

แบบจำลองการทำนายความเสี่ยงของโรคหลอดเลือดในสมองด้วยอัลกอริทึม Ant-Miner

จิรัฐติกาล สุตวนิชย์¹ อุทิศ ดวงผาสุก² สฤญชัย ปรีดาวัลย์³

¹โรงพยาบาลอากาศอำนวย จังหวัดสกลนคร

²โรงพยาบาลสมเด็จพระพุทธเลิศหล้า จังหวัดสมุทรสงคราม

³วิทยาลัยการสาธารณสุขสิรินธร จังหวัดชลบุรี คณะสาธารณสุขและสหเวชศาสตร์ สถาบันพระบรมราชชนก

Received: December 19, 2024

Revised: April 10, 2025

Accepted: April 14, 2025

บทคัดย่อ

ปัจจุบันโรคหลอดเลือดในสมองเป็นสาเหตุสำคัญของการเสียชีวิตและสร้างความพิการของคนทั่วโลก ส่งผลกระทบต่อระบบสาธารณสุขอย่างมาก การประเมินความเสี่ยงและวินิจฉัยในระยะเริ่มแรกที่ถูกต้อง มีความสำคัญอย่างมากในการป้องกันโรคและการรักษาโรคที่มีประสิทธิภาพ การศึกษาค้นคว้าครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์ข้อมูลปัจจัยเสี่ยงทางการแพทย์ของโรคหลอดเลือดในสมองจากฐานข้อมูลด้านการแพทย์สากล คัดแยกปัจจัยเสี่ยงของโรคหลอดเลือดในสมองที่เป็นประโยชน์ต่อการวินิจฉัยคัดกรองการเกิดโรค และนำมาสร้างแบบจำลองสำหรับการวินิจฉัยโรคหลอดเลือดในสมองด้วยเทคนิคการทำเหมืองข้อมูล ใช้ข้อมูลจากฐานข้อมูล “Stroke Prediction Dataset” ประกอบด้วย ข้อมูลผู้ป่วย 5,110 ราย มีคุณลักษณะปัจจัยเสี่ยงทางคลินิก 11 ประการ ใช้สถิติการถดถอยโลจิสติกแบบพหุตัวแปรเพื่อระบุคุณลักษณะปัจจัยเสี่ยงที่สำคัญ และใช้อัลกอริทึม Ant-Miner ที่เลียนแบบพฤติกรรมค้นหาเส้นทางการหาอาหารของมดมาพัฒนาสร้างแบบจำลองการวินิจฉัยโรคหลอดเลือดในสมอง ใช้วิธี 5-fold cross-validation เพื่อประเมินประสิทธิภาพของแบบจำลอง และเปรียบเทียบกับเทคนิคการเรียนรู้ของเครื่องแบบอื่นๆ ได้แก่ decision trees และ random forest ผลการศึกษพบว่า คุณลักษณะของปัจจัยเสี่ยงที่สำคัญที่สุดของโรคหลอดเลือดในสมอง ได้แก่ อายุมากกว่า 65 ปี (OR = 3.72, ความเชื่อมั่นร้อยละ 95 = 2.86-4.84) ความดันโลหิตสูง (OR = 2.98, ความเชื่อมั่นร้อยละ 95 = 2.31-3.84) โรคหัวใจ (OR = 2.75, ความเชื่อมั่นร้อยละ 95 = 2.07-3.66) และระดับน้ำตาลในเลือดสูง (OR = 2.45, ความเชื่อมั่นร้อยละ 95 = 1.89-3.17 สำหรับ >150 mg/dL) แบบจำลองการวินิจฉัยโรคหลอดเลือดในสมองที่พัฒนาด้วยอัลกอริทึม Ant-Miner แสดงประสิทธิภาพที่โดดเด่นในการจัดการข้อมูลทางการแพทย์ที่ไม่สมดุล โดยสร้างกฎทางคลินิกที่ตีความได้ง่าย ซึ่งมีคุณค่าต่อการตัดสินใจทางคลินิก แบบจำลองนี้มีความถูกต้องร้อยละ 98.24 ความแม่นยำร้อยละ 92.00 ความไวร้อยละ 90.20 และค่า F1-score ร้อยละ 91.09 กฎที่อธิบายได้ชัดเจนเหล่านี้ เช่น “หากอายุ >65 ปี มีระดับน้ำตาลเฉลี่ย >150 mg/dL มีความดันโลหิตสูงจะมีความเสี่ยงต่อโรคหลอดเลือดสมอง 27.12 เท่า” สามารถนำไปประยุกต์ใช้ในการวินิจฉัยโรคหลอดเลือดในสมองในสถานการณ์จริงได้โดยสรุป แบบจำลองด้วยอัลกอริทึม Ant-Miner ในรูปของกฎสามารถระบุปัจจัยเสี่ยงและกฎการวินิจฉัยอย่างง่ายที่ช่วยในการตรวจพบและป้องกันโรคหลอดเลือดในสมองในระยะเริ่มต้น เป็นส่วนสนับสนุนการตัดสินใจทางคลินิกผ่านนวัตกรรมการทำเหมืองข้อมูลนี้ ควรมีการค้นคว้าเพิ่มเติมเพื่อเพิ่มประสิทธิภาพของแบบจำลองในกลุ่มโรคของประชากรที่หลากหลาย นำไปประยุกต์ใช้ทางคลินิกและการสาธารณสุขด้านอื่นๆ การใช้แบบจำลองนี้ร่วมกับระบบบันทึกสุขภาพอิเล็กทรอนิกส์ การพัฒนาแอปพลิเคชันที่ใช้งานง่ายสำหรับบุคลากรทางการแพทย์ อาจเป็นการช่วยเพิ่มประสิทธิภาพในการประเมินและจัดการความเสี่ยงของโรคหลอดเลือดในสมองได้อย่างมีนัยสำคัญ

คำสำคัญ: แบบจำลองการทำนาย; โรคหลอดเลือดในสมอง; อัลกอริทึม Ant-Miner

ผู้นิพนธ์ประสานงาน:

สฤญชัย ปรีดาวัลย์

วิทยาลัยการสาธารณสุขสิรินธร จังหวัดชลบุรี คณะสาธารณสุขและสหเวชศาสตร์ สถาบันพระบรมราชชนก

29 หมู่ 4 ถนนชิรปราการ ตำบลบ้านสวน อำเภอเมือง จังหวัดชลบุรี 20000

อีเมล: saritchai@scphc.ac.th

Predictive model for cerebrovascular disease risk factors using ant-miner algorithm

Jirattikal Suttawanit¹, Uthid Duangphasuk², Saritchai Predawan³

¹Akatumnua Hospital, Sakon Nakhon

²Somdet Phraphutthalertla Hospital, Samut Songkhram

³Sirindhorn College of Public Health Chonburi, Faculty of Public Health and Allied Health Sciences,
Praboromarajchanok Institute

Abstract

Stroke remains a leading cause of mortality and disability worldwide, with a significant impact on public health. Early accurate risk assessment and diagnosis are crucial for effective prevention and treatment strategies. This study aimed to analyze medical data of cerebrovascular disease from an international database, extract beneficial risk factors related to diagnosis, and create an advanced model for stroke diagnosis using data mining techniques. This study was analyzed data from the Stroke Prediction Dataset, comprising 5,110 patient records with 11 clinical attributes. Multivariate logistic regression analysis was employed to identify significant risk factors. The Ant-Miner Algorithm, inspired by ant colony behavior, was utilized to develop a diagnostic model for stroke. The model's performance was evaluated using 5-fold cross-validation and compared with other machine learning techniques, including decision trees and random forest. The results found that identified age over 65 years (OR = 3.72, 95% CI = 2.86-4.84), hypertension (OR = 2.98, 95% CI = 2.31-3.84), heart disease (OR = 2.75, 95% CI = 2.07-3.66), and high blood glucose levels (OR = 2.45, 95% CI = 1.89-3.17 for >150 mg/dL) as the most significant risk factors for stroke. The Ant-Miner Algorithm-based model excelled in handling imbalanced medical data, generating interpretable clinical rules, and valuable for clinical decisions. It achieved 98.24% accuracy, 92.00% precision, 90.20% sensitivity, and 91.09% F1-score. These explainable rules, like "IF (age >65) AND (avg_glucose >150) AND (hypertension = 1) THEN high stroke risk with OR = 27.12", are applicable in real-world stroke diagnostics. In conclusion, the Ant-Miner algorithm model proved to be an accurate and stroke diagnostic model which simply identifies risk factors and diagnostic rules that can aid in early detection and prevention of stroke, potentially improving clinical decision-making through this data mining innovation. Further studies are recommended to validate the model's performance in other diseases of diverse populations and to explore its potential applications in clinical practice and public health interventions. The integration of this model with electronic health records and the development of user-friendly applications for healthcare professionals could significantly enhance stroke risk assessment and management tools.

Keywords: predictive model; cerebrovascular disease; ant-miner algorithm

Corresponding Author:

Saritchai Predawan

Sirindhorn College of Public Health Chonburi, Faculty of Public Health and Allied Health Sciences, Praboromarajchanok Institute

29 Village No.4, Wachiraprakarn Road, Bansuan Sub-district, Muang District, Chonburi 20000, Thailand

E-mail: saritchai@scphc.ac.th

บทนำ

โรคหลอดเลือดในสมอง (cerebrovascular disease, stroke) เป็นปัญหาสำคัญทางสุขภาพที่ส่งผลกระทบต่อประชากรทั่วโลกอย่างมาก พบได้บ่อยที่สุดในโรคระบบประสาท ในปี 2024 โรคนี้ยังคงเป็นสาเหตุหลักของการเสียชีวิตและความพิการในหลายประเทศ สาเหตุโรคหลอดเลือดในสมองแบ่งออกตามลักษณะทางพยาธิวิทยาได้เป็นสองประเภทหลัก ได้แก่ โรคหลอดเลือดในสมองตีบ (ischemic stroke) เกิดจากการอุดตันของหลอดเลือดในสมอง ทำให้เลือดไม่สามารถไปเลี้ยงสมองได้เพียงพอ ประมาณร้อยละ 87 ของโรคหลอดเลือดในสมองทั้งหมดเป็นชนิดนี้¹ และโรคหลอดเลือดในสมองแตก (hemorrhagic stroke) เกิดจากการแตกของหลอดเลือดในสมอง อาจเกิดจากความดันโลหิตสูงหรือการผิดปกติของหลอดเลือดเอง ประมาณร้อยละ 13 อาการต่างๆ มักเกิดขึ้นอย่างรวดเร็ว และคงอยู่นานเกิน 24 ชั่วโมง ความรุนแรงของโรคขึ้นอยู่กับขนาด ตำแหน่งที่เกิดของสมอง และระยะเวลาการเกิดโรค โดยปกติสมองจะทนต่อภาวะขาดเลือดได้ชั่วขณะหนึ่ง แต่สมองจะสูญเสียหน้าที่อย่างถาวรเมื่อเกิดภาวะขาดเลือดนานกว่า 3-6 ชั่วโมง อาการที่อาจพบได้ทันทีและพบบ่อยๆ คือ กล้ามเนื้อบริเวณหน้า แขน และขา ข้างใดข้างหนึ่งมีอาการอ่อนแรง เดินเซ สูญเสียการทรงตัว พูดไม่ชัดเจน พูดสับสนจับใจความไม่ได้มองเห็นภาพได้ไม่ชัด อาจเกิดข้างเดียวหรือทั้งสองข้าง หรือปวดศีรษะอย่างรุนแรงโดยไม่ทราบสาเหตุ อาการของผู้ป่วยอาจกลับมารู้สึกตัวเป็นปกติได้ประมาณ 1-14 วัน แต่อาจพบภาวะแขนขาไม่มีแรง อ่อนปวกเปียกไม่สามารถควบคุมการเคลื่อนไหวได้ เรียกว่า อัมพาตองค์การอนามัยโลก² รายงานว่า โรคหลอดเลือดในสมองมีผู้เสียชีวิตอยู่ใน 10 อันดับแรกจากทั่วโลก และเป็นสาเหตุการตายอันดับที่ 3 ในประเทศสหรัฐอเมริกา โดยจะพบผู้ป่วยเพิ่มขึ้น 7.8 ล้านคนในปี ค.ศ. 2030 สำหรับประเทศไทย พบว่า เพศชาย มีการสูญเสียปีสุขภาวะจากการตายก่อนวัยอันควรประมาณ 9.1 ล้านปี มีสาเหตุหลักจากโรคหัวใจและหลอดเลือดเป็นอันดับแรก ร้อยละ 21 และเพศหญิง มีการสูญเสียปีสุขภาวะจาก

การตายก่อนวัยอันควรประมาณ 5.4 ล้านปี มีสาเหตุหลักจากโรคหัวใจและหลอดเลือดเป็นอันดับที่สอง ร้อยละ 21 เมื่อพิจารณาการตายจำแนกตามรายโรค 175 รายโรค เฉพาะประชากรไทยทุกอายุในปี พ.ศ. 2562 พบว่า เพศชาย มีการตายจากโรคหลอดเลือดในสมองมากที่สุด คิดเป็นร้อยละ 12 ของการตายทั้งหมดในเพศชาย ส่วนเพศหญิงมีการตายจากโรคหลอดเลือดในสมองมากที่สุด เช่นเดียวกับเพศชาย คิดเป็นร้อยละ 14.9 และสาเหตุสามอันดับแรกของการสูญเสียปีสุขภาวะในแต่ละช่วงวัย พบว่า กลุ่มอายุ 30-59 ปี ในเพศชาย ได้แก่ การบาดเจ็บทางถนน โรคหัวใจขาดเลือด และโรคหลอดเลือดในสมอง ส่วนในเพศหญิง ได้แก่ โรคเบาหวาน การบาดเจ็บทางถนน และโรคมะเร็งเต้านม กลุ่มอายุ 60 ปีขึ้นไป ในเพศชาย ได้แก่ โรคหลอดเลือดในสมอง โรคเบาหวาน และโรคหัวใจขาดเลือด ส่วนเพศหญิง ได้แก่ โรคเบาหวาน โรคหลอดเลือดในสมอง และโรคหัวใจขาดเลือด จากรายงานแผนงานพัฒนาดัชนีภาวะโรคแห่งประเทศไทย และมูลนิธิเพื่อการพัฒนาคุณภาพพระหว่างประเทศ³ ได้ร่วมกับหน่วยงานที่เกี่ยวข้องศึกษาประเมินภาระโรคและการบาดเจ็บของประชากรไทย จากการศึกษาและสำรวจครั้งล่าสุด พบว่า คนไทยทั้งประเทศสูญเสียปีสุขภาวะ (disability-adjusted life years) รวมกว่า 18.5 ล้านปี ซึ่งจำนวนปีที่สูญเสียไปอย่างน่าเสียดายนี้เพราะสุขภาพที่ไม่ดี พิกการ หรือเสียชีวิตก่อนวัย โดยกว่าร้อยละ 72 เป็นการสูญเสียจากการตายก่อนวัยอันควรจากกลุ่มโรคไม่ติดต่อเรื้อรัง ได้แก่ การบาดเจ็บทางถนน โรคหลอดเลือดในสมอง และโรคเบาหวาน พบสูงสุดในกลุ่มอายุ 45-59 ปี ทั้งในเพศชายและหญิง ซึ่งเป็นกลุ่มวัยทำงาน โดยในเพศชายสูญเสียจากการบาดเจ็บทางถนนสูงสุด รองลงมา คือ โรคหลอดเลือดในสมอง และโรคเบาหวาน ส่วนในเพศหญิง พบการสูญเสียสูงสุดจากโรคเบาหวาน รองลงมา คือ โรคหลอดเลือดในสมอง และการบาดเจ็บทางถนน นอกจากนี้ ยังพบการสูญเสียปีสุขภาวะจากการติดเชื้อเอชไอวี วัณโรค รวมถึงโรคมะเร็งชนิดต่างๆ อีกด้วย

ปัจจุบัน เทคโนโลยีสารสนเทศและการประมวลผลข้อมูลทางการแพทย์ได้เข้ามามีบทบาทสำคัญในวงการแพทย์ การทำเหมืองข้อมูล (data mining) เป็นทฤษฎีการค้นหาคำตอบจากข้อมูลขนาดใหญ่ในรูปแบบ (knowledge discovery in database, KDD) มีบทบาทสำคัญในการวิเคราะห์ข้อมูลทางการแพทย์ เช่น ข้อมูลการรักษา ข้อมูลการวินิจฉัยโรค และข้อมูลการตรวจวินิจฉัยทางคลินิก ข้อมูลเหล่านี้สามารถนำมาใช้ในการปรับปรุงการรักษาและการวินิจฉัยโรคให้มีความแม่นยำและมีประสิทธิภาพมากขึ้น ตัวอย่างเช่น การทำเหมืองข้อมูลสามารถใช้ในการวิเคราะห์ข้อมูลผู้ป่วยเพื่อตรวจจับแนวโน้มการเกิดโรคและพัฒนาวิธีการรักษาใหม่ๆ⁴ Ant-miner algorithm เป็นหนึ่งในเทคนิคการขุดค้นข้อมูล (data mining) ที่พัฒนามาจากทฤษฎีการหาเส้นทางของมด (ant colony optimization, ACO) จากพฤติกรรมกรรมการหาอาหารของมดในธรรมชาติ อัลกอริทึมนี้จะใช้กลุ่มมดเสมือนในการสำรวจและค้นหาข้อมูลเพื่อสร้างกฎในการจำแนกประเภทของข้อมูล⁵ เช่น ข้อมูลทางการแพทย์จากแหล่งต่างๆ เช่น การตรวจเลือด การถ่ายภาพทางการแพทย์ และอาการทางคลินิก เป็นต้น การประเมินความเสี่ยงเริ่มแรกของโรคหลอดเลือดในสมองนี้ต้องอาศัยความรู้จากหลายด้าน ทั้งการตรวจพบและการพยากรณ์โรคที่แม่นยำ การป้องกันโรคที่ถูกต้องเหมาะสมและการรักษาโรคที่มีประสิทธิภาพ ผู้ศึกษาจึงสนใจการใช้เทคนิคเหมืองข้อมูล Ant-miner algorithm มาประยุกต์ใช้ในการศึกษาวิจัยสำหรับการวินิจฉัยโรคหลอดเลือดในสมอง ใช้ข้อมูลคุณลักษณะด้านสุขภาพและอาการปัจจัยเสี่ยงต่างๆ ทางคลินิก มาสร้างโมเดลสำหรับการวินิจฉัยและทำนายโรคหลอดเลือดในสมองได้อย่างแม่นยำ สามารถตรวจพบรอยโรคได้ในระยะแรกเริ่ม ป้องกันการเกิดโรคและผู้ป่วยได้รับการรักษาอย่างทันท่วงที ลดความพิการและอัตราการเสียชีวิตช่วยเพิ่มประสิทธิภาพในการตรวจคัดกรองและการรักษาโรคนี้ได้

การศึกษานี้มีวัตถุประสงค์เพื่อวิเคราะห์ข้อมูลทางการแพทย์ของโรคหลอดเลือดในสมองจาก

ฐานข้อมูลสากล สกัดข้อมูลปัจจัยเสี่ยงที่เป็นประโยชน์และสัมพันธ์ต่อการวินิจฉัยโรค รวมทั้งสร้างแบบจำลองสำหรับการวินิจฉัยโรคหลอดเลือดในสมอง

วิธีการศึกษา

การศึกษานี้ได้ผ่านการพิจารณาจากจริยธรรมโครงการวิจัยในมนุษย์แล้ว จากคณะกรรมการพิจารณาจริยธรรมการวิจัยในมนุษย์ สำนักงานสาธารณสุขจังหวัดสกลนคร (เลขที่รับรอง SKN REC 2024-097)

จากการทบทวนวรรณกรรมและทฤษฎีสำคัญที่เกี่ยวข้องในการทำเหมืองข้อมูล (data mining) เป็นวิธีการวิเคราะห์หรือสืบค้นรูปแบบ (model) ความรู้และความสัมพันธ์ของข้อมูลอันเป็นประโยชน์ที่หลบซ่อนอยู่ฐานข้อมูลขนาดใหญ่ เครื่องมือที่ใช้ในการศึกษาประกอบด้วย ฐานข้อมูลเปิดสากลเกี่ยวกับประวัติผู้ป่วยเฉพาะโรคหลอดเลือดในสมอง (stroke prediction dataset) ของผู้มาใช้บริการในโรงพยาบาลคลินิกซานการโลส ประเทศสเปน ในปี ค.ศ. 2020 จำนวน 5,110 ระเบียบ มีคุณลักษณะสำคัญทางคลินิกที่ใช้ศึกษา 11 แบบคุณลักษณะ ฐานข้อมูลได้รับอนุญาตให้เข้าใช้ถึงข้อมูลแบบ creative commons เป็นแหล่งข้อมูลสาธารณะทางวิชาการที่สำคัญ สำหรับใช้ในการศึกษาวิจัย หรือเป็นตัวอย่างทดลอง สำหรับฝึกสอนระบบปัญญาประดิษฐ์ (artificial intelligence) และการเรียนรู้ของเครื่อง (machine learning) จากฐานข้อมูลศึกษาที่ <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/download?DatasetVersionNumber=1>⁶ ผู้ศึกษามีกรอบแนวคิด ดังแสดงในรูปที่ 1 (Figure 1) และดำเนินการศึกษาวิจัยด้วยวิธีการทำเหมืองข้อมูล (data mining) เป็นกระบวนการสร้างแบบจำลองด้วยการสกัดข้อมูลที่มีค่าและความหมายจากฐานข้อมูลขนาดใหญ่ (KDD) ผ่านการใช้เทคนิคทางสถิติ และการวิเคราะห์ทางคณิตศาสตร์ ประกอบด้วยขั้นตอนหลักๆ ดังนี้

1. การเก็บรวบรวมข้อมูล (data collection) ขั้นตอนนี้หมายถึงการรวบรวมข้อมูลจากแหล่งต่างๆ

เช่น ฐานข้อมูลทางการแพทย์ เวชระเบียนอิเล็กทรอนิกส์ ข้อมูลการวินิจฉัยทางคลินิก และข้อมูลจากอุปกรณ์ติดตามสัญญาณชีพของผู้ป่วย ข้อมูลที่ได้ต้องมีความ

ครบถ้วนและถูกต้อง การศึกษานี้ใช้ฐานข้อมูลที่รวบรวมไว้ใน <https://www.kaggle.com>⁶ อยู่ในไฟล์ “healthcare-dataset-stroke-data.csv”

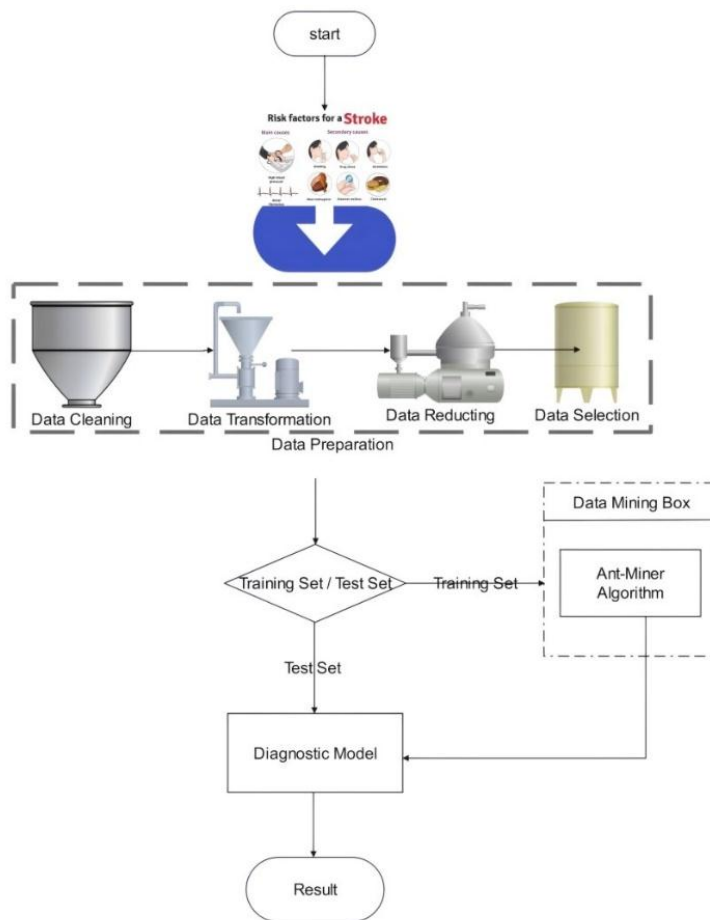


Figure 1 Research Framework and Methodology

2. การเตรียมข้อมูล (data preparation)
การเตรียมข้อมูล (data preparation) เป็นขั้นตอนสำคัญที่ผู้วิจัยดำเนินการเพื่อให้ได้ข้อมูลที่มีคุณภาพดังในตารางที่ 1 (Table 1) สำหรับการวิเคราะห์ในการศึกษานี้ ผู้วิจัยได้ดำเนินการดังนี้

2.1 การคัดเลือกข้อมูล (data selection)
ผู้วิจัยคัดเลือกกระเป๋นข้อมูลที่มีความสมบูรณ์จาก

ฐานข้อมูลเปิดสากล “Stroke Prediction Dataset” โดยพิจารณาจากความครบถ้วนของข้อมูลในแต่ละคุณลักษณะที่สำคัญ คัดเลือกคุณลักษณะที่มีความสัมพันธ์กับโรคหลอดเลือดในสมองตามการทบทวนวรรณกรรมได้คุณลักษณะที่ใช้ในการศึกษาทั้งสิ้น 11 คุณลักษณะ (ยกเว้นรหัสผู้ป่วย-ID ที่ไม่นำมาใช้ในการวิเคราะห์)

2.2 การลดทอนมิติข้อมูล (data reduction) ผู้วิจัยได้วิเคราะห์และจัดกลุ่มตัวแปรต่อเนื่อง เช่น อายุ ระดับน้ำตาลในเลือด และดัชนีมวลกาย เป็นช่วงที่มีความหมายทางคลินิก เพื่อลดความซับซ้อนและเพิ่มความสามารถในการตีความของแบบจำลอง ตัวอย่างเช่น แบ่งกลุ่มอายุเป็น <45 ปี, 45-65 ปี และ >65 ปี ตามระดับความเสี่ยงที่แตกต่างกัน

2.3 การทำความสะอาดข้อมูล (data cleaning) ผู้วิจัยตรวจสอบและจัดการกับค่าที่ขาดหายไป (missing values) โดยเฉพาะในคุณลักษณะ BMI ที่มีค่าเป็น “N/A” และ smoking_status ที่มีค่าเป็น “Unknown” โดยใช้วิธีการประมาณค่าที่เหมาะสม และคัดกรองค่าผิดปกติ (outliers) โดยใช้วิธี Interquartile Range (IQR) สำหรับตัวแปรเชิงปริมาณ

2.4 การแปลงรูปแบบข้อมูล (data transformation) ผู้วิจัยแปลงข้อมูลจากไฟล์ CSV เป็นรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ โดยนำเข้าสู่โปรแกรมวิเคราะห์ทางสถิติ และจัดเตรียมข้อมูลให้พร้อมสำหรับการใช้งานกับอัลกอริทึม Ant-Miner ข้อมูลที่รวบรวมมาอาจมีข้อผิดพลาดหรือไม่สมบูรณ์

ขั้นตอนการเตรียมข้อมูล ประกอบด้วย การทำความสะอาดข้อมูล การแปลงข้อมูล และการคัดเลือกข้อมูลจากการรวมข้อมูลจากแหล่งต่างๆ ให้เป็นข้อมูลที่พร้อมสำหรับการวิเคราะห์ ปัจจัยเสี่ยงในการสร้างแบบจำลองสำหรับวินิจฉัยโรคหลอดเลือดในสมอง มีจำนวน 12 แบบคุณลักษณะ (attribute) ประกอบด้วย รหัสข้อมูล (Id) เพศ (gender) อายุ (age) ความดันโลหิต (hypertension) โรคหัวใจ (heart_disease) สถานะแต่งงาน (ever_married) ประเภทการทำงาน (work_type) ประเภทพลเมือง (residence_type) ระดับน้ำตาลเฉลี่ย (avg_glucose_level) ค่าดัชนีมวลกาย (BMI) สถานะสูบบุหรี่ (smoking_status) และผลลัพธ์การเป็นโรคหลอดเลือดในสมอง (stroke label) ดังในตารางที่ 1 (Table 1) ในการศึกษาครั้งนี้ รหัสข้อมูล (Id) เป็นตัวกำหนดความไม่ซ้ำกันของระเบียนข้อมูล จะไม่นำไปใช้สร้างกฎของการทำเหมืองข้อมูล สำหรับข้อมูลคุณสมบัติผลลัพธ์การเป็นโรคหลอดเลือดในสมอง (stroke label) เป็นผลลัพธ์การวินิจฉัยระเบียนข้อมูลที่ไม่เป็น/เป็นโรคหลอดเลือดในสมอง (label value = “0”, “1”)

Table 1 Data attribute ["healthcare-dataset-stroke-data.csv"]⁶

No.	Attribute name.	Remark note
1	id	unique identifier
2	gender	“Male”, “Female” or “Other”
3	age	age of the patient
4	hypertension	0 if the patient doesn’t have hypertension, 1 if the patient has hypertension
5	heart_disease	0 if the patient doesn’t have any heart diseases, 1 if the patient has a heart disease
6	ever_married	“No” or “Yes”
7	work_type	“Children”, “Govt_jov”, “Never_worked”, “Private” or “Self-employed”
8	Residence_type	“Rural” or “Urban”
9	avg_glucose_level	average glucose level in blood
10	bmi	body mass index
11	smoking_status	“Formerly smoked”, “never smoked”, “smokes” or “Unknown”*
12	Stroke (Label)	1 if the patient had a stroke or 0 if not

*Note “Unknown” in smoking_status means that the information is unavailable for this patient

3. การค้นหาความรู้ และการวิเคราะห์ข้อมูล (knowledge discovery and data analysis)

ค้นหาความรู้ (knowledge discovery) ด้วยเทคนิคการทำเหมืองข้อมูล เช่น การจำแนกประเภท (classification) การจัดกลุ่ม (clustering) และการวิเคราะห์ความสัมพันธ์ (association analysis) เพื่อค้นหาความรู้จากข้อมูล⁸ ซึ่งในการศึกษาครั้งนี้ ผู้ศึกษา

ใช้เทคนิคทางสถิติต่างๆ เพื่อหาความสัมพันธ์ของคุณลักษณะ จากการทำเหมืองข้อมูลด้วย Ant-miner algorithm ค้นหาคุณลักษณะทั้งหมดจากฐานข้อมูลขนาดใหญ่ มาคัดแยกค้นหาคุณลักษณะเด่นที่สำคัญ และเป็นประโยชน์ต่างๆ ตามแบบกฎ IF-Then ในรูป (Term) If<Term1^Term2^Term3^...> and Then <Diagnosis> ดังแสดงในรูปที่ 2 (Figure 2)

```
Training Set = all training cases;
DiscoveredRuleList = []; /*initialized with empty list */
WHILE (No. of uncovered cases in the Training set > max_uncovered_cases)
    l = 0;
    REPEAT
        l=l+1;
        Anti incrementally constructs a classification rule;
        Prune the just constructed rule;
        Update the pheromone of the trail followed by Anti;
    UNTIL (l≥No_of_Ants) or (Anti constructed the same rule as
        the previous No_Rules_Converg-1Ants)
        Select the best rule among all constructed rules;
        Add the chosen rule to DiscoveredRuleList;
        Remove the cases correctly covered by the selected rule from the training set;
    END WHILE
```

Figure 2 Pseudocode of Ant-Miner Algorithm⁵

จากรูปที่ 2 (Figure 2) เป็นขั้นตอนทำงานของ Ant-miner algorithm ประกอบด้วย ขั้นตอนหลักๆ ดังนี้

1. Initialization เริ่มต้นด้วยการสร้างกลุ่มมด และการตั้งค่าพารามิเตอร์เบื้องต้น เช่น จำนวนมด และอัตราการระเหยของฟีโรโมน

2. Construction มดแต่ละตัวจะสำรวจข้อมูล และสร้างเส้นทางหรือกฎในการจำแนกประเภทของข้อมูล

3. Pheromone update ปรับปรุงเส้นทางที่ดีที่สุดโดยการเพิ่มฟีโรโมนบนเส้นทางนั้น ซึ่งจะช่วยให้มดในรอบถัดไปสามารถค้นพบเส้นทางที่ดีขึ้น

4. Convergence ทำซ้ำขั้นตอนการสำรวจ และการปรับปรุงเส้นทางจนกว่าจะได้กฎที่มีความแม่นยำ และประสิทธิภาพสูง

การวิเคราะห์ข้อมูล (data analysis) ด้วยการวิเคราะห์ความถดถอยโลจิสติกแบบพหุตัวแปร (multivariate logistic regression)⁹ เป็นการศึกษาปัจจัยเสี่ยงด้วยหลักการทางสถิติ logistic regression ($\text{logis}(p) = \ln\left(\frac{p}{1-p}\right)$) ที่ใช้สำหรับโมเดลความน่าจะเป็นของเหตุการณ์แบบไบนารี (มี 2 ผลลัพธ์ เช่น ความเสี่ยงสูง หรือ ความเสี่ยงต่ำ) ตามสมการที่ (1) ในการศึกษาครั้งนี้

ข้อมูลตัวแปรอิสระ (independent variable) เช่น อายุ, ความดันโลหิต, ดัชนีมวลกาย, การสูบบุหรี่ เป็นต้น

ตัวแปรตาม (dependent variable) เช่น การเกิด Stroke (0 = ไม่เป็น, 1 = เป็น)

$$\text{logis}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \dots\dots\dots(1)$$

p คือ ความน่าจะเป็นที่จะเกิด Stroke

β_0 คือ ค่าคงที่

$\beta_1, \beta_2, \dots, \beta_n$ คือ ค่าสัมประสิทธิ์ของตัวแปรอิสระ

$X_1, X_2, \dots, X_n$ คือ ตัวแปรอิสระ

Odds Ratio (OR) คือ อัตราส่วนระหว่างโอกาสที่เหตุการณ์จะเกิดขึ้นกับโอกาสที่เหตุการณ์จะไม่เกิดขึ้น
คำนวณจากค่าสัมประสิทธิ์ β ของตัวแปรอิสระ จากสูตร $OR = e^\beta$

ตัวอย่าง เช่น β_1 ของอายุ (age) = 0.85, ค่า $OR = e^{0.85} = 2.34$

Confidence Interval (CI) ให้ช่วงค่าที่คาดว่า OR จะอยู่ภายในช่วงนั้นในระดับความมั่นใจที่กำหนด (เช่น 95%) คำนวณจากค่า OR และค่าความแปรปรวนของ β จากสูตร ตามสมการที่ 2

$$CI = OR \times (e^{(SE \times Z)}, e^{(-SE \times Z)}) \dots\dots\dots (2)$$

SE คือ Standard Error ของ β

Z คือ Z-score สำหรับระดับความเชื่อมั่น (เช่น 1.96 สำหรับ 95% CI)

*หาก CI = 1 แสดงว่าความสัมพันธ์ไม่มีนัยสำคัญทางสถิติที่ระดับความเชื่อมั่น 95%

4. การตรวจสอบแบบจำลองและการทดสอบ (model validation and testing) ในขั้นตอนนี้ Ant-miner algorithm มีการสร้างแบบจำลองโดยใช้เทคนิคทางสถิติและการเรียนรู้ของเครื่องแบบจำลอง (term) ที่ได้ผ่านการทดสอบจะถูกประเมินและปรับปรุงเพื่อให้มีความแม่นยำและมีประสิทธิภาพ

สูงสุด¹⁰ การศึกษานี้ ใช้วิธีการการตรวจสอบแบบไขว้ (cross validation) เพื่อตรวจสอบประสิทธิภาพผลลัพธ์ของกฎในแบบจำลองการวินิจฉัยโรคหลอดเลือด ในสมองด้วยการคาดการณ์ค่าความผิดพลาดของแบบจำลองในส่วนการเรียนรู้ (learning) จากชุดข้อมูลสำหรับฝึกสอน ดังแสดงในรูปที่ 3 (Figure 3)

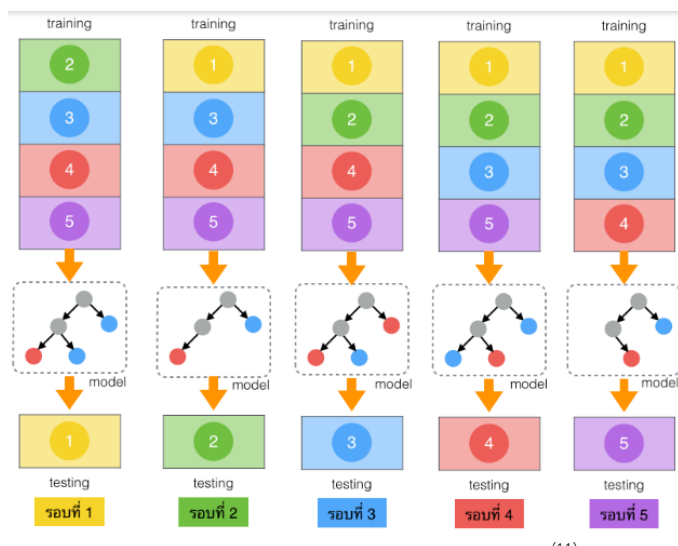


Figure 3 5-Fold Cross Validation Sample¹¹

ผลลัพธ์ของแบบจำลอง คือ กฎ (term) ที่ได้จากการตรวจสอบไขว้ จะนำไปหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน เพื่อเป็นข้อมูลนำเข้าการประเมินประสิทธิภาพของกฎในแบบจำลอง ในการศึกษา ผู้วิจัยประเมินประสิทธิภาพของแบบจำลองด้วยวิธี Confusion Matrix ซึ่งเป็นขั้นตอนมาตรฐานในการประเมินแบบจำลองการจำแนกประเภท โดยพิจารณาค่าต่างๆ ได้แก่ ความถูกต้อง (accuracy) คือสัดส่วนของการทำนายที่ถูกต้องทั้งหมด, ความแม่นยำ (precision) คือสัดส่วนของการทำนายผลบวกที่ถูกต้อง, ความไว (sensitivity หรือ recall) คือสัดส่วนของกรณีบวกจริงที่ถูกทำนายถูกต้อง และ F1-score ซึ่งเป็นค่าเฉลี่ยแบบ harmonic mean ระหว่าง precision และ recall ค่าสูงของมาตรวัดเหล่านี้บ่งชี้ว่าแบบจำลองการวินิจฉัยโรคหลอดเลือดในสมอง (diagnostic decision model) มีประสิทธิภาพหรือแบบจำลองที่ได้จากการทำเหมืองข้อมูลดี

5. นำแบบจำลองจากผลลัพธ์ที่ได้จาก Ant-miner algorithm ไปประยุกต์ใช้เป็นเครื่องมือสำหรับช่วยการวินิจฉัยโรคหลอดเลือดในสมองในรูปแบบของกฎที่เข้าใจง่าย แสดงในรูปกราฟ, รายงาน หรือโปรแกรมแอปพลิเคชันต่างๆ

ผลการศึกษา

การศึกษานี้ใช้ชุดข้อมูล Stroke Prediction Dataset ซึ่งประกอบด้วยข้อมูลผู้ป่วยจำนวน 5,110 ราย ประกอบด้วย คุณลักษณะ gender (เพศ), age (อายุ), hypertension (ความดันโลหิตสูง), heart_disease (โรคหัวใจ), ever_married (เคยแต่งงาน), work_type (ประเภทงาน), residence_type (ประเภทที่อยู่อาศัย), avg_glucose_level (ระดับน้ำตาลในเลือดเฉลี่ย), bmi (ดัชนีมวลกาย), smoking_status (สถานะการสูบบุหรี่), stroke (การเกิดโรคหลอดเลือดในสมอง - ตัวแปรเป้าหมาย) ทำการวิเคราะห์ข้อมูลเบื้องต้นและสร้างแบบจำลอง มีผลการศึกษา ดังนี้

1. ผลการเตรียมข้อมูล (data preparation) ผู้วิจัยได้เตรียมข้อมูลตามขั้นตอนที่วางแผนไว้ โดยมีผลการดำเนินการดังนี้

1.1 การคัดเลือกข้อมูล (data selection)

ใช้ฐานข้อมูล “Stroke Prediction Dataset”

จำนวน 5,110 ระเบียบ คุณลักษณะสำคัญ 11 คุณลักษณะที่มีความสัมพันธ์กับโรคหลอดเลือดในสมองตามการทบทวนวรรณกรรม (ยกเว้น รหัสผู้ป่วย-ID ที่ไม่นำมาใช้ในการวิเคราะห์)

1.2 การลดทอนมิติข้อมูล (data reduction) ผู้วิจัยได้จัดกลุ่มตัวแปรต่อเนื่องเป็นช่วงที่มีความหมายทางคลินิก ดังนี้

- อายุ แบ่งเป็น <40 ปี, 40-50 ปี, 50-65 ปี และ >65 ปี

- ระดับน้ำตาลในเลือด แบ่งเป็น <120 mg/dL, 120-150 mg/dL และ >150 mg/dL

- ดัชนีมวลกาย แบ่งเป็น <25 (ปกติ), 25-30 (น้ำหนักเกิน) และ >30 (อ้วน)

การจัดกลุ่มนี้ช่วยลดความซับซ้อนของข้อมูลและทำให้ผลลัพธ์ของการวิเคราะห์มีความหมายทางคลินิกมากขึ้น

1.3 การทำความสะอาดข้อมูล (data cleaning) ผู้วิจัยได้ดำเนินการทำความสะอาดข้อมูล โดย

- จัดการกับค่าที่ขาดหาย (missing values) ใน BMI ที่มีค่าเป็น “N/A” โดยแทนที่ด้วยค่าเฉลี่ยตามกลุ่มอายุและเพศ

- สำหรับ smoking_status ที่มีค่าเป็น “Unknown” ได้เก็บไว้เป็นกลุ่มแยกเพื่อวิเคราะห์ว่ามีผลต่อความเสี่ยงของโรคหรือไม่

- ตรวจสอบและกำจัดค่าผิดปกติ (outliers) โดยใช้วิธี Interquartile Range (IQR) สำหรับตัวแปรเชิงปริมาณ ได้แก่ อายุ, ระดับน้ำตาลในเลือดเฉลี่ย และ BMI ซึ่งพบค่าผิดปกติ 48 ระเบียบและได้ทำการปรับค่าให้อยู่ในพิสัยที่ยอมรับได้

1.4 การแปลงรูปแบบข้อมูล (data transformation) ผู้วิจัยได้แปลงข้อมูลจากไฟล์ CSV เป็นรูปแบบที่เหมาะสมสำหรับการวิเคราะห์โดย

- นำเข้าข้อมูลสู่โปรแกรมวิเคราะห์ทางสถิติ สำหรับการวิเคราะห์ความถดถอยโลจิสติกแบบพหุตัวแปร

- แปลงตัวแปรเชิงคุณภาพให้อยู่ในรูปแบบตัวแปรหุ่น (dummy variables) สำหรับการวิเคราะห์

- เตรียมข้อมูลในรูปแบบที่เหมาะสมสำหรับการใช้กับ Ant-miner algorithm

2. ลักษณะพื้นฐานของข้อมูลที่ศึกษาหลังจากเตรียมข้อมูลแล้ว มีอายุเฉลี่ย 43.22 ± 22.61 ปี ระดับน้ำตาลในเลือดเฉลี่ย 106.14 ± 45.28 mg/dL ดัชนีมวลกายเฉลี่ย 28.89 ± 7.85 ส่วนใหญ่เป็นเพศชาย (ร้อยละ 58.59) มีความดันโลหิตสูงร้อยละ 9.73 มีโรคหัวใจร้อยละ 5.40 และไม่เคยสูบบุหรี่ร้อยละ 37.03

2.1 ลักษณะการกระจายของข้อมูลและปัญหาความไม่สมดุลของข้อมูล จากข้อมูลทั้งหมด 5,110 ราย พบผู้ที่เป็นโรคหลอดเลือดในสมอง 249 ราย (ร้อยละ 4.87) และไม่เป็นโรค 4,861 ราย (ร้อยละ 95.13) แสดงให้เห็นถึงความไม่สมดุลของข้อมูลที่อาจส่งผลกระทบต่อประสิทธิภาพของแบบจำลองในการทำนายกลุ่มผู้ป่วยที่เป็นโรค ข้อมูลมีการกระจายของตัวแปรตาม attributes ที่สำคัญ ดังแสดงในตารางที่ 2 (Table 2)

Table 2 Distribution of key attributes by stroke outcome

Attributes	Category	No stroke (%)	Stroke (%)	p-value
Gender	Male	58.82	53.41	0.035
	Female	41.18	46.59	
Age	< 50 years	68.73	19.28	<0.001
	50-65 years	20.04	28.51	
	> 65 years	11.23	52.21	
Hypertension	No	92.59	54.22	<0.001
	Yes	7.41	45.78	
Heart disease	No	96.48	65.46	<0.001
	Yes	3.52	34.54	
Blood glucose level	< 120 mg/dL	78.85	44.98	<0.001
	120-150 mg/dL	11.91	22.49	
	> 150 mg/dL	9.24	32.53	
Body mass index	< 25	36.82	24.50	<0.001
	25-30	32.03	31.33	
	> 30	31.15	44.17	
Smoking status	Never smoked	37.91	23.29	<0.001
	Formerly smoked	16.89	26.10	
	Currently smokes	14.65	31.33	
	Unknown	30.55	19.28	

จากตารางที่ 2 (Table 2) ผลการวิเคราะห์พบว่า ตัวแปรอายุ ความดันโลหิตสูง โรคหัวใจ ระดับน้ำตาลในเลือด ดัชนีมวลกาย และการสูบบุหรี่ มีความสัมพันธ์อย่างมีนัยสำคัญทางสถิติกับการเกิดโรคหลอดเลือดในสมอง ($p < 0.001$) โดยกลุ่มผู้ป่วยที่เป็นโรคมีสัดส่วนของผู้ที่มีอายุมากกว่า 65 ปี มีความดันโลหิตสูง มีโรคหัวใจ มีระดับน้ำตาลในเลือดสูงกว่า 150 mg/dL มีดัชนีมวลกายสูงกว่า 30 และมีประวัติการสูบบุหรี่ในสัดส่วนที่สูงกว่ากลุ่มที่ไม่เป็นโรคอย่างชัดเจน

2.2 วิธีการแก้ไขปัญหาค่าความไม่สมดุลของข้อมูล เนื่องจากข้อมูลมีความไม่สมดุลอย่างมาก เพื่อแก้ไขปัญหาค่าความไม่สมดุลของข้อมูล ได้ใช้เทคนิค SMOTE (synthetic minority over-sampling

technique)¹² เพิ่มจำนวนตัวอย่างในกลุ่มผู้ป่วยโรคหลอดเลือดในสมองจาก 224 ราย (ร้อยละ 4.87) เป็น 2,187 ราย (ร้อยละ 50) และใช้ random under-sampling¹³ ลดจำนวนตัวอย่างในกลุ่มไม่เป็นโรคจาก 4,375 ราย เป็น 2,187 ราย ทำให้ชุดฝึกสอนมีสัดส่วนของทั้งสองกลุ่มเท่ากัน

2.3 การแบ่งข้อมูลและการทดสอบแบบจำลอง ในการศึกษาครั้งนี้ ข้อมูลทั้งหมด 5,110 ราย ถูกแบ่งออกเป็นชุดฝึกสอน (training Set) จำนวน 4,599 ราย (ร้อยละ 90) และชุดทดสอบ (test Set) จำนวน 511 ราย (ร้อยละ 10) โดยการสุ่มแบบแบ่งชั้นภูมิ (stratified sampling) เพื่อให้ทั้งสองชุดมีสัดส่วนของผู้ป่วยที่เป็นโรคและไม่เป็นโรคใกล้เคียงกับข้อมูลต้นฉบับ สำหรับชุดฝึกสอนได้ใช้วิธีการตรวจสอบ

แบบไขว้ 5 ส่วน (5-fold cross-validation) ซึ่งแบ่งชุดฝึกสอนออกเป็น 5 ส่วนเท่าๆ กัน และทำการฝึกสอนแบบจำลอง 5 รอบ โดยแต่ละรอบจะใช้ข้อมูล 4 ส่วนในการฝึกสอนและอีก 1 ส่วนในการทดสอบ

3. ผลการวิเคราะห์ทางสถิติในการศึกษาครั้งนี้แบบจำลองที่ได้จากการทำเหมืองข้อมูล จะใช้เทคนิคการวิเคราะห์ตัวแปรเชิงพหุ เพื่อเพิ่มประสิทธิภาพการทำนายเหตุการณ์ หรือประเมินความเสี่ยง โดย

ตัวแปรเกณฑ์ (ตัวแปรตาม) ที่เป็นตัวแปรแบบพหุนาม และตัวแปรอิสระ (ตัวแปรทำนาย) ที่มีหลายตัวที่เป็นไปได้ทั้งตัวแปรกลุ่ม (categorical variable) หรือตัวแปรต่อเนื่อง (continuous variable)

3.1 ผลการวิเคราะห์ความถดถอยโลจิสติกแบบพหุตัวแปรสำหรับปัจจัยเสี่ยงโรคหลอดเลือดในสมอง มีผลการศึกษาดังในตารางที่ 3 (Table 3)

Table 3 Multivariable logistic regression analysis results for stroke risk factors

Factor	β	SE(β)	Odds ratio	95% CI	p-value
Age > 65 years	1.3139	0.1335	3.72	2.86-4.84	<0.001
Hypertension	1.0919	0.1287	2.98	2.31-3.84	<0.001
Heart disease	1.0116	0.1462	2.75	2.07-3.66	<0.001
Blood glucose level > 150 mg/dL	0.8961	0.1317	2.45	1.89-3.17	<0.001
BMI > 30	0.6259	0.1332	1.87	1.44-2.43	<0.001
Current smoker	0.6575	0.1439	1.93	1.46-2.56	<0.001
Age 50-65 years	0.7608	0.1333	2.14	1.65-2.78	<0.001
Blood glucose level 120-150 mg/dL	0.5188	0.1345	1.68	1.29-2.19	<0.001
BMI 25-30	0.3716	0.1325	1.45	1.12-1.88	0.005
Former smoker	0.4187	0.1374	1.52	1.16-1.99	0.002
Work type (Private/Self-employed)	0.3293	0.1336	1.39	1.07-1.81	0.014
Male gender	0.2469	0.117	1.28	1.02-1.61	0.035
Ever married	0.3001	0.1285	1.35	1.05-1.74	0.019
Urban residence	0.174	0.114	1.19	0.95-1.49	0.127
Family history of stroke	0.6366	0.1446	1.89	1.42-2.51	<0.001

จากตารางที่ 3 (Table 3) พบว่า

1. อายุ > 65 ปี มีความเสี่ยงสูงที่สุด (OR = 3.72) และมีนัยสำคัญทางสถิติ ($p < 0.001$)

2. ความดันโลหิตสูง และโรคหัวใจ เป็นปัจจัยเสี่ยงสำคัญรองลงมา (OR = 2.98 และ 2.75 ตามลำดับ)

3. ระดับน้ำตาลในเลือดสูง เพิ่มความเสี่ยงอย่างมีนัยสำคัญ (OR = 2.45 สำหรับ > 150 mg/dL)

4. ดัชนีมวลกายสูง และการสูบบุหรี่ เพิ่มความเสี่ยงปานกลาง (OR = 1.87 และ 1.93 ตามลำดับ)

5. ประวัติ Stroke ในครอบครัว เป็นปัจจัยเสี่ยงที่สำคัญ (OR = 1.89)

6. เพศชาย มีความเสี่ยงเพิ่มขึ้นเล็กน้อย (OR = 1.28) เมื่อเทียบกับเพศหญิง

7. ประเภทที่อยู่อาศัย (เขตเมือง) ไม่มีนัยสำคัญทางสถิติ ($p = 0.127$)

3.2 การจำแนกตัวแปรเกณฑ์ระดับความเสี่ยงสูง กลาง และต่ำ แบบจำลองที่ได้จากการทำเหมืองข้อมูล ผู้ศึกษาวิเคราะห์ค่า odds ratio (OR) และ Confidence Interval (CI) ใช้ในการกำหนดเกณฑ์ความเสี่ยงในแต่ละระดับ สามารถจัดระดับความเสี่ยงตามค่า OR ได้ดังนี้ ความเสี่ยงสูง (OR > 2.0) ได้แก่ อายุ > 65 ปี (OR = 3.72), ความดันโลหิตสูง (OR = 2.98), โรคหัวใจ (OR = 2.75), และระดับน้ำตาลในเลือด > 150 mg/dL (OR = 2.45) ความเสี่ยงปานกลาง (OR = 1.5-2.0) ได้แก่ BMI > 30 (OR = 1.87), การสูบบุหรี่ในปัจจุบัน (OR = 1.93), และอายุ 50-65 ปี (OR = 2.14) ความเสี่ยงต่ำ (OR < 1.5) ได้แก่ อายุ < 50 ปี (OR = 1.00), ไม่มีความดันโลหิตสูง (OR = 1.00), ระดับน้ำตาลในเลือด < 120 mg/dL (OR = 1.00), BMI < 25 (OR = 1.00), และไม่เคยสูบบุหรี่ (OR = 1.00), เพื่อนำไปใช้สร้างกฎ (term)

4. ประสิทธิภาพของแบบจำลองสำหรับชุดฝึกสอนและชุดทดสอบ แบบจำลองการวินิจฉัยโรค

หลอดเลือดในสมองด้วย Ant-miner algorithm (5-fold cross validation) เริ่มต้นด้วยการสร้างกลุ่มมดในการศึกษา ผู้ศึกษาตั้งค่าพารามิเตอร์เบื้องต้น ดังนี้ จำนวนมด 1,000 ตัว และอัตราการระเหยของฟีโรโมน 0.1 และค่า convergence เท่ากับ 10 รอบ การทดสอบประสิทธิภาพของแบบจำลองด้วยวิธี 5-fold cross validation บนชุดฝึกสอน 4,599 ราย พบว่ามีค่า accuracy เฉลี่ยร้อยละ 95.58 (95.32-95.87), precision เฉลี่ยร้อยละ 90.88 (90.45-91.23), recall เฉลี่ยร้อยละ 92.53 (92.18-92.76), และ F1-score เฉลี่ยร้อยละ 91.66 (91.31-91.99)

4.1 ผลการทำนายความเสี่ยงของโรคหลอดเลือดในสมองด้วย Ant-miner algorithm และประสิทธิภาพของแบบจำลองการวินิจฉัยโรคหลอดเลือดในสมองด้วย Ant-miner algorithm สำหรับชุดทดสอบ (test set) ดังแสดงในตารางที่ 4 (Table 4)

Table 4 Confusion matrix and performance measures of the ant-miner model on test set

Actual \ Predicted	Positive	Negative
Positive	46 (True Positive TP)	4 (False Negative FN)
Negative	5 (False Positive FP)	456 (True Negative TN)

Confusion Matrix (N = 511)

จากตารางที่ 4 (Table 4) พบว่า แบบจำลองการวินิจฉัยโรคหลอดเลือดในสมองด้วย Ant-miner algorithm เมื่อทดสอบด้วยข้อมูล test set จำนวน 511 ราย มีประสิทธิภาพด้าน accuracy เฉลี่ยร้อยละ 98.24 แสดงว่าแบบจำลองสามารถทำนายได้ถูกต้องมากกว่าร้อยละ 95 ของกรณีทั้งหมด ด้าน precision เฉลี่ยร้อยละ 92.00 บ่งชี้ว่าเมื่อแบบจำลองทำนายว่า มีความเสี่ยงจะถูกต้องประมาณร้อยละ 92 ของกรณีทั้งหมด ด้าน recall เฉลี่ยร้อยละ 90.20 แสดงว่าแบบจำลองสามารถระบุผู้ที่มีความเสี่ยงจริงได้ประมาณ

ร้อยละ 90 ของกรณีทั้งหมด และ F1-score เฉลี่ยร้อยละ 91.09 แสดงถึงความสมดุลที่ดีระหว่าง precision และ recall

4.2 แบบจำลองสำหรับการวินิจฉัยโรคหลอดเลือดในสมองที่สร้างด้วย Ant-miner algorithm ได้สร้างกฎการวินิจฉัย (term) จำนวนทั้งสิ้น 25 กฎ โดยสามารถจัดกลุ่มตามระดับความเสี่ยง แสดงกฎที่สำคัญและมีค่า OR สูงในแต่ละกลุ่มความเสี่ยง ดังแสดงในตารางที่ 5 (Table 5)

Table 5 Key diagnostic rules generated by ant-miner algorithm

Term	Odds ratio
A. High Risk Rules (OR > 15)	
1. IF (age > 65) AND (hypertension = 1) AND (heart_disease = 1) AND (ever_married = 'Yes')	41.43
2. IF (gender = 'Female') AND (age > 75) AND (hypertension = 1) AND (heart_disease = 1)	34.01
3. IF (age > 65) AND (avg_glucose_level > 150) AND (hypertension = 1)	27.12
4. IF (age > 70) AND (bmi > 28) AND 25.54 (heart_disease = 1) AND (ever_married = 'Yes')	25.54
5. IF (gender = 'Male') AND (age > 70) AND (hypertension = 1) AND (ever_married = 'Yes')	21.32
B. Moderate Risk Rules (5 < OR < 15)	
1. IF (age > 50) AND (bmi > 35) AND (avg_glucose_level > 150) AND (ever_married = 'Yes')	14.23
2. IF (age > 60) AND (avg_glucose_level > 120) AND (hypertension = 1) AND (Residence_type = 'Urban')	11.96
3. IF (age > 50) AND (avg_glucose_level > 130) 10.43 AND (hypertension = 1) AND (work_type = 'Private')	10.43
C. Low Risk Rules (OR < 5)	
1. IF (age < 45) AND (bmi < 25) AND (smoking_status = 'never smoked')	1.00
2. IF (age < 30) AND (hypertension = 0) AND (heart_disease = 0) AND (bmi < 22)	0.88
3. IF (age < 25) AND (avg_glucose_level < 90) 0.64 AND (bmi < 20) AND (smoking_status = 'never smoked')	0.64

จากตารางที่ 5 (Table 5) แสดงกฎการวินิจฉัยที่สำคัญที่ได้จาก Ant-miner algorithm ซึ่งสามารถสรุปลักษณะสำคัญได้ดังนี้

1. กฎ (term) ความเสี่ยงสูง [A. high risk rules (OR > 15)] แสดงให้เห็นว่าปัจจัยที่มีความสำคัญสูงสุดคือ การมีปัจจัยเสี่ยงหลายอย่างร่วมกัน โดยเฉพาะอย่างยิ่ง อายุมากกว่า 65 ปี ร่วมกับความดันโลหิตสูงและโรคหัวใจ ซึ่งเพิ่มความเสี่ยงได้สูงถึง 41.43 เท่า กฎเหล่านี้มีความสอดคล้องกับความรู้ทางการแพทย์ที่ว่าผู้สูงอายุที่มีโรคประจำตัวมีความเสี่ยงสูงต่อการเกิดโรคหลอดเลือดในสมอง

2. กฎ (term) ความเสี่ยงปานกลาง [B. moderate risk rules (5 < OR < 15)] แสดงถึงบทบาทของปัจจัยด้านอาชีพและที่อยู่อาศัย ร่วมกับปัจจัยทางสุขภาพ โดยพบว่าผู้ที่ทำงานภาคเอกชนหรืออาศัยในเขตเมืองที่มีอายุมากกว่า 50 ปี และมีปัจจัยเสี่ยงทางสุขภาพอย่างน้อย 2 ปัจจัย มีความเสี่ยงต่อโรคหลอดเลือดในสมองในระดับปานกลาง

3. กฎ (term) ความเสี่ยงต่ำ [C. low risk rules (OR < 5)] ยืนยันว่า บุคคลที่มีอายุน้อย (ต่ำกว่า 45 ปี) ไม่มีภาวะความดันโลหิตสูงหรือโรคหัวใจ มีดัชนีมวลกายปกติ และไม่สูบบุหรี่ มีความเสี่ยงต่ำต่อการเกิดโรคหลอดเลือดในสมอง

กฎ (term) เหล่านี้แสดงให้เห็นถึงความซับซ้อนของปัจจัยเสี่ยงที่ส่งผลต่อการเกิดโรคหลอดเลือดในสมอง และความสามารถของ Ant-miner algorithm ในการค้นหารูปแบบความสัมพันธ์ที่ซับซ้อนระหว่างปัจจัยเสี่ยงต่างๆ ซึ่งสามารถนำไปประยุกต์ใช้ในการคัดกรองและประเมินความเสี่ยงของผู้ป่วยในทางคลินิกได้ จากกฎการวินิจฉัยโรคหลอดเลือดในสมองทั้งหมด พบว่า

อายุ (age) มีความสำคัญที่สุด ทุกกฎมีเงื่อนไขเกี่ยวกับอายุ โดยส่วนใหญ่จะเริ่มที่อายุมากกว่า 40 ปีขึ้นไป และความเสี่ยงจะเพิ่มขึ้นตามอายุที่มากขึ้น

ปัจจัยทางสุขภาพ ความดันโลหิตสูง (hypertension) และโรคหัวใจ (heart_disease) ปรากฏในหลายกฎ โดยเฉพาะในกฎที่ระบุความเสี่ยงสูง

ดัชนีมวลกาย (BMI) และระดับน้ำตาลในเลือด (avg_glucose_level) เป็นปัจจัยสำคัญที่ปรากฏในหลายกฎ โดยค่าที่สูงมักจะเพิ่มความเสี่ยง

พฤติกรรมการสูบบุหรี่ (smoking_status) ทั้งผู้ที่สูบบุหรี่ปัจจุบันและเคยสูบในอดีตมีความเสี่ยงเพิ่มขึ้น โดยเฉพาะเมื่อร่วมกับปัจจัยอื่นๆ

ประเภทงาน (work_type) งานภาคเอกชน (private) และอาชีพอิสระ (self-employed) ปรากฏในหลายกฎ อาจเกี่ยวข้องกับความเร็วหรือลักษณะการทำงาน

เพศ (sex) มีการแยกกฎสำหรับเพศชายและหญิงในบางกรณี แสดงถึงความแตกต่างของปัจจัยเสี่ยงระหว่างเพศ

สถานภาพสมรส (ever_married) การแต่งงานปรากฏในบางกฎ โดยเฉพาะในกลุ่มอายุมาก อาจเกี่ยวข้องกับปัจจัยทางสังคมและการดูแลสุขภาพ

ประเภทที่อยู่อาศัย (residence_type) ปรากฏในบางกฎ แสดงถึงความแตกต่างของปัจจัยเสี่ยงระหว่างเขตเมืองและชนบท

กฎ (term) ที่ระบุความเสี่ยงสูงมักจะรวมหลายปัจจัยเสี่ยงเข้าด้วยกัน เช่น อายุมาก ความดันโลหิตสูง และโรคหัวใจ ขณะที่ปัจจัยด้านพฤติกรรมและสังคม เช่น การสูบบุหรี่ อาชีพ สถานภาพสมรส มีบทบาทเมื่อพิจารณาร่วมกับปัจจัยทางสุขภาพและอายุ ความซับซ้อนของกฎ กฎส่วนใหญ่ประกอบด้วย 3-4 เงื่อนไข แสดงถึงความซับซ้อนของการประเมินความเสี่ยง Stroke

การศึกษานี้ได้เปรียบเทียบประสิทธิภาพของ Ant-miner algorithm กับวิธี decision tree และ random forest ในการวินิจฉัยโรคหลอดเลือดในสมอง โดยใช้ข้อมูลจาก Stroke Prediction Dataset เดียวกัน ผลการเปรียบเทียบแสดงให้เห็นถึงประสิทธิภาพของ Ant-miner algorithm ที่สูงกว่าในทุกด้านที่ประเมิน แสดงในตารางที่ 6 (Table 6)

Table 6 Comparison of performance measures across different algorithms

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Ant-miner (My proposal)	98.24	92.00	90.20	91.09
Decision tree	91.20	89.70	86.50	88.07
Random forest	92.70	91.50	87.60	89.51

ผลการเปรียบเทียบนี้แสดงว่า Ant-miner algorithm มีประสิทธิภาพเหนือกว่า decision tree และ random forest ในทุกด้าน โดยเฉพาะอย่างยิ่งด้านความถูกต้องและความแม่นยำ ซึ่งมีความสำคัญอย่างยิ่งในการวินิจฉัยทางการแพทย์ ทั้งนี้ควรพิจารณาข้อดีของแต่ละวิธีประกอบกันด้วย เช่น ความเร็วในการประมวลผลและความง่ายในการอธิบายผลลัพธ์ เพื่อเลือกวิธีที่เหมาะสมที่สุดสำหรับการนำไปใช้จริงในทางคลินิก

อภิปรายผล

การวิเคราะห์ข้อมูลทางการแพทย์และการสกัดปัจจัยเสี่ยง ใช้ข้อมูลสำหรับการศึกษาจาก Stroke Prediction Dataset ประกอบด้วย ข้อมูลผู้ป่วย 5,110 ราย ผลการศึกษาจากการวิเคราะห์ความถดถอยโลจิสติกแบบพหุตัวแปร พบว่า ปัจจัยเสี่ยงที่สำคัญที่สุดของโรคหลอดเลือดในสมอง คือ อายุมากกว่า 65 ปี (OR = 3.72, 95% CI 2.86-4.84) สอดคล้องกับการศึกษาของ Wang et al. (2020) ที่พบว่า อายุเป็นปัจจัยเสี่ยงหลักของโรคหลอดเลือดในสมอง¹⁴ นอกจากนี้ ความดันโลหิตสูง (OR = 2.98) และโรคหัวใจ

(OR = 2.75) ก็เป็นปัจจัยเสี่ยงสำคัญ สอดคล้องกับการศึกษาของ Boehme et al. (2017)¹⁵ และ Suwanwela N.C. (2014)¹⁶ ที่ระบุว่า ความดันโลหิตสูงและโรคหัวใจ เป็นปัจจัยเสี่ยงหลักของโรคหลอดเลือดในสมองทั้งในระดับสากลและในประเทศไทย ระดับน้ำตาลในเลือดสูง (OR = 2.45 สำหรับ > 150 mg/dL) เป็นปัจจัยเสี่ยงที่สำคัญ สอดคล้องกับการศึกษาของ Zheng et al. (2018) ที่พบว่า ภาวะน้ำตาลในเลือดสูงเพิ่มความเสี่ยงต่อการเกิดโรคหลอดเลือดในสมองอย่างมีนัยสำคัญ¹⁷ ปัจจัยด้านดัชนีมวลกายสูง (OR = 1.87 สำหรับ BMI > 30) และการสูบบุหรี่ (OR = 1.93 สำหรับผู้ที่สูบบุหรี่ในปัจจุบัน) ก็เป็นปัจจัยเสี่ยงสำคัญที่ไม่ควรมองข้าม ประสิทธิภาพของแบบจำลองที่สร้างด้วย Ant-miner algorithm มีประสิทธิภาพสูงกว่าวิธีการเรียนรู้ของเครื่องแบบดั้งเดิม โดยมีค่า accuracy 98.24%, precision 92.00%, recall 90.20%, และ F1-score 91.09% สูงกว่า decision tree (accuracy 91.20%) และ random forest (accuracy 92.70%) อย่างมีนัยสำคัญ ผลลัพธ์นี้สอดคล้องกับการศึกษาของ Shen et al. (2016) ที่พบว่า วิธีการเรียนรู้ของเครื่องที่อิงตามพฤติกรรมของแมลง เช่น Ant colony optimization มีประสิทธิภาพสูงในการวิเคราะห์ข้อมูลทางการแพทย์ที่มีความซับซ้อน¹⁸ Ant-miner algorithm มีจุดเด่นที่สำคัญคือ สามารถสร้างกฎการวินิจฉัยที่เข้าใจง่ายและความหมายทางคลินิก เช่น “ถ้า (อายุ > 65) และ (ระดับน้ำตาลในเลือดเฉลี่ย > 150) และ (มีความดันโลหิตสูง) แล้วความเสี่ยงต่อโรคหลอดเลือดในสมอง = สูง” (OR = 27.12) ซึ่งแตกต่างจาก neural networks ที่มักถูกมองว่าเป็น “black box” เนื่องจากความซับซ้อนในการอธิบายผลลัพธ์ (Yu et al., 2018) (29) ความสามารถในการสร้างกฎที่อธิบายได้นี้ สอดคล้องกับแนวคิดของ Cynthia R. (2019) ที่เน้นย้ำถึงความสำคัญของ interpretable machine learning ในการตัดสินใจทางการแพทย์¹⁹

เมื่อเปรียบเทียบกับวิธีอื่นๆ ที่ใช้ในการศึกษานี้พบว่า Ant-miner algorithm มีข้อได้เปรียบเหนือ

decision tree ในด้านการรับมือกับข้อมูลที่มีความซับซ้อนและมีหลายมิติ โดย decision tree มักมีข้อจำกัดในการจัดการกับพื้นที่ค้นหาที่ซับซ้อนและมีแนวโน้มที่จะเกิด overfitting เมื่อต้นไม่มีความลึกมาก ในขณะที่ random forest แม้จะช่วยลดปัญหา overfitting ของ decision tree แต่ก็มีข้อเสียคือความซับซ้อนในการตีความและความยากในการทำความเข้าใจกฎการตัดสินใจที่ซ่อนอยู่ภายใน ตามที่ Molnar et al. (2020) ได้วิเคราะห์ไว้²⁰ ส่วน neural networks แม้จะมีประสิทธิภาพสูงในการจำแนกรูปแบบที่ซับซ้อน แต่จำเป็นต้องใช้ข้อมูลจำนวนมากในการฝึกสอนและมีความยากในการอธิบายกระบวนการตัดสินใจ ซึ่งเป็นข้อจำกัดสำคัญในการประยุกต์ใช้ทางคลินิก (Topol, 2019)²¹

ข้อได้เปรียบสำคัญอีกประการของ Ant-miner algorithm คือความสามารถในการดำเนินการกับข้อมูลที่ไม่สมดุล ซึ่งเป็นลักษณะทั่วไปของข้อมูลโรคหลอดเลือดในสมอง โดยมีค่า recall สูงถึง 90.20% แสดงถึงความสามารถในการตรวจจับปัจจัยเสี่ยงกรณีที่เป็นโรคได้ดี สอดคล้องกับการศึกษาของ Chawla et al. (2002) ที่เน้นย้ำถึงความสำคัญของการจัดการกับข้อมูลไม่สมดุลในการวินิจฉัยทางการแพทย์²² ในการจัดการกับข้อมูลไม่สมดุลของโรคหลอดเลือดในสมองซึ่งมีลักษณะเป็นโรคที่พบไม่บ่อยแต่มีความรุนแรงสูง การศึกษานี้ได้ประยุกต์ใช้เทคนิค Synthetic Minority Over-sampling Technique ตามที่ Fernandez et al. (2018) แนะนำไว้¹² เพื่อเพิ่มจำนวนตัวอย่างในกลุ่มผู้ป่วยโรคหลอดเลือดในสมองที่มีจำนวนน้อยกว่า และใช้เทคนิค random under-sampling ในกลุ่มไม่เป็นโรคที่มีจำนวนมากกว่า ตามแนวทางของ Wang et al. (2016)¹³ เพื่อให้ได้ชุดข้อมูลที่มีความสมดุลมากขึ้น วิธีการเหล่านี้ช่วยเพิ่มประสิทธิภาพของการทำนาย โดยเฉพาะในกลุ่มผู้ป่วยที่เป็นโรคซึ่งมีความสำคัญทางคลินิกมากกว่า นอกจากนี้ยังมีการใช้เทคนิค cost-sensitive learning ตามที่ Johnson และ Khoshgoftaar (2019) นำเสนอ²³

ร่วมด้วย โดยกำหนดค่าถ่วงน้ำหนักให้การทำนายผิดในกลุ่มผู้ป่วยโรคหลอดเลือดในสมอง (false negative) มีค่าเสียหายสูงกว่าการทำนายผิดในกลุ่มไม่เป็นโรค (false positive) เนื่องจากผลเสียของการไม่ได้รับการรักษาในผู้ป่วยโรคหลอดเลือดในสมองมีความรุนแรงมากกว่า แบบจำลองนี้สามารถสร้างกฎที่มีความซับซ้อนมากขึ้น เช่น “IF (age > 65) AND (hypertension = 1) AND (heart_disease = 1) AND (ever_married = ‘Yes’) THEN ความเสี่ยง Stroke = สูง” (OR = 41.43) กฎเหล่านี้สะท้อนถึงความสัมพันธ์ที่ซับซ้อนระหว่างปัจจัยเสี่ยงต่างๆ สอดคล้องกับการศึกษาของ Liu et al. (2018) ที่เน้นย้ำถึงความสำคัญของการพิจารณาความสัมพันธ์ระหว่างปัจจัยเสี่ยงในการประเมินความเสี่ยงของโรคหลอดเลือดในสมอง²⁴

อย่างไรก็ตาม Ant-miner algorithm ก็มีข้อจำกัดที่ควรพิจารณา เช่นเดียวกันกับการศึกษาของ Zhang et al. (2017) ที่พบว่าวิธีการเรียนรู้ของเครื่องที่เรียนแบบพฤติกรรมธรรมชาติ มีประสิทธิภาพสูงในการวิเคราะห์ข้อมูลทางการแพทย์ที่ซับซ้อน²⁵ แต่ก็มีความเสี่ยงต่อปัญหา overfitting ในการใช้ algorithms ที่ซับซ้อนกับข้อมูลทางการแพทย์สอดคล้องกับข้อสังเกตของ Domingos (2012)²⁶ ซึ่งอาจส่งผลต่อ generalizability ของแบบจำลองเมื่อนำไปใช้กับข้อมูลจริงในบริบทที่แตกต่างออกไป ตามที่ Khosla et al. (2020) ได้ศึกษาไว้²⁷ โดยเฉพาะเมื่อต้องนำไปใช้กับประชากรไทยที่มีลักษณะทางพันธุกรรม วิถีชีวิต พฤติกรรมการบริโภคอาหาร และปัจจัยแวดล้อมที่แตกต่างจากข้อมูลในฐานข้อมูลที่ใช้ในการศึกษานี้ ซึ่งเป็นข้อมูลจากโรงพยาบาลในประเทศสเปน นอกจากนี้ ปัจจัยสำคัญอื่นๆ เช่น รายได้ การเข้าถึงระบบฐานข้อมูล EHRs ของโรงพยาบาลในประเทศไทย และปัจจัยเอื้อต่อการดูแลตนเองหรือออกกำลังกายที่ไม่ได้รับการเก็บข้อมูลมาในการศึกษานี้ อาจส่งผลต่อความแม่นยำของแบบจำลองเมื่อนำไปใช้ในบริบทที่แตกต่างออกไป

สำหรับการพัฒนาการศึกษาในอนาคต มีข้อพิจารณาหลายประการที่ควรคำนึงถึง การนำผล

การศึกษานี้ไปประยุกต์ใช้ในบริบทของประชากรไทย ควรทำด้วยความระมัดระวัง เนื่องจากข้อมูลที่ใช้ในการศึกษามาจากประชากรในประเทศสเปน ซึ่งมีปัจจัยที่อาจแตกต่างกัน ทั้งลักษณะทางพันธุกรรม พฤติกรรม การบริโภคอาหาร การเข้าถึงระบบสุขภาพ และปัจจัยเอื้อต่อการดูแลตนเอง จึงควรมีการศึกษาเพิ่มเติมโดยใช้ข้อมูลจากประชากรไทยโดยตรงเพื่อสร้างแบบจำลองที่เหมาะสมกับบริบทของประเทศไทย นอกจากนี้ การเปรียบเทียบประสิทธิภาพของ Ant-miner algorithm กับโมเดลอื่นๆ ที่ได้รับความนิยมในปัจจุบัน เช่น XGBoost, LightGBM หรือโมเดล deep learning ต่างๆ จะช่วยให้เห็นถึงข้อได้เปรียบและข้อเสียเปรียบของแต่ละโมเดลอย่างชัดเจน อีกทั้งอาจพิจารณาใช้เทคนิค ensemble methods ที่ผสมผสานจุดแข็งของหลายโมเดลเข้าด้วยกัน การเปรียบเทียบผลลัพธ์ที่ได้จากแบบจำลอง Ant-miner กับการวินิจฉัยของแพทย์ผู้เชี่ยวชาญในข้อมูลชุดเดียวกันก็มีความสำคัญ เพื่อประเมินว่าแบบจำลองสามารถให้ผลลัพธ์ที่สอดคล้องกับการวินิจฉัยทางคลินิกจริงหรือไม่ ซึ่งจะช่วยสร้างความเชื่อมั่นในการนำแบบจำลองไปใช้เป็นเครื่องมือสนับสนุนการตัดสินใจทางคลินิก การทดสอบแบบจำลองกับข้อมูลจากหลายแหล่ง โดยเฉพาะข้อมูลจากโรงพยาบาลในประเทศไทย จะช่วยประเมินความสามารถในการทำนายในบริบทที่แตกต่าง การพัฒนาแบบจำลองให้อยู่ในรูปแบบที่สามารถนำไปใช้งานได้จริงในระบบสาธารณสุข เช่น การพัฒนาเป็น web application หรือ mobile application ที่ใช้งานง่ายสำหรับบุคลากรทางการแพทย์ หรือการบูรณาการเข้ากับระบบเวชระเบียนอิเล็กทรอนิกส์ที่มีอยู่แล้ว จะช่วยให้การคัดกรองและประเมินความเสี่ยงของโรคหลอดเลือดในสมองทำได้อย่างมีประสิทธิภาพมากขึ้น ประเด็นเรื่องข้อมูลที่ไม่สมดุลยังคงเป็นความท้าทาย ควรมีการศึกษาและพัฒนาเทคนิคการจัดการข้อมูลที่ไม่สมดุลเพิ่มเติม เช่น การใช้เทคนิค hybrid sampling ที่ผสมผสานระหว่าง over-sampling และ under-sampling หรือการพัฒนาอัลกอริทึมใหม่ที่ยืดหยุ่น

เฉพาะสำหรับข้อมูลที่ไม่สมดุลในบริบทของโรคหลอดเลือดในสมอง ท้ายที่สุด เทคนิค Ant-miner algorithm อาจถูกนำไปประยุกต์ใช้กับโรคอื่นๆ ที่มีปัจจัยเสี่ยงซับซ้อนและมีความสำคัญทางสาธารณสุข เช่น โรคหัวใจ โรคเบาหวาน หรือโรคมะเร็ง เพื่อขยายประโยชน์ของการใช้เทคนิคเหมืองข้อมูลในการวินิจฉัยและคัดกรองโรค

สรุปผล

การวิเคราะห์ข้อมูลทางการแพทย์ของโรคหลอดเลือดในสมองจาก Stroke Prediction Dataset สามารถระบุปัจจัยเสี่ยงที่สำคัญได้ โดยอายุมากกว่า 65 ปี ความดันโลหิตสูง โรคหัวใจ และระดับน้ำตาลในเลือดสูงเป็นปัจจัยหลักที่มีความเสี่ยง นอกจากนี้ยังพบความสัมพันธ์ที่น่าสนใจระหว่างปัจจัยต่างๆ เช่น อายุ กับระดับน้ำตาลในเลือด และอายุกับความดันโลหิตสูง แบบจำลองการวินิจฉัยโรคหลอดเลือดในสมองที่สร้างด้วย Ant-miner algorithm มีประสิทธิภาพสูง โดยมีค่า accuracy ร้อยละ 98.24, precision ร้อยละ 92.00, recall ร้อยละ 90.20 และ F1-score ร้อยละ 91.09 ซึ่งสูงกว่าวิธีการอื่นๆ ที่นำมาเปรียบเทียบ แบบจำลองที่พัฒนาขึ้นนี้มีความสามารถในการจัดการกับข้อมูลที่ไม่สมดุล อันเป็นลักษณะสำคัญของข้อมูลทางการแพทย์ กฎการวินิจฉัยที่ได้จากแบบจำลองสามารถนำไปประยุกต์ใช้ในการคัดกรองและประเมินความเสี่ยงของโรคหลอดเลือดในสมองในเวชปฏิบัติได้อย่างมีประสิทธิภาพ นำไปพัฒนาเป็นเครื่องมือสนับสนุนการตัดสินใจทางคลินิก และประยุกต์พัฒนาเป็น application software ได้ อย่างไรก็ตาม การนำแบบจำลองนี้ไปใช้ในทางปฏิบัติควรคำนึงถึงข้อจำกัดและความแตกต่างของประชากรในแต่ละพื้นที่ รวมถึงการปรับปรุงแบบจำลองอย่างต่อเนื่องเพื่อให้สอดคล้องกับข้อมูลและแนวโน้มทางสุขภาพที่เปลี่ยนแปลงไป

เอกสารอ้างอิง

1. American Stroke Association [Internet]. Types of stroke and treatment; [cited 2024 Apr 11]. Available from: <https://www.stroke.org/en/about-stroke/types-of-stroke>.
2. World Health Organization [Internet]. Cardiovascular diseases (CVDs); [cited 2024 Apr 12]. Available from: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
3. International Health Policy Program, Ministry of Public Health. Disability-adjusted life years DALYs 2019. Bangkok: Handy Press; 2023. (in Thai)
4. Durairaj M, Ranjani VA. Data mining applications in healthcare sector: A study. Int J Sci Technol Res 2013;2(10):29-35.
5. Parpinelli AA, Lopes HS, Freitas AA. Data mining with an ant colony optimization algorithm. IEEE Trans Evol Comput 2002; 6(4):321-32. doi:10.1109/TEVC.2002.802452.
6. Soriano F. Stroke prediction dataset [Internet]; [cited 2024 Apr 13]. Available from: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>.
7. Han J, Kamber M, Pei J. Data mining concepts and techniques. 3rd ed. Massachusetts: Morgan Kaufmann; 2012.
8. Witten IH, Frank E, Hall MA, et al. Data mining practical machine learning tools and techniques. 4th ed. Massachusetts: Morgan Kaufmann; 2016.

9. Kankhajane C. Accuracy assessment remote sensing technical note no. 3 [Internet]; Faculty of Forestry, Kasetsart University; 2018 [cited 2024 May 13]. Available from: <https://forest-admin.forest.ku.ac.th/304xxx/?q=system/files/book/3%282018%29%20Accuracy%20Assessment.pdf>. [in Thai]
10. Hartshorn S. Machine learning with random forests and decision trees: A visual guide for beginners. Amazon Kindle; 2016.
11. Pacharawongsakda E. An introduction to data mining techniques. 2nd ed. Bangkok: ASIA Digital PRESS; 2014.
12. Fernandez A, Garcia S, Herrera F, et al. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J Artif Intell Res* 2018;61:863-905. doi:10.1613/jair.1.11192.
13. Wang S, Liu W, Wu J, et al. Training deep neural networks on imbalanced data sets. In: 2016 International Joint Conference on Neural Networks (IJCNN); 2016 Jul 24-29; Vancouver, BC, Canada. IEEE; 2016; 4368-74. doi:10.1109/IJCNN.2016.7727770.
14. Wang Y, Xingquan Z, Yong J, et al. Prevalence, knowledge, and treatment of transient ischemic attacks in China. *Neurology* 2015;84(23):2354-61. doi:10.1212/WNL.0000000000001665.
15. Boehme AK, Esenwa C, Elkind MS. Stroke risk factors, genetics, and prevention. *Circ Res* 2017;120(3):472-95. doi:10.1161/CIRCRESAHA.116.308398.
16. Suwanwela NC. Stroke epidemiology in Thailand. *J Stroke* 2014;11-7. (In Thai)
17. Zheng Y, Ley SH, Hu FB. Global etiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol* 2018;14(2):88-98. doi:10.1038/nrendo.2017.151.
18. Shen L, Chen H, Yu Z, et al. Evolving support vector machines using fruit fly optimization for medical data classification. *Knowl Based Syst* 2016;96:61-75. doi:10.1016/j.knosys.2016.01.002.
19. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;1(5):206-15. doi:10.1038/s42256-019-0048-x.
20. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning – a brief history, state-of-the-art and challenges. In: Koprinska I, et al., editors. *ECML PKDD 2020 Workshops*. Cham: Springer; 2020. p. 417–31. (Communications in Computer and Information Science; vol. 1323). doi:10.1007/978-3-030-65965-3_28.
21. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7.
22. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321-57. doi:10.1613/jair.953.

23. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data* 2019;6(27):1-54. doi:10.1186/s40537-019-0192-5.
24. Liu L, Chen W, Zhou H, et al. Chinese Stroke Association guidelines for clinical management of cerebrovascular disorders executive summary and 2019 update of clinical management of ischemic cerebrovascular diseases. *Stroke Vasc Neurol* 2020;5(2):159-76. doi:10.1136/svn-2020-000378.
25. Zhang C, Liu C, Zhang X, et al. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Syst Appl* 2017;82:128-50. doi:10.1016/j.eswa.2017.04.003.
26. Domingos P. A few useful things to know about machine learning. *Commun ACM* 2012;55(10):78-87. doi:10.1145/2347736.2347755.
27. Khosla A, Cao Y, Lin CCY, et al. An integrated machine learning approach to stroke prediction. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*. New York (NY): Association for Computing Machinery; 2010. p. 183-92. doi:10.1145/1835804.1835830.