม่กี่อัลกอริทึมที่สามารถจัดการตัวแปรจำนวนหลายแสนได้อย่างมีประสิทธิภาพ [18]

ในปี 2010 มีการศึกษาการนำ RF เข้ามาใช้กับ GWAS เป็นครั้งแรกด้วยการใช้ข้อมูลโรค Multiple sclerosis ซึ่งมีจำนวน SNP มากกว่า 300,000 ตำแหน่ง พบว่าค่าพารามิเตอร์พื้นฐานของ RF ไม่เหมาะสมกับข้อมูล GWAS ขนาดใหญ่ และพบว่า การทำการสุ่มตัวอย่างของข้อมูล การทำการ pruning บนพื้นฐานของ Linkage disequilibrium คือ SNP ที่อยู่คนละตำแหน่งบน โครโมโซมเดียวกันที่อาจจะมีความสัมพันธ์กัน และการนำตัวแปรที่มีผลรุนแรงออกจากโมเดล นั้นไม่เหมาะสมกับการใช้วิธีนี้ ดังนั้นควรทำการปรับค่าพารามิเตอร์ให้เหมาะสมด้วย [4]

มีการนำ RF มาใช้โดยใช้การเลือกตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Sampling) เพื่อเลือกตัวแปร SNP โดยการสุ่มจำนวน SNP จากแต่ละกลุ่มโดยนำมาผสมกันเพื่อนำมาสร้างต้นไม้ ซึ่งมีข้อดีของการใช้วิธีนี้คือสามารถ สร้างชุดข้อมูลที่แน่ใจได้ว่ามีตัวแปร SNP ที่มีประโยชน์ในการคำนวณ และสามารถหลบเลี่ยงปัญหาที่ต้องใช้การคำนวณเป็นจำนวนมากในการตั้งค่าพารามิเตอร์ได้ [19]

บางครั้ง RF ก็ถูกนำมาใช้เป็นโมเดลที่ใช้ทำการจำแนกโรคร่วมกับการหาความสัมพันธ์ของ SNP ในแนวทางอื่น เช่น มีการทดลอง 2 แนวทางกับข้อมูลโรคอัลไซเมอร์ด้วยการใช้ logistic regression เพื่อกรองข้อมูลด้วย p-value ใช้การเลือกตัวแปรที่มีค่า p-value ที่น้อยที่สุดที่น้อยกว่าค่าขีดแบ่ง Bonferroni แล้วนำสับเซตของ SNP มาทำการจำแนกด้วย RF แล้วเปรียบเทียบผลลัพธ์กับการใช้ข้อมูลการคัดเลือก SNP จากความรู้ทางชีววิทยาที่มีอยู่ พบว่าวิธีการแรกนั้นได้ผลลัพธ์ที่มีค่าผิดพลาดเฉลี่ย 9.8% ซึ่งมีประสิทธิภาพดีกว่าแบบใช้ข้อมูลทางชีววิทยาที่มีอยู่ก่อนซึ่งมีความผิดพลาดเฉลี่ยอยู่ที่ 17.5% [20]

การนำค่า p-value ที่เป็น filter method มาเพื่อทำการคัดเลือกตัวแปร SNP เพื่อแบ่งเป็น 2 กลุ่ม กลุ่มที่ 1 เลือกตัวแปรที่มีค่า informative สูง และกลุ่มที่ 2 เลือกตัวแปรกลุ่มที่มีค่า informative ต่ำ เมื่อทำการสุ่มตัวแปรเพื่อสร้างต้นไม้สำหรับ RF แล้วจะพบว่าตัวแปรที่มีค่า informative สูงจะถูกใช้ในการ split node ของต้นไม้ ซึ่งทำให้สามารถสร้างวิธีการที่มีความแม่นยำมีค่าความผิดพลาดต่ำ และสามารถหลบเลี่ยงการ overfitting ได้ โดยทดลองกับข้อมูลโรคพาร์กินสันและโรคอัลไซเมอร์[3]

6. Hybrid method

เป็นการรวมข้อดีของ Filter method และ Wrapper method เช่นกัน โดยช่วงการทำ filter นั้นเพื่อการนำตัวแปรตัวที่มีความสัมพันธ์น้อยออกไปก่อนจะนำมาใช้ Wrapper method เพื่อหาสับเซตของชุดตัวแปร เนื่องจากการใช้ Filter method นั้นใช้เวลาและ

การคำนวณที่น้อย ซึ่งการใช้ Wrapper method มักจะให้ผลลัพธ์เมื่อทำไปทำการคำนวณได้ดีกว่า

มีการทดลองใช้ Symmetrical uncertainly ระหว่างชุดตัวแปรและคลาสคำตอบ เพื่อใช้หาค่า goodness of feature ซึ่งเลือกตัวแปรที่มีค่า SU มากกว่าค่าขีดแข่ง เพื่อนำมาสร้างเป็นประชากรเริ่มต้นของ GA ซึ่งได้ผลลัพธ์ที่ดีกว่าการใช้วิธีเดี่ยวๆ หรือใช้ตัวแปรทั้งหมด [25]

F-score และ Information gain ถูกนำมาใช้เพื่อนำข้อมูลที่มีความซ้ำซ้อนและไม่เกี่ยวข้องออกไปจากข้อมูล ซึ่งวิธีนี้เป็นวิธีการแบบ Filter method แล้วนำตัวแปรที่มีความสัมพันธ์ของทั้งสองวิธีมารวมกันสร้างเป็นสับเซตของชุดข้อมูล เรียกว่า Combination model นำข้อมูลที่ได้มาคัดเลือกอีกครั้งด้วย การทำ Fine tuning ด้วย Inversed sequential floating search method เพื่อสร้างเซตข้อมูลที่เหมาะสมชุดสุดท้าย [26] ยังมีการใช้ F-score เพื่อสร้างข้อมูลที่ถูก Filter ชุดเริ่มต้นเพื่อสร้างข้อมูลการเลือกตัวแปรชุดแรกก่อนที่จะนำไปใช้ใน Wrapper method อย่างเช่น Supported Sequential Forward Search เพื่อเลือกตัวแปรก่อนที่จะนำไปจำแนกคลาส [27]

นอกจาก F-score แล้วการใช้สหสัมพันธ์ก็เป็นวิธีหนึ่งที่ถูกนำมาใช้ในการเลือกตัวแปรแบบ Hybrid method ในการวินิจฉัยโรคมะเร็งจาก Microarray ด้วยการรวมการใช้การหาสหสัมพันธ์เพื่อ Filter ความสัมพันธ์ของตัวแปรแล้วใช้ Particle swarm optimization เพื่อสร้างสับเซตของชุดตัวแปรเพื่อนำมาจำแนกคลาสคำตอบด้วยอัลกอริทึม Extreme learning machines[28]

7. สรุป

บทความนี้ได้นำเสนองานวิจัยที่เกี่ยวข้องในการหาความสัมพันธ์ของ SNP กับโรคที่เกี่ยวข้อง ด้วยหลักวิธีที่เป็นที่นิยม 3 แบบ คือ 1. Filter method เช่น Chi-squard ซึ่งเป็นวิธีพื้นฐานได้รับความนิยมในการหาความสำคัญของตัวแปรกับคำตอบ หรือ วิธีการที่ใช้ค่าระยะทางในการหาความสัมพันธ์ เช่น Feature Similarity หรือ ReliefF ถูกนำไปใช้งานในด้าน Sensors ที่มักจะมีมิติขนาดใหญ่ [39,40] 2. Wrapper method เช่น PSO และ GA ซึ่งเป็นที่นิยมในอัลกอริทึมแบบ Bio-inspired ที่ถูกนำไปใช้เป็นอย่างมากในการทำ Optimization [36,37,38,41] ซึ่งทำการสุ่มชุดตัวอย่างเพื่อวนรอบหาชุดตัวแปรที่มีความสมพันธ์กับตอบที่สุด 3. Embedded Method ซึ่งมีข้อดีข้อเสียและประสิทธิภาพที่แตกต่างกัน RF เป็นวิธีหนึ่งซึ่งได้รับความนิยมอย่างมากในช่วงไม่กี่ปีที่ผ่านมาเนื่องจากมีประสิทธิภาพที่สูง ถูกใช้ในหลายๆอุตสาหกรรม เช่น การทำนายการเลิกใช้งาน [42] หรือตรวจจับผู้บุกรุกในระบบ [43] เนื่องจากมีการสร้างเครื่องมือให้ใช้นำไปใช้งานได้ง่าย มีความเร็วสูง ประมวลผลแบบขนานได้ ทำให้สามารถทำงานกับข้อมูลขนาดใหญ่ได้ง่าย การนำวิธีหาตัวแปรที่มีความสัมพันธ์กับโรคหรือลักษณะเด่นนั้น แต่ละแบบมีอัลกอริทึมที่เป็นที่นิยมใช้และเกี่ยวข้องกับงานวิจัยด้าน GWAS ซึ่งบางครั้งมีการใช้วิธีเหล่านี้ร่วมกันเพื่อเพิ่มประสิทธิภาพ ลดระยะเวลาและกำลังที่ต้องใช้ในการคำนวณ ปัญหาในการคำนวณที่ต้องใช้กำลังในการคำนวณมากมีอีกวิธีหนึ่งซึ่งเรียกว่า Dimensionality Reduction ซึ่งเป็นการลดมิติในการคำนวณแต่ยังสามารถเป็นตัวแทนในชุดข้อมูลได้ แต่เนื่องจากจุดมุ่งหมายในการทำ GWAS นั้นเพื่อจะหา SNP หรือตัวแปรทางพันธุศาสตร์ที่มีความเกี่ยวข้องกับคุณลักษณะพิเศษหรือโรค ทำให้วิธีนี้มักจะไม่ตรงกับความต้องการในการหาผลลัพธ์ ในช่วงหลังงานวิจัยที่เกี่ยวกับ GWAS มักจะพบว่าอัลกอริทึม

ประเภท Random Forest ได้รับความนิยมสูงในการทำ GWAS เนื่องจากสามารถใช้เป็นตัวจำแนกข้อมูลได้อีกทั้งยังสามารถหาความสัมพันธ์ของ SNP กับคุณลักษณะพิเศษได้ นอกจากนี้มีการใช้ Hybrid method ซึ่งเป็นการนำข้อดีของ Filter method ที่มีความเร็วในการทำงาน และใช้กำลังในการคำนวณที่น้อย ร่วมกับการใช้ Wrapper method ที่มีประสิทธิภาพเมื่อนำสับเซตของข้อมูลเพื่อไปใช้กับโมเดลเพื่อจำแนกคลาส สามารถนำอัลกอริทึ่มของแต่ละแบบมาผสมกันแล้วหาวิธีที่เหมาะสมกับข้อมูลแต่ละแบบได้ ข้อมูล GWAS มีลักษณะที่เป็น HDLSS ทำให้คำนวณได้ยากลำบากหรือไม่สามารถคำนวณได้ ถ้าเครื่องมือที่มี ไม่มีประสิทธิภาพเพียงพอเมื่อนำมาใช้กับ Wrapper method การใช้เฉพาะ Filter method ก็อาจจะไม่มีประสิทธิภาพเพียงพอหรือไม่สามารถค้นหาความสัมพันธ์ที่แฝงได้เทียบเท่า ดังนั้นการนำวิธี Filter method มาใช้เพื่อกรองข้อมูลส่วนที่ซ้ำซ้อนหรือมีสหสัมพันธ์กับคำตอบต่ำ ทำสับเซตข้อมูลที่มีความสัมพันธ์กับโรคหรือลักษณะเด่นมาใช้กับ Wrapper method ซึ่งมีประสิทธิภาพสูงกว่า ซึ่งสามารถลดตัวแปรในขั้นตอนแรกได้แล้ว ทำให้การทำงานในขั้นที่สองนั้นทำได้ง่ายขึ้น และเร็วขึ้น

เอกสารอ้างอิง

[1] ศาสตราจารย์เกียรติคุณ แพทย์หญิงพรสวรรค์ วสันต์. ผศ.ดร. ว่าที่ร้อยตรี เจษฎา เด่นดวง บริพันธ์. “โรคพันธุกรรมในเด็ก” สารานุกรมไทยสำหรับเยาวชน. [Online]. Available: http://kanchanapisek.or.th/kp6/Ebook/BOOK38/pdf/book38_9.pdf

[2] National Human Genome Research Institute. (201, Aug 27). Genome-Wide Association Studies [Online]. Available: https://www.genome.gov/20019523/

[3.] Nguyen, Thanh-Tung, et al. "Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests." *BMC genomics*. Vol. 16. No. Suppl 2. BioMed Central Ltd, 2015. HsinChu, Taiwan, Jan. 2015.

[4] Goldstein, Benjamin A., et al. "An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings." *BMC genetics* 11.1 (2010): 49.

[5] Krawczuk, Jerzy, and Tomasz Łukaszuk. "The feature selection bias problem in relation to high-dimensional gene data." *Artificial intelligence in medicine* 66 (2016): 63-71.

[6] Sánchez-Maroño, Noelia, Amparo Alonso-Betanzos, and María Tombilla-Sanromán. "Filter methods for feature selection–a comparative study." *Intelligent Data Engineering and Automated Learning-IDEAL 2007* (2007): 178-187.

[7] Suzuki, David T., and Anthony JF Griffiths. *An introduction to genetic analysis*. WH Freeman and Company., 1976.

[8] Vafaie, Haleh, and Ibrahim F. Imam. "Feature selection methods: genetic algorithms vs. greedy-like search." *Proceedings of the International Conference on Fuzzy and Intelligent Control Systems*. Vol. 51. 1994.

[9] Briones, Natalia, and Valentin Dinu. "Data mining of high density genomic variant data for prediction of Alzheimer's disease risk." *BMC medical genetics* 13.1 (2012): 7.

[10] Eppstein, Margaret J., and Paul Haake. "Very large scale ReliefF for genome-wide association analysis." *Computational*

Intelligence in Bioinformatics and Computational Biology, 2008. CIBCB'08. IEEE Symposium on. IEEE, 2008.

[11]Chandra, B., and Manish Gupta. "An efficient statistical feature selection approach for classification of gene expression data." *Journal of biomedical informatics* 44.4 (2011): 529-535.

[12]Niel, Clément, et al. "A survey about methods dedicated to epistasis detection." *Frontiers in genetics* 6 (2015).

[13]Phuong, Tu Minh, Zhen Lin, and Russ B. Altman. "Choosing SNPs using feature selection." *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*. IEEE, 2005.

[14]Hira, Zena M., and Duncan F. Gillies. "A review of feature selection and feature extraction methods applied on microarray data." *Advances in bioinformatics* 2015 (2015).

[15] Maldonado, Sebastián, and Richard Weber. "A wrapper method for feature selection using support vector machines." *Information Sciences* 179.13 (2009): 2208-2217.

[16] Wei, Bin, et al. "USVM: Selection of SNPs in Diseases Association Study Using UMDA and SVM." *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*. IEEE, 2010.

[17] Shah, Shital C., and Andrew Kusiak. "Data mining and genetic algorithm based gene/SNP selection." *Artificial intelligence in medicine* 31.3 (2004): 183-196.

[18] Goldstein, Benjamin A., Eric C. Polley, and Farren Briggs. "Random forests for genetic association studies." *Statistical applications in genetics and molecular biology* 10.1 (2011).

[19] Wu, Qingyao, et al. "SNP selection and classification of genome-wide SNP data using stratified sampling random forests." *IEEE transactions on nanobioscience* 11.3 (2012): 216-227.

[2 0] Briones, Natalia, and Valentin Dinu. "Data mining of high density genomic variant data for prediction of Alzheimer's disease risk." *BMC medical genetics* 13.1 (2012): 7.

[21] Bessa, M. A., et al. "A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality." *Computer Methods in Applied Mechanics and Engineering* 320 (2017): 633-667.

[22] Sammut, Claude, and Geoffrey I. Webb, eds. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[23] Pohjalainen, Jouni, Okko Räsänen, and Serdar Kadioglu. "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits." *Computer Speech & Language* 29.1 (2015): 145-171.

[24] Song, Qinbao, Jingjie Ni, and Guangtao Wang. "A fast clustering-based feature subset selection algorithm for high-dimensional data." *IEEE transactions on knowledge and data engineering* 25.1 (2013): 1-14.

[25] Jiang, Bai-ning, et al. "A hybrid feature selection algorithm: Combination of symmetrical uncertainty and genetic algorithms." *The Second International*

Symposium on Optimization and Systems Biology. 2008.

[26] Hsu, Hui-Huang, Cheng-Wei Hsieh, and Ming-Da Lu. "Hybrid feature selection by combining filters and wrappers." *Expert Systems with Applications* 38.7 (2011): 8144-8150.

[27] Lee, Ming-Chi. "Using support vector machine with a hybrid feature selection method to the stock trend prediction." *Expert Systems with Applications* 36.8 (2009): 10896-10904.

[28] Chinnaswamy,Arunkumar, and Ramakrishnan Srinivasan. "Hybrid Feature Selection Using Correlation Coefficient and Particle Swarm Optimization on Microarray Gene Expression Data." *Innovations in Bio-Inspired Computing and Applications*. Springer International Publishing, 2016. 229-239.

[29] Bengio, Samy, and Yoshua Bengio. "Taking on the curse of dimensionality in joint distributions using neural networks." *IEEE Transactions on Neural Networks* 11.3 (2000): 550-557.

[30] Verleysen, Michel, and Damien François. "The curse of dimensionality in data mining and time series prediction." *International Work-Conference on Artificial Neural Networks*. Springer Berlin Heidelberg, 2005.

[31] Apte, Chid, Sholom Weiss, and Gordon Grout. "Predicting defects in disk drive manufacturing: A case study in high-dimensional classification." *Artificial Intelligence for Applications, 1993. Proceedings., Ninth Conference on*. IEEE, 1993.

[32] Subramanian, Jyothi, and Richard Simon. "Overfitting in prediction models–Is it a problem only in high dimensions?." *Contemporary clinical trials* 36.2 (2013): 636-641.

[33] Qian, Weiyi, et al. "Particle swarm optimization for SNP haplotype reconstruction problem." *Applied mathematics and Computation* 196.1 (2008): 266-272.

[34] Chuang, Li-Yeh, et al. "An improved PSO algorithm for generating protective SNP barcodes in breast cancer." *PLoS One* 7.5 (2012): e37018.

[35] R. Kennedy J., Eberhart, R. C. (2006). PSO tutorial [Online]. Available: http://www.swarmintelligence.org/tutorials.php

[36] Nourmohammadzadeh, Abtin, and Sven Hartmann. "The Fuel-Efficient Platooning of Heavy Duty Vehicles by Mathematical Programming and Genetic Algorithm." *Theory and Practice of Natural Computing: 5th International Conference, TPNC 2016, Sendai, Japan, December 12-13, 2016, Proceedings* 5. Springer International Publishing, 2016.

[37] Poli, L., G. Oliveri, and A. Massa. "An integer genetic algorithm for optimal clustering in phased array antenna." *Applied Computational Electromagnetics Society Symposium-Italy (ACES), 2017 International*. IEEE, 2017.

[38] Lai, Chunyan, et al. "Genetic algorithm based current optimization for torque ripple reduction of interior PMSMs." *Electrical Machines (ICEM), 2016 XXII International Conference on*. IEEE, 2016.

[39] Zhang, Jianhai, et al. "ReliefF-based EEG sensor selection methods for emotion recognition." *Sensors* 16.10 (2016): 1558.

[40] Wang, Zhi, et al. "Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image." *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*. IEEE, 2016.

[41] Soesanti, Indah, and Ramadoni Syahputra. "Batik Production Process Optimization Using Particle Swarm Optimization Method." *Journal of Theoretical and Applied Information Technology* 86.2 (2016): 272.

[42] Li, Hui, et al. "Enhancing telco service quality with big data enabled churn analysis: infrastructure, model, and deployment." *Journal of Computer Science and Technology* 30.6 (2015): 1201-1214.

[43] Zhang, Jiong, Mohammad Zulkernine, and Anwar Haque. "Random-forests-based network intrusion detection systems." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.5 (2008): 649-659.

ข้อมูลประวัติคณะผู้วิจัย

ประวัติส่วนตัว

ชื่อ-สกุล ดร. กิติ์สุชาต พสุภา

ตำแหน่งปัจจุบัน รองศาสตราจารย์

ประวัติการศึกษา

ปริญญา	สาขา	สถาบันที่จบ	ปีที่จบ
PhD	Automatic Control & Systems Engineering	The University of Sheffield	2008
MSc(Eng)	Control Systems	The University of Sheffield	2004
BEng	Electrical Engineering	Sirindhorn International Institute of Technology, Thammasat University	2003

สาขาวิจัยที่มีความชำนาญพิเศษ

Machine Learning, Data Science, Pattern Recognition, Eye Movements

รางวัลด้านวิชาการ/ด้านวิจัย ที่ได้รับ

ค.ศ.	ชื่อรางวัล	สถาบันที่ให้
2014	The Honorable Fame Award for Outstanding Academic Staffs	King Mongkut's Institute of Technology Ladkrabang, Thailand
2014	Best Paper Award	10th National Conference on Computing and Information Technology (NCCIT 2014), 8-9 May 2014, Phuket, Thailand
2015	Best Paper Award	7th National Conference on Information Technology (NCIT 2015), 29-30 October 2015, Chiang Mai, Thailand
2016	The Honorable Fame Award for Outstanding Academic Staffs	King Mongkut's Institute of Technology Ladkrabang, Thailand
2016	Special Award	NAPROCK 8th International Programming Contest, 8-9 October 2016, Ise-shi, Japan
2017	Excellence Teaching Award	King Mongkut's Institute of Technology Ladkrabang, Thailand
2017	Third Prize	19th National Software Contest - Mobile Application (NSC 2017), 15-17 March 2017, Bangkok, Thailand
2018	Third Prize	20th National Software Contest - Artificial Intelligence Application (NSC 2018), 14-16 March 2018, Bangkok, Thailand
2018	Third Prize	20th National Software Contest - Human Detection Contest (NSC 2018), 14-16 March 2018, Bangkok, Thailand
2018	Best Paper Award	14th International Conference on Computing and Information Technology (IC2IT 2018), 5-6 July 2018, Chiang Mai, Thailand
2019	Best Paper Award	11th International Conference on Knowledge and Smart Technology (KST 2019), 23-26 January 2019, Phuket, Thailand

ทุนวิจัยที่เคยได้รับ

ระยะเวลา	หน้าที่	โครงการ	แหล่งทุน	จำนวนเงิน
06/2019-07/2019	Principal Investigator	DeepWaste: Waste Classification and Recycling Rate Estimation Based on Deep Learning Technique	ASEA-UNINET	120,000 THB
19/2017-09/2018	Principal Investigator	An Approach to Select a Group of Associated SNP with Genetic Traits	IT.KMITL	50,000 THB

ระยะเวลา	หน้าที่	โครงการ	แหล่งทุน	จำนวนเงิน
10/2017-09/2018	Researcher	Development of an Information Processing and Data Analytics System for Electronic Examination	KMITL (Annual Government Statement of Expenditure)	
10/2016-09/2017	Principal Investigator	Prediction of Human Emotions Stimulated by Content in Image with Eye Movement Data	IT.KMITL	100,000 THB
10/2014-01/2015	Principal Investigator	The Development of Career Roadmap System	Siam Commercial Bank PCL	100,000 THB
07/2014-08/2014	Principal Investigator	Prediction of Eye Gaze Behavior with Image Features	AUN/SEED-Net	170,000 THB
09/2013-08/2014	Principal Investigator	The Development of Next Generation Sequencing Tool for Annotation and Variant Filtering of Human Nucleotide Alteration	Ramathibodi Hospital	200,800 THB
06/2013-05/2015	Principal Investigator	Predicting where Humans Look at in Images by Machine Learning Techniques	Thailand Research Fund TRG5680090	480,000 THB
10/2012-09/2013	Principal Investigator	The Development of Online Submission System for Computer Programming Teaching Tool	IT.KMITL 2556-0206002	50,000 THB
01/2012-05/2012	Principal Investigator	The Development of Submission System for Computer Programming Teaching Tool	IT.KMITL	20,000 THB
10/2011-09/2012	Principal Investigator	Dimensionality Reduction for Data Mining	IT.KMITL 2555-0206003	50,000 THB
06/2008-05/2010	Researcher	Personal Information Navigator Adapting Through Viewing (PinView)	EU FP7 216529	668,558.58 EUR

ผลงานวิจัย

Books (1)

1. Kitsuchart Pasupa, “Introduction to Machine Learning (In Thai)”, Mean Service Supply, Bangkok, 2017.

National Journal Manuscripts (5)

1. Kitsuchart Pasupa, Sanparith Marukatat, Thavida Maneewarn, “The Development of Artificial Intelligence, Robotics, Big Data Analytics in Thailand. (In Thai)”, In TCEL White Paper - Ethical Perspective on Science, Technology: Artificial Intelligence, Robotics, Big Data, pp. 1-20, accepted.

2. Kitsuchart Pasupa, Panawee Chatkamjuncharoen, Chotiros Wuttilertdeshar, “Prediction of Human Emotions toward Abstract Images by Image Features and Eye Tracking Device (In Thai)”, In Jounal of Information Science and Technology, vol. 5, no. 2, pp. 1-8, 2015.

3. Suthasinee Nopparit, Natapon Pantuwong, Kitsuchart Pasupa, “Behavioural Analysis in Dairy Cow from Video. (In Thai)”, In KMITL Journal of Information Technology, vol. 3, no. 2, pp. 51-58, 2014.

4. Kitsuchart Pasupa, “The Review of Virtual Screening Techniques”, In KMITL Journal of Information Technology, vol. 1, no. 1, pp. 60-82, 2012.

5. Pattheera Janiam, Kitsuchart Pasupa, “The Possibilities of Broadcasting TV Shows in 3D. (In Thai)”, In KMITL Journal of Information Technology, vol. 1, no. 1, pp. 14-23, 2012.

International Journal Manuscripts (14)

1. Wanthanee Rathasamuth, Kitsuchart Pasupa, Sissades Tongsima, “Selection of a Minimal Number of Significant Porcine SNPs by an Information Gain and Genetic Algorithm Hybrid”, In Malaysian Journal of Computer Science, vol. , pp. XX-XX, accepted.

2. Ritthikrai Thaweecharoen, Warut Chaiwong, Chulathip Boonma, Kitsuchart Pasupa, Piyawan Massa-Ard, Chaiyos Kunanusont, “Diagnosis of Metabolic Syndrom using Radar Chart”, In Bangkok Medical Journal, vol. 15, no. 1, pp. 11-18, 2019.

3. Kitsuchart Pasupa, Wisuwat Sunhem, Chu Kiong Loo, “A Hybrid Approach to Build Face Shape Classifier for Hairstyle Recommender System”, In Expert Systems With Applications, vol. 120, pp. 14-32, 2019.

4. Habeebah Adamu Kakudi, Chu Kiong Loo, Foong Ming Moy, Naoki Masuyama, Kitsuchart Pasupa, "Diagnosing Metabolic Syndrome using Genetically Optimised Bayesian ARTMAP", In IEEE Access, vol. 7, pp. 8437-8453, 2019.

5. Ungsumalee Suttapakti, Kitsuchart Pasupa, Kuntpong Woraratpanya, "Empirical Monocomponent Image Decomposition", In IEEE Access, vol. 6, pp. 38706-38735, 2018.

6. Kitsuchart Pasupa, Wasu Kudisthalert, "Virtual Screening by a Novel Clustering-Based Weighted Similarity Extreme Learning Machine Approach", In PLOS One, vol. 13, no. 4, pp. e0195478, 2018.

7. Zongying Liu, Chu Kiong Loo, Naoki Masuyama, Kitsuchart Pasupa, "Recurrent Kernel Extreme Reservoir Machine for Time Series Prediction", In IEEE Access, vol. 6, pp. 19583-19596, 2018.

8. Kitsuchart Pasupa, Sandor Szedmak, "Utilising Kronnecker Decomposition and Tensor-based Multi-view Learning to Predict Where People are Looking in Images", In Neurocomputing, vol. 248, pp. 80-93, 2017.

9. Ponruedee Netisopakul, Kitsuchart Pasupa, Ratthawut Lertsuksakda, "Hypothesis Testing based on Observation from Thai Sentiment Classification", In Artificial Life and Robotics, vol. 22, no. 2, pp. 184-190, 2017.

10. Kitsuchart Pasupa, Ponruedee Netisopakul, Ratthawut Lertsuksakda, "Sentiment Analysis on Thai Children Stories", In Artificial Life and Robotics, vol. 21, no. 3, pp. 357-364, 2016.

11. Kitsuchart Pasupa, Natapon Pantuwong, Suthasinee Nopparit, "A Comparative Study of Automatic Dairy Cow Detection Using Image Processing Techniques", In Artificial Life and Robotics, vol. 20, no. 4, pp. 320-326, 2015.

12. Robert F. Harrison, Kitsuchart Pasupa, "A Simple Iterative Algorithm for Parsimonious Binary Kernel Fisher Discrimination", In Pattern Analysis & Applications, vol. 13, no. 1, pp. 15-22, 2010.

13. Robert F. Harrison, Kitsuchart Pasupa, "Sparse Multinomial Kernel Discriminant Analysis (sMKDA)", In Pattern Recognition, vol. 42, no. 9, pp. 1795-1802, 2009.

14. Beining Chen, Robert F. Harrison, Kitsuchart Pasupa, Peter Willett, David J. Wilton, David J. Wood, Xiao Qing Lewell, "Virtual Screening Using Binary Kernel Discrimination: Effect of Noisy Training Data and the Optimization of Performance", In Journal of Chemical Information and Modeling, vol. 46, no. 2, pp. 478-486, 2006.

National Conference Proceeding Manuscripts (7)

1. Wisuwat Sunhem, Kitsuchart Pasupa, Piyakorn Jansiripitikul, "Hairstyle Recommendation System for Women (In Thai)", In Proceeding of the 12th National Conference on Computing and Information Technology (NCCIT 2016), 7-8 July 2016, Khon Kaen, Thailand, pp. 1-6, 2016.

2. Trust Rukkunnatham, Chodok Posew, Kitsuchart Pasupa, "Driving Simulator System (In Thai)", In Proceeding of the 7th National Conference on Information Technology (NCIT 2015), 29-30 Oct 2015, Chiang Mai, Thailand, pp. 57-62, 2015.

3. Kitsuchart Pasupa, Panawee Chatkamjuncharoen, Chotiros Wuttilertdeshar, "Prediction of Human Emotions by Image Features and Eye Movements (In Thai)", In Proceeding of the 7th National Conference on Information Technology (NCIT 2015), 29-30 Oct 2015, Chiang Mai, Thailand, pp. 298-303, 2015.

4. Sarawuth Rungcharoenkit, Kitsuchart Pasupa, "The Development of Automatic Programming Exercise Verification System (In Thai)", In Proceeding of the 10th National Conference on Computing and Information Technology (NCCIT 2014), 8-9 May 2014, Phuket, Thailand, pp. 258-263, 2014.

5. Kittaboon Panjarattankorn, Ubolwan Chaovanakij, Kitsuchart Pasupa, "The Development of Next Generation Sequencing Tool for Annotation and Variant Filtrating of Human Nucleotide Alteration (In Thai)", In Proceeding of the 10th National Conference on Computing and Information Technology (NCCIT 2014), 8-9 May 2014, Phuket, Thailand, pp. 115-120, 2014.

6. Kittipun Khantitrirat, Chayaphol Prapaipornlert, Kitsuchart Pasupa, "The Development of Music Information Retrieval System on iOS (In Thai)", In Proceeding of the 10th National Conference on Computing and Information Technology (NCCIT 2014), 8-9 May 2014, Phuket, Thailand, pp. 683-688, 2014.

7. Krit Thangchoeywilai, Kitsuchart Pasupa, "Wonderland of Laddie: An Adventure Game on Windows 8 and Windows RT (In Thai)", In Proceeding of the 5th Conference on Application Research and Development (ECTI-CARD 2013), 8-10 May 2013, Nakornratchasima, Thailand, pp. 127-132, 2013.

International Conference Proceeding Manuscripts (55)

1. Boonyarith Piriyothinkul, Kitsuchart Pasupa, Masanori Sugimoto, "Detecting Text in Manga using Stroke Width", In Proceeding of the 11th International Conference on Knowledge and Smart Technology (KST 2019), 23-26 January 2019, Phuket, Thailand, pp. 142-147, 2019.

2. Supawit Vatathanavaro, Suchat Tungjitnob, Kitsuchart Pasupa, “White Blood Cell Classification: A Comparison between VGG16 and ResNet50 Model”, In Proceeding of the 6th Joint Symposium on Computational Intelligence (JSCI6), 12 December 2018, Bangkok, Thailand, pp. 4-5, 2018.

3. Ek Thamwiwatthana, Kitsuchart Pasupa, Sissades Tongsima, “Selection of SNP Subsets for Severity of Beta-thalassaemia Classification Problem”, In Proceeding of the 9th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2018), 10-13 December 2018, Bangkok, Thailand, pp. 1-7, 2018.

4. Thititorn Seneewong Na Ayutthaya, Kitsuchart Pasupa, “Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features”, In Proceeding of the 13th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2018), 15-17 November 2018, Pattaya, Thailand, pp. 84-89, 2018.

5. Wanthanee Rathasamuth, Kitsuchart Pasupa, Sissades Tongsima, “SNP selection for Porcine breed classification by a hybrid information gain and genetic algorithm”, In Proceeding of the 4th Joint Symposium on Computational Intelligence (JSCI2018), 2 February 2018, Bangkok, Thailand, pp. 10-11, 2018.

6. Boonyarith Piriyothinkul, Patcharapon Joksamut, Kitsuchart Pasupa, “The Disaster Victim Locating System with a Smartphone Without a Telecommunication System”, In Proceeding of the 15th International Joint Conference on Computer Science and Software Engineering (JCSSE2018), 11-13 July 2018, Nakhon Pathom, Thailand, pp. 327-332, 2018.

7. Thanawat Lodkaew, Weeruhputt Supsohmboon, Kitsuchart Pasupa, Chu Kiong Loo, “Fashion Finder: A System for Locating Online Stores on Instagram from Product Images”, In Proceeding of the 10th International Conference on Information Technology and Electrical Engineering (ICITEE 2018), 24-26 July 2018, Bali, Indonesia, pp. 288-293, 2018.

8. Zongying Liu, Chu Kiong Loo, Kitsuchart Pasupa, “Handling Concept Drift in Time-series Data: Meta-cognitive Recurrent Recursive-Kernel OS-ELM”, In Proceeding of the 24nd International Conference on Neural Information Processing (ICONIP 2018), 14-16 Dec 2018, Siem Reap, Cambodia (Long Cheng, Andrew C.S. Leung, Seiichi Ozawa, eds.), vol. 11306, pp. 3-13, 2018.

9. Chanawee Chavaltada, Kitsuchart Pasupa, David R. Hardoon, “Combining Multiple Features for Product Categorisation by Multiple Kernel Learning”, In Proceeding of the 14th International Conference on Computing and Information Technology (IC2IT2018), 5-6 July 2018, Chiang Mai, Thailand, pp. 3-12, 2018.

10. Kittipop Peuwnuan, Kuntpong Woraratpanya, Kitsuchart Pasupa, Yoshimitsu Kuroki, “Local Variance Image-Based for Scene Text Binarization under Illumination Effects”, In Proceeding of the 2nd International Conference on Image, Vision and Computing (ICIVC 2017), 2-4 June 2017, Chengdu, China, pp. 798-802, 2017.

11. Kitsuchart Pasupa, Wisuwat Sunhem, Chu Kiong Loo, Yoshimitsu Kuroki, “Can Eye Movement Information Improve Prediction Performance of Human Emotional Response to Images?”, In Proceeding of the 23nd International Conference on Neural Information Processing (ICONIP 2017), 14-18 Nov 2017, Guangzhou, China (Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, El-Sayed M. El-Alfy, eds.), vol. 10637, pp. 830-838, 2017.

12. ZongYing Liu, Chu Kiong Loo, Kitsuchart Pasupa, “Recurrent Kernel Online Sequential Extreme Learning Machine with Kernel Adaptive Filtering for Time Series Prediction”, In Proceeding of the 10th IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2017), 27 Nov-1 Dec 2017, Honolulu, Hawaii, USA, pp. 1447-1453, 2017.

13. ZongYing Liu, Chu Kiong Loo, Naoki Masuyama, Kitsuchart Pasupa, “Multiple Steps Time Series Prediction by A Novel Recurrent Kernel Extreme Learning Machine Approach”, In Proceeding of the 9th International Conference on Information Technology and Electrical Engineering (ICITEE 2017), 12-13 October 2017, Phuket, Thailand, pp. SIG5.5, 2017.

14. Habeebah Adamu Kakudi, Chu Kiong Loo, Kitsuchart Pasupa, “Risk Quantification of Metabolic Syndrome with Quantum Particle Swarm Optimisation”, In Proceeding of the 26th International Conference on World Wide Web (WWW 2017) - Companion, 3-7 April 2017, Perth, Australia, pp. 1141-1147, 2017.

15. Chanawee Chavaltada, Kitsuchart Pasupa, David R. Hardoon, “A Comparative Study of Machine Learning Techniques in Automatic Product Categorisation”, In Proceeding of the 14th International Symposium on Neural Networks (ISNN 2017), 21-23 June 2017, Hokkaido, Japan (Fengyu Cong, Andrew Leung, Qinglai Wei, eds.), vol. 10261, pp. 10-17, 2017.

16. Wisuwat Sunhem, Kitsuchart Pasupa, Piyakorn Jansiripitikul, “Hairstyle Recommendation System for Women”, In Proceeding of the 5th ICT International Student Project Conference (ICT-ISPC 2016), 27-28 May 2016, Nakhon Pathom, Thailand, pp. 166-169, 2016.

17. Wisuwat Sunhem, Kitsuchart Pasupa, “An Approach to Face Shape Classification for Hairstyle Recommendation System”, In Proceeding of the 8th International Conference on Advanced Computational Intelligence (ICACI2016), 14-16 February 2016, Chiang Mai, Thailand, pp. 390-394, 2016.

18. Kittipop Peuwnuan, Kuntpong Woraratpanya, Kitsuchart Pasupa, “Modified Adaptive Thresholding Using Integral Image”, In Proceeding of the 13th International Joint Conference on Computer Science and Software Engineering (JCSSE 2016), 13-15 July 2016, Khon Kaen, Thailand, pp. 1-5, 2016.

19. Kitsuchart Pasupa, Wisuwat Sunhem, “A Comparison between Shallow and Deep Architecture Classifiers on Small Dataset”, In Proceeding of the 8th International Conference on Information Technology and Electrical Engineering (ICITEE 2016), 5-6 October 2016, Yogyakarta, Indonesia, pp. 390-395, 2016.

20. Kitsuchart Pasupa, Siripen Jungjareantrat, “Water Levels Forecast In Thailand: A Case Study Of Chao Phraya River”, In Proceeding of the 14th International Conference on Control, Automation, Robotics and Vision (ICARCV 2016), 13-15 November 2016, Phuket, Thailand, pp. 1-6, 2016.

21. Kitsuchart Pasupa, Panawee Chatkamjuncharoen, Chotiros Wuttilertdeshar, Masanori Sugimoto, “Using Image features and Eye Tracking Device to Predict Human Emotions Toward Abstract Images”, In Proceeding of the 7th Pacific Rim Symposium on Image and Video Technology (PSIVT 2015), 23-27 Nov 2015, Auckland, New Zealand (Thomas Bräunl, Brendan McCane, Mariano Rivers, Xinguo Yu, eds.), vol. 9431, pp. 419–-430, 2016.

22. Ponrudee Netisopakul, Ratthawut Lertsuksakda, Kitsuchart Pasupa, “Hypothesis Testing based on Observation from Thai Sentiment Classification”, In Proceeding of the 21th International Symposium on Artificial Life and Robotics (AROB 2016), 20-22 Jan 2016, Beppu, Japan, pp. 8-13, 2016.

23. Wasu Kudisthalert, Kitsuchart Pasupa, “Clustering-based Weighted Extreme Learning Machine for Classification in Drug Discovery Process”, In Proceeding of the 23nd International Conference on Neural Information Processing (ICONIP 2016), 16-21 Oct 2016, Kyoto, Japan (Akira Hirose, Seiichi Ozawa, Kenji Doya, Kazushi Ikeda, Minho Lee, Derong Liu, eds.), vol. 9948, pp. 441-450, 2016.

24. Wasu Kudisthalert, Kitsuchart Pasupa, “A Coefficient Comparison of Weighted Similarity Extreme Learning Machine for Drug Screening”, In Proceeding of the 8th International Conference on Knowledge and Smart Technology (KST 2016), 3-6 February 2016, Chiang Mai, Thailand, pp. 43-48, 2016.

25. Phassarun Iamamphai, Jeerana Noymanee, Wimol San-Um, Kitsuchart Pasupa, “Investigations and Comparisons of Government Open Data Websites through Systematic Functional Analysis and Efficient Promotion Approach”, In Proceeding of the 3rd Management and Innovation Technology International Conference (MITiCON 2016), 12-14 October 2016, Bang-Saen, Thailand, pp. 142-147, 2016.

26. Sarawut Bussadee, Sittipong Suwannatria, Arnon Chonrawut, Ek Thamwiwatthana, Kitsuchart Pasupa, “Inside Me: A Proposal Healthcare Mobile Application”, In Proceeding of the 5th ICT International Student Project

Conference (ICT-ISPC 2016), 27-28 May 2016, Nakhon Pathom, Thailand, pp. 85-88, 2016.

27. Syukron Abu Ishaq Alfarozi, Kuntpong Woraratpanya, Kitsuchart Pasupa, "Hinge Loss Projection for Classification", In Proceeding of the 23nd International Conference on Neural Information Processing (ICONIP 2016), 16-21 Oct 2016, Kyoto, Japan (Akira Hirose, Seiichi Ozawa, Kenji Doya, Kazushi Ikeda, Minho Lee, Derong Liu, eds.), vol. 9948, pp. 250-258, 2016.

28. Syukron Abu Ishaq Alfarozi, Noor Akhmad Setiawan, Teguh Bharata Adji, Kuntpong Woraratpanya, Kitsuchart Pasupa, "Analytical Incremental Learning: Fast Constructive Learning Method for Neural Network", In Proceeding of the 23nd International Conference on Neural Information Processing (ICONIP 2016), 16-21 Oct 2016, Kyoto, Japan (Akira Hirose, Seiichi Ozawa, Kenji Doya, Kazushi Ikeda, Minho Lee, Derong Liu, eds.), vol. 9948, pp. 259-268, 2016.

29. Panupon Usachokcharoen, Yoshikazu Washizawa, Kitsuchart Pasupa, "Sign Language Recognition with Depth and Colour Sensor by Using Microsoft Kinect", In Proceeding of the International Conference on Signal and Image Processing Applications (ICSIPA 2015), 19-21 Oct 2015, Kuala Lumpur, Malaysia, pp. 186-190, 2015.

30. Nutchaphon Rewik, Kittipop Peuwnuan, Kuntpong Woraratpanya, Kitsuchart Pasupa, "Particle-Flow Interactive Animation for Painting Image", In Proceeding of the 7th International Conference on Information Technology and Electrical Engineering (ICITEE 2015), 29-30 Oct 2015, Chiang Mai, Thailand, pp. 237-240, 2015.

31. Kitsuchart Pasupa, Sandor Szedmak, "Learning to Predict Where People Look with Tensor-based Multiview Learning", In Proceeding of the 22nd International Conference on Neural Information Processing (ICONIP 2015), 9-12 Nov 2015, Istanbul, Turkey (Sabri Arik, Tingwen Huang, Weng Kin Lai, Qingshan Liu, eds.), vol. 9489, pp. 432–-441, 2015.

32. Suthasinee Nopparit, Natapon Pantuwong, Kitsuchart Pasupa, "A Comparative Study of Automatic Dairy Cow Detection Using Image Processing Techniques", In Proceeding of the 20th International Symposium on Artificial Life and Robotics (AROB 2015), 21-23 Jan 2015, Beppu, Japan, pp. 445-450, 2015.

33. Ratthawut Lertsuksakda, Kitsuchart Pasupa, Ponruedee Netisopakul, "Sentiment Analysis on Thai Children Stories with Support Vector Machine", In Proceeding of the 20th International Symposium on Artificial Life and Robotics (AROB 2015), 21-23 Jan 2015, Beppu, Japan, pp. 138-142, 2015.

34. Sittipong Apichartstaporn, Kitsuchart Pasupa, Yoshikazu Washizawa, "Vibrotactile Brain-Computer Interface with Error-Detecting Codes", In Proceeding of the 5th International Conference on Cognitive Neurodynamics (ICCN 2015), 3-7 Jun 2015, Sanya, China (Rubin Wang, Xiaochuan Pan, eds.), vol. 5, pp. 355-361, 2015.

35. Ungsumalee Suttapakti, Kuntpong Woraratpanya, Kitsuchart Pasupa, Pimlak Boonchukusol, Taravichet Titijaroonroj, Rattaphon Hokking, Yoshimitsu Kuroki, Yasushi Kato, "Text-Background Decomposition for Thai Text Localization and Recognition in Natural Scenes", In Proceeding of the 6th International Conference on Information Technology and Electrical Engineering (ICITEE 2014), 7-8 Oct 2014, Yogyakarta, Indonesia, pp. 138-143, 2014.

36. Suthasinee Nopparit, Natapon Pantuwong, Kitsuchart Pasupa, "Automatic Detection Algorithm of Dairy Cows in Freestall Barns using Feature Points Matching", In Proceeding of the 2014 Regional Conference on Computer and Information Engineering (RCCIE 2014), 7-8 Oct 2014, Yogyakarta, Indonesia, pp. 70-74, 2014.

37. Ratthawut Lertsuksakda, Ponrudee Netisopakul, Kitsuchart Pasupa, "Thai Sentiment Terms Construction using the Hourglass of Emotions", In Proceeding of the 6th International Conference on Knowledge and Smart Technology (KST 2014), 30-31 January 2014, Chonburi, Thailand, pp. 46-50, 2014.

38. Krit Thangchoeywilai, Kitsuchart Pasupa, "Wonderland of Laddie: An Adventure Game on Windows 8 and Windows RT", In Proceeding of the 2nd ICT International Senior Project Conference (ICT-ISPC 2013), 28-29 March 2013, Nakhon Pathom, Thailand, pp. 90-93, 2013.

39. Ungsumalee Suttapakti, Kuntpong Woraratpanya, Kitsuchart Pasupa, "Font Descriptor Construction for Printed Thai Character Recognition", In Proceeding of the 13th IAPR Conference on Machine Vision Applications (MVA 2013), 21-23 May 2013, Kyoto, Japan, pp. 45-48, 2013.

40. Kitsuchart Pasupa, Ek Thamwiwatthana, "Prediction of Reference Evapotranspiration with Missing Data in Thailand", In Proceeding of the 5th International Conference on Information Technology and Electrical Engineering (ICITEE 2013), 7-8 October 2013, Yogyakarta, Indonesia, pp. 181-186, 2013.

41. Kitsuchart Pasupa, Zakria Hussain, John Shawe-Taylor, Peter Willett, "Drug Screening with Elastic-Net Multiple Kernel Learning", In Proceeding of the 13th IEEE International Conference on BioInformatics and BioEngineering (BIBE 2013), 10-13 November 2013, Chania, Greece, pp. 1-5, 2013.

42. Kitsuchart Pasupa, "A Comparison of Dimensionality Reduction Techniques in Virtual Screening", In Proceeding of the 12th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2013), Part II, 9-13 June 2013, Zakopane, Poland (Leszek Rutkowski, Marcin Korytkowski, Rafal Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, Jacek M. Zurada, eds.), vol. 7895, pp. 297-308, 2013.

43. Kitsuchart Pasupa, "Sparse Fisher Discriminant Analysis with Jeffrey's Hyperprior", In Proceeding of the 1st International Conference on Control, Automation & Information Sciences (ICCAIS 2012), 26-29 November

2012, Ho Chi Minh City, Vietnam, pp. 36-41, 2012.

44. Kitsuchart Pasupa, "Prediction by Posterior Estimation in Virtual Screening", In Proceeding of the 2nd International Conference on Engineering, Applied Sciences, and Technology (ICEAST 2012), 21-24 November 2012, Bangkok, Thailand, pp. 165-170, 2012.

45. Kitsuchart Pasupa, Ponrudee Netisopakul, "Thai Paragraph Shortening Based on Binary Classification Model", In Proceeding of the Joint International Symposium on Natural Language Processing and Agricultural Ontology Service (SNLP-AOS 2011), 9-10 February 2012, Bangkok, Thailand, pp. 181-185, 2011.

46. Zakria Hussain, Kitsuchart Pasupa, John Shawe-Taylor, "Learning Relevant Eye Movement Feature Spaces Across Users", In Proceeding of the 6th Biennial Symposium on Eye Tracking Research & Applications (ETRA 2010), 22-24 March 2010, Austin, USA (Carlos Hitoshi Morimoto, Howell O. Istance, Aulikki Hyrskykari, Qiang Ji, eds.), pp. 181-185, 2010.

47. Zakria Hussain, Alex Po Leung, Kitsuchart Pasupa, David R. Hardoon, Peter Auer, John Shawe-Taylor, "Exploration-Exploitation of Eye Movement Enriched Multiple Feature Spaces for Content-Based Image Retrieval", In Proceeding of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2010), Part I, 20-24 September 2010, Barcelona, Spain (José L. Balcázar, Francesco Bonchi, Aristides Gionis, Michèle Sebag, eds.), vol. 6321, pp. 554-569, 2010.

48. David R. Hardoon, Kitsuchart Pasupa, John Shawe-Taylor, "Image Ranking with Implicit Feedback from Eye Movements", In Proceeding of the 6th Biennial Symposium on Eye Tracking Research & Applications (ETRA 2010), 22-24 March 2010, Austin, USA (Carlos Hitoshi Morimoto, Howell O. Istance, Aulikki Hyrskykari, Qiang Ji, eds.), pp. 291-298, 2010.

49. Peter Auer, Zakria Hussain, Samuel Kaski, Arto Klami, Jussi Kujala, Jorma Laaksonen, Alex Po Leung, Kitsuchart Pasupa, John Shawe-Taylor, "Pinview: Implicit Feedback in Content-Based Image Retrieval", In Proceeding of the Workshop on Applications of Pattern Analysis (WAPA 2010), 1-2 September 2010, Cumberland Lodge, UK (Tom Diethe, Nello Cristianini, John Shawe-Taylor, eds.), vol. 11, pp. 51-57, 2010.

50. Peter Auer, Zakria Hussain, Samuel Kaski, Arto Klami, Jussi Kujala, Jorma Laaksonen, Alex Po Leung, Kitsuchart Pasupa, John Shawe-Taylor, "Pinview: Implicit Feedback in Content-Based Image Retrieval", In Proceeding of the International Conference on Machine Learning (ICML 2010) Workshop on Reinforcement Learning and Search in Very Large Spaces, 25 June 2010, Haifa, Israel, pp. 1-4, 2010.

51. Kitsuchart Pasupa, Sandor Szedmak, David R. Hardoon, “Image Ranking with Eye Movements”, In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009) Workshop on Advance in Rankings, 11 December 2009, Whistler, Canada (Shivani Agarwal, Chris Burges, Koby Crammer, eds.), pp. 37-42, 2009.

52. Kitsuchart Pasupa, Craig J. Saunders, Sandor Szedmak, Arto Klami, Samuel Kaski, Steve R. Gunn, “Learning to Rank Images from Eye Movements”, In Proceeding of the IEEE 12th International Conference on Computer Vision (ICCV 2009) Workshops on Human-Computer Interaction (HCI 2009), 27 September-4 October 2009, Kyoto, Japan, pp. 2009-2016, 2009.

53. Kitsuchart Pasupa, Robert F. Harrison, Peter Willett, “Parsimonious Kernel Fisher Discrimination”, In Proceeding of the 3rd Iberian Conference on Pattern Recognition and Image Analysis, Part I, 6-8 June 2007, Girona, Spain (Joan Martí, José-Miguel Benedí, Ana Maria Mendonça, Joan Serrat, eds.), vol. 4477, pp. 531-538, 2007.

54. Komsan Hongesombut, Yasunori Mitani, Sanchai Dechanupaprittha, Issarachai Ngamroo, Kitsuchart Pasupa, Jarurote Tippayachai, “Power System Stabilizer Tuning Based on Multiobjective Design Using Hierarchical and Parallel Micro Genetic Algorithm”, In Proceedings of the International Conference on Power System Technology (POWERCON 2004), 21-24 November 2004, Singapore, pp. 402-407, 2004.

55. Sanchai Dechanupaprittha, Issarachai Ngamroo, Kitsuchart Pasupa, Jarurote Tippayachai, Komsan Hongesombut, Yasunori Mitani, “New Heuristic-based Design of Robust Power System Stabilizers”, In Proceedings of the International Conference on Power System Technology (POWERCON 2004), 21-24 November 2004, Singapore, pp. 618-623, 2004.

Abstract (2)

1. Kitsuchart Pasupa, Arto Klami, Craig J. Saunders, Teofilo de Campos, Samuel Kaski, “Can relevance of images be inferred from eye movements?”, pp. 50, 2009.

2. Kitsuchart Pasupa, “Prediction by Nonparametric Posterior Estimation in Virtual Screening”, 2007.

Technical Reports (9)

1. Zakria Hussain, Arto Klami, Jussi Kujala, Alex Po Leung, Kitsuchart Pasupa, Peter Auer, Samuel Kaski, Jorma Laaksonen, John Shawe-Taylor, “Pinview: Implicit Feedback in Content-Based Image Retrieval”, 1410.0471, ArXiv, no. 1410.0471, pp. 1-12, 2014.

2. Peter Auer, Steve R. Gunn, Zakria Hussain, Samuel Kaski, Arto Klami, Jussi Kujala, Jorma Laaksonen, Alex Po Leung, Sasan Mahmoodi, Ivan Markovsky, Kitsuchart Pasupa, Amirthalingam Ramanan, John Shawe-Taylor, Sandor Szedmak, "Summary of Personal Information Navigator experiments", Technical report, Pinview Deliverable D.8.5.3, 2010.

3. Antti Ajanki, Zakria Hussain, Jorma Laaksonen, Kitsuchart Pasupa, Alex Po Leung, Teemu Ruokolainen, Rudolf Traunmüller He Zhang, "Prototype framework implementation, stage II", Technical report, Pinview Deliverable D.8.5.2, 2010.

4. Teofilo de Campos, Gabriela Csurka, Florent Perronnin, Julian McAuley, Martin Antenreiter, Ronald Ortner, Peter Auer, Ville Viitaniemi, Jorma Laaksonen, Kitsuchart Pasupa, Craig J. Saunders, Zakria Hussain, John Shaew-Taylor, "Description and evaluation of techniques for transfer learning across sub-categories", Technical report, Pinview Deliverable D.6.3, 2009.

5. Kitsuchart Pasupa, Craig J. Saunders, Sandor Szedmak, Steve R. Gunn, David R. Hardoon, Arto Klami, Samuel Kaski, Alex Po Leung, Peter Auer, "Ranking algorithms for implicit feedback", Technical report, Pinview Deliverable D.5.1, 2009.

6. Arto Klami, Samuel Kaski, Kitsuchart Pasupa, Sandor Szedmak, Steve R. Gunn, David R. Hardoon, Gabriela Csurka, "Predicting relevance of parts of an image", Technical report, Pinview Deliverable D.2.2, 2009.

7. Teofilo de Campos, Gabriela Csurka, Florent Perronnin, Zakria Hussain, John Shaew-Taylor, Kitsuchart Pasupa, Craig J. Saunders, Haider Ali, Martin Antenreiter, Ronald Ortner, Peter Auer, Ville Viitaniemi, Jorma Laaksonen, "Description, analysis and evaluation of confidence estimation procedures for sub-categorisation", Technical report, Pinview Deliverable D.6.2.1, 2008.

8. Arto Klami, Samuel Kaski, Kitsuchart Pasupa, Craig J. Saunders, Teo de Campos, "Prediction of relevance of an image from a scan pattern", Technical report, Pinview Deliverable D.2.1, 2008.

9. Zakria Hussain, John Shawe-Taylor, Craig J. Saunders, Kitsuchart Pasupa, "Basic metric learning", Technical report, Pinview Deliverable D.3.1, 2008.