แบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง

Automatic Diagnosis Model for Risk of Systomatic Venous Thromboembolism based on Machine Learning

อัจฉราภรณ์ สุขเพิ่ม
สาขาวิทยาการคอมพิวเตอร์,
คณะวิทยาศาสตร์
มหาวิทยาลัยศรีนครินทรวิโรฒ
กรุงเทพมหานคร, ประเทศไทย
autcharaporn.sukperm@g.swu.ac.th

พลภัทร โรจน์นครินทร์
ภาควิชาอายุรศาสตร์,
คณะแพทยศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย
กรุงเทพมหานคร, ประเทศไทย
rojnuckarinp@gmail.com

เบญจพร อัคควัฒน์
ภาควิชาอายุรศาสตร์,
คณะแพทยศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย
กรุงเทพมหานคร, ประเทศไทย
smedbam@gmail.com

วีระ สอึ้ง
สาขาวิทยาการคอมพิวเตอร์,
คณะวิทยาศาสตร์
มหาวิทยาลัยศรีนครินทรวิโรฒ
กรุงเทพมหานคร, ประเทศไทย
vera@g.swu.ac.th

***บทคัดย่อ*—ภาวะลิ่มเลือดอุดตันในหลอดเลือดดำเป็นโรคหนึ่งที่สำคัญที่มีผู้ป่วยเพิ่มมากขึ้นในประเทศไทยซึ่งเกิดจากการปิดกั้นการไหลเวียนของเลือดในหลอดเลือดดำ นอกจากนี้แบบจำลองการประเมินความเสี่ยงของภาวะลิ่มเลือดอุดตันในหลอดเลือดดำที่มีประสิทธิภาพจึงเป็นสิ่งสำคัญที่สุดเพื่อช่วยแพทย์วินิจฉัย งานวิจัยนี้ได้สร้างแบบจำลองการวินิจฉัยความเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ โดยอาศัยหลักการเรียนรู้ของเครื่อง จากการเก็บรวบรวมข้อมูลผู้ป่วยในหอผู้ป่วยอายุรศาสตร์ โรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย งานวิจัยนี้ได้เตรียมข้อมูลทั้งหมด 1290 แถว 65 คอลัมน์ และตรวจสอบค่าที่ขาดหายไปพร้อมทั้งแปลงข้อมูลให้พร้อมสำหรับนำไปสร้างแบบจำลองการทำนายในลำดับต่อไป จากนั้นแบ่งข้อมูลสำหรับการฝึกสอนและข้อมูลสำหรับการทดสอบในอัตราส่วน 70:30 ผลการทดลองของงานวิจัยได้เปรียบเทียบประสิทธิภาพของแต่ละ 3 อัลกอริทึม ประกอบด้วย Decision Tree, Logistic regression และ Neural network จากผลการทดลองแบบจำลอง Decision tree มีประสิทธิภาพที่สุด มีค่าความถูกต้องสูงที่สุด 96.6% โดยการปรับสมดุลของข้อมูลด้วยวิธี Class Weight**

Keywords— venous thromboembolism, machine learning, automatic diagnosis

I. บทนำ

ภาวะหลอดเลือดดำอุดตัน (Venous Thromboembolism : VTE) เป็นปัญหาสำคัญที่ทำให้มีผู้ป่วยเพิ่มมากขึ้นในประเทศไทยเพราะขาดการได้รับความรู้ที่ถูกต้อง [1] ภาวะแทรกซ้อนที่พบบ่อยของภาวะหลอดเลือดดำอุดตันเกิดขึ้นระหว่างและหลังการรักษาในโรงพยาบาล เนื่องจากมีอาการเจ็บป่วยเฉียบพลันซึ่งจะมี 5-10% ของผู้ป่วยที่เสียชีวิตในโรงพยาบาล นอกจากนี้ภาวะลิ่มเลือดอุดตันในหลอดเลือดดำยังเป็นการรักษาระยะยาวของกลุ่มหลังเกิดอาการ ดังนั้นภาวะแทรกซ้อนเหล่านี้ทำให้เกิดโรคและมีต้นทุนในการรักษาเกิดขึ้น VTE เกิดจากการที่มีลิ่มเลือดไปขัดขวางการไหลเวียนของเลือดในหลอดเลือดดำ ซึ่งการจับตัวกันเป็นลิ่มเลือดจะขัดขวางการไหลเวียนของเลือดอย่างช้าๆ ประกอบด้วยภาวะหลอดเลือดดำชั้นลึกอุดตันที่ขา (Deep Vein Thrombosis, DVT) ซึ่งส่วนใหญ่เกิดที่หลอดเลือดดำชั้นลึกที่ขา และอาจเกิดลิ่มเลือดหลุดไปอุดหลอดเลือดแดงที่ปอดทำให้เกิดภาวะลิ่มเลือดอุดกั้นในหลอดเลือดแดงที่ปอด (Pulmonary Embolism, PE) เป็นสาเหตุให้เกิดการเสียชีวิตอย่างเฉียบพลันได้ งานวิจัยนี้จึงได้ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อสร้างแบบจำลองของการทำนาย (Prediction Model) ถึงปัจจัยเสี่ยงที่อาจจะทำให้เกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ เพื่อใช้ในการช่วยให้แพทย์ตัดสินใจในการวินิจฉัยและรักษาได้แม่นยำมากขึ้น

II. งานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ผู้วิจัยได้ศึกษางานวิจัยที่เกี่ยวข้องกับการพัฒนาแบบจำลองเกี่ยวกับภาวะการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำโดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning)

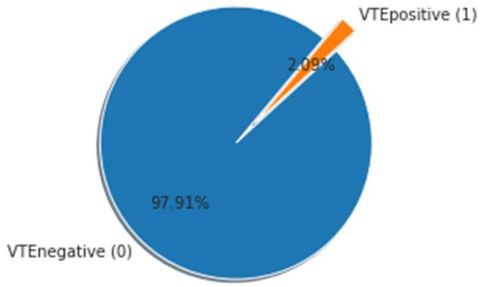
ศึกษาบทความวิจัยที่ใช้เทคนิคการเรียนรู้ของเครื่องดังต่อไปนี้ Agharezaei และคณะ [4] ได้ทำนายระดับความเสี่ยงของการเกิดลิ่มเลือดอุดตันที่ปอดของผู้ป่วย โดยใช้โครงข่ายประสาทเทียม 2 ชนิด ได้แก่ Feed-Forward Back Propagation และ Elman Back Propagation ด้วย software MATLAB ในการวิเคราะห์ข้อมูล ผลการศึกษานี้เสนอแบบจำลองโครงข่ายประสาทเทียมแสดงถึงความแม่นยำและคัชนีระดับความเสี่ยงที่ 93.23 เปอร์เซ็นต์ นอกจากนี้ Fei, Hu, Li และคณะ [5,6] ยังได้ใช้โครงข่ายประสาทเทียม (ANNs) ในการทำนายการเกิดลิ่มเลือดอุดตันของหลอดเลือดดำ (Portosplenomesenteric Venous Thrombosis, PSMVT) เปรียบเทียบความสามารถในการทำนายของ ANNs กับ logistic regression พบว่า ANNs ให้ความแม่นยำมากกว่า logistic regression ในการทำนายการเกิดลิ่มเลือดอุดตัน (PSMVT) หลังจากมีดับอ่อนอักเสบเฉียบพลัน ในงานวิจัยของ Qatawneh และคณะ [7] ยังเป็นงานวิจัยที่สร้างระบบช่วยสนับสนุนการตัดสินใจทางคลินิกของการจำแนกความเสี่ยงของการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ ได้ใช้ระบบโครงข่ายประสาทเทียม Multilayer Perceptron (MLP) โดยใช้ Resilient Backpropagation algorithm (Rprop) ผลลัพธ์ output ที่ได้ออกมาจะเป็นระดับคะแนนความเสี่ยงของการเกิดโรค โดยแบ่งเป็น 5 ระดับความเสี่ยง ได้แก่ low, lower-mild, higher-mild, moderate และ high เพื่อช่วยแพทย์ประกอบการตัดสินใจในการวินิจฉัยผู้ป่วย ซึ่งผลการประเมินประสิทธิภาพความถูกต้องของระบบดังกล่าวอยู่ที่ 81% นอกจากนี้ Ferroni Z.และคณะ[8,9] ได้ใช้เทคนิค Machine Learning (ML) และ Random Optimization (OR) พัฒนาในเรื่องของการทำนายความเสี่ยงของการเกิดหลอดเลือดอุดตันเพื่อที่จะเป็นแนวทางในการออกแบบเว็บอินเตอร์เฟสเพื่อแยกความเสี่ยงของการเกิดลิ่มเลือดอุดตันในผู้ป่วยมะเร็งที่ได้รับเคมีบำบัด และเปรียบเทียบประสิทธิภาพของทั้ง 2 เทคนิค ดังกล่าวกับเทคนิคที่ใช้ในปัจจุบันซึ่งก็คือ Khorana Score (KS) ผลลัพธ์ของงานวิจัยนี้พบว่าวิธี ML-RO มีประสิทธิภาพกว่าวิธี Khorana Score (KS) จึงเหมาะและเป็นประโยชน์ในการออกแบบ web service ที่สามารถอินเตอร์เฟสแก่แพทย์เพื่อช่วยในขั้นตอนการตัดสินใจรักษาผู้ป่วย Liu S.และคณะ [10] ใช้เทคนิค machine learning ประเมินประสิทธิภาพที่ระบุความสามารถในการทำนายความเสี่ยงของการเกิดลิ่มเลือดอุดตันในผู้ป่วยมะเร็งที่ใส่สายสวน PICC ได้ โดยใช้ 5 โมเดล ได้แก่ Seely, Seely-RF, Seely-LASSO-RF, RF และ LASSO-RF เพื่อเปรียบเทียบผลลัพธ์การประเมินประสิทธิภาพของโมเดล และเปรียบเทียบกับผลลัพธ์ของโมเดลที่ใช้อยู่ในปัจจุบัน (Seely) กับ machine learning พบว่า machine learning ให้

ประสิทธิภาพที่ดีกว่า เพื่อช่วยในการตัดสินใจในการป้องกันโรคและลดอัตราการเกิดลิ่มเลือดอุดตันในผู้ป่วยมะเร็งที่ใส่สายสวน PICC ได้อย่างมีประสิทธิภาพ และงานวิจัย Nafee T. และคณะ [11] ได้ใช้ Super learner model (ML) และ Reduced model (rML) เทียบกับ IMPROVE score (International Medical Prevention Registry on Venous Thromboembolism) ในการทำนายการเกิดลิ่มเลือดอุดตัน พบว่าวิธี Super learner model (ML) ได้ค่า c-statistic สูงสุดสำหรับการทำนายการเกิดลิ่มเลือดอุดตัน

ดังนั้นงานวิจัยนี้ได้ประยุกต์ใช้การเรียนรู้ของเครื่องเพื่อหาประสิทธิภาพที่ดีที่สุดสำหรับทำนายความเสี่ยงของการเกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ โดยรวบรวมข้อมูลจากหอผู้ป่วยในโรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย เพื่อช่วยแพทย์ในการตัดสินใจและวินิจฉัยในการรักษาที่แม่นยำยิ่งขึ้น ส่วนที่เหลือของงานวิจัยนี้จะกล่าวดังต่อไปนี้ : ส่วนที่ 3 อธิบายถึงการรวบรวมข้อมูล ส่วนที่ 4 วิธีการดำเนินงานวิจัย ส่วนที่ 5 แสดงผลลัพธ์ของงานวิจัย สุดท้ายส่วนที่ 6 สรุปและอภิปรายผลงานวิจัย

III. ชุดข้อมูล

งานวิจัยนี้นำข้อมูลที่ถูกเก็บรวบรวมโดยแพทย์ผู้ทำวิจัยหาปัจจัยเสี่ยงของการเกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ โดยเก็บข้อมูลจากจากผู้ป่วยที่เข้ารับการรักษาในหอผู้ป่วยแผนกอายุรศาสตร์ ของโรงพยาบาลจุฬาลงกรณ์ ภายในปี 2009 ที่ได้รับการอนุมัติจากคณะกรรมการจริยธรรมของคณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยแล้ว ซึ่งข้อมูลประกอบด้วยทั้งหมด 1290 แถว คือจำนวนผู้ป่วยที่ถูกสำรวจ และมี 65 คอลัมน์ คือปัจจัยต่างๆ ที่นำมาประกอบการพิจารณาหาความเสี่ยงของการเกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ ซึ่งจะรวมถึงคอลัมน์ที่เป็นคำตอบ (Label) ด้วย



รูปที่ 1 แสดงจำนวนผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตัน (VTE positive) และไม่เป็นโรคลิ่มเลือดอุดตัน (VTE negative) จากข้อมูลทั้งหมด 1290 แถว 65 คอลัมน์

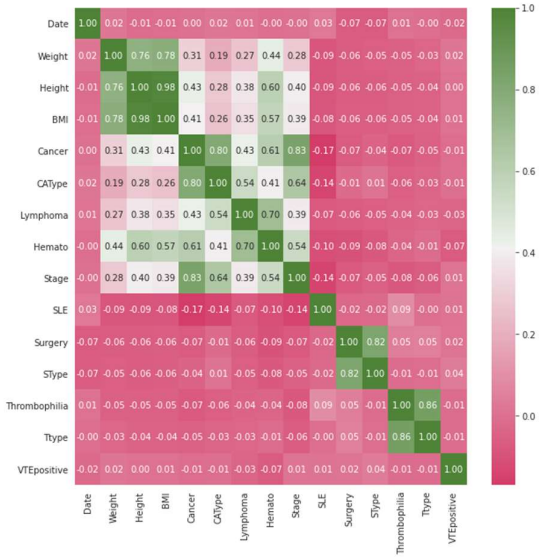
จากรูปที่ 1 กำหนดให้ 0 คือ ผู้ป่วยที่ไม่เป็นโรคลิ่มเลือดอุดตันในหลอดเลือดดำ (VTE negative) มี 2.09% หรือ 27 ราย ของข้อมูลทั้งหมด และ 1 คือ ผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตันในหลอดเลือดดำ (VTE positive) มี 97.91% หรือ 1263 ราย ของข้อมูลทั้งหมด

IV. โมเดลและวิธีการ

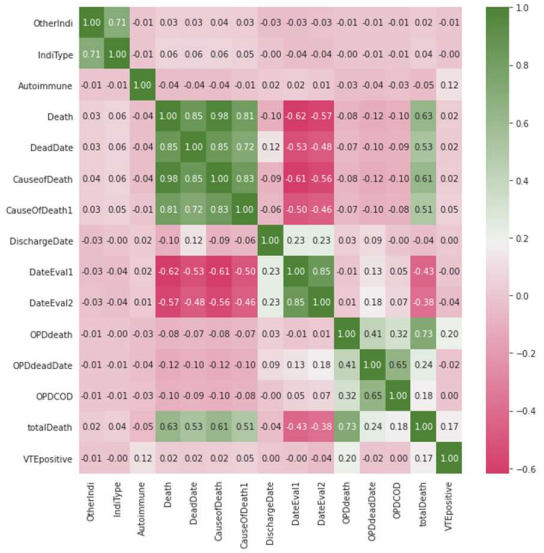
งานวิจัยนี้สร้างแบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง มีหลักการสร้างและเปรียบเทียบประสิทธิภาพของแบบจำลองดังต่อไปนี้

A. สำรวจข้อมูล Exploratory Data Analysis (EDA)

วิเคราะห์ความสัมพันธ์ระหว่างข้อมูลปัจจัยเสี่ยงที่ทำให้เกิดภาวะลิ่มเลือดอุดตันทั้งหมดว่าสัมพันธ์กันในระดับใด สัมพันธ์กันมากน้อยเพียงใด พบว่าตัวแปรแต่ละคอลัมน์มีความสัมพันธ์กันน้อยมากเมื่อเทียบกับ Class Label เนื่องจากข้อมูลทั้งหมดที่แพทย์ผู้เชี่ยวชาญได้เก็บข้อมูลไว้นั้นเป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Variable) เช่น เทียบกลุ่มอายุ >40 ปี หรือ <40 ปี ไม่ได้ใช้อายุที่เป็นปีจริงๆ กับการเป็น VTE positive และ VTE negative เป็นต้น จึงทำให้ไม่มีความสัมพันธ์กันระหว่างตัวแปรในแต่ละคอลัมน์กับ Class Label จากข้อมูลทั้งหมด 1290 แถว 65 คอลัมน์ ผู้วิจัยได้เลือกพลอตกราฟเฉพาะคอลัมน์ที่มีความสัมพันธ์กันมากที่สุดจากข้อมูลทั้งหมด ดังรูปที่ 2

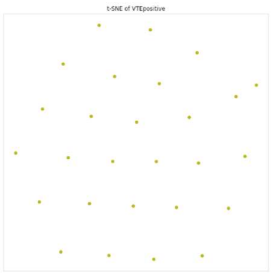


รูปที่ 2 แสดงความสัมพันธ์ของข้อมูลปัจจัยเสี่ยงของการเกิดลิ่มเลือดอุดตัน



รูปที่ 2 (ต่อ) แสดงความสัมพันธ์ของข้อมูลปัจจัยเสี่ยงของการเกิดลิ่มเลือดอุดตัน

เนื่องจากข้อมูลนี้เป็นปัญหาข้อมูลไม่สมดุลกัน (Imbalanced data) ผู้วิจัยจึงได้ทำการสำรวจความกระจายของข้อมูลเฉพาะผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตันในหลอดเลือดดำ ซึ่งมี 27 ราย ของข้อมูลทั้งหมด โดยแสดงผลข้อมูลด้วยวิธี t-distributed Stochastic Neighbor Embedding : t-SNE ซึ่งเป็นวิธีการลดมิติของข้อมูลให้สามารถวิเคราะห์ออกมาง่ายขึ้น ดังรูปที่ 3



รูปที่ 3 แสดงข้อมูลของผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตันในหลอดเลือดดำ 27 ราย

จากนั้นดูความสัมพันธ์ของข้อมูลคนที่เป็นโรคลิ่มเลือดอุดตันกับปัจจัยเสี่ยงทั้งหมด พบว่าปัจจัยเสี่ยงที่จะเป็นโรคลิ่มเลือดอุดตันมีหลากหลายปัจจัยมาก ซึ่งคนที่จะเป็นโรคนี้เกิดจากหลายปัจจัยไม่มีปัจจัยที่เฉพาะเจาะจง แสดงผลข้อมูลด้วยวิธี t-distributed Stochastic Neighbor Embedding : t-SNE ดังรูปที่ 4 ซึ่งมีความกระจายตัวกันมาก

รูปที่ 4 แสดงความสัมพันธ์ของผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตันเทียบกับข้อมูลทั้งหมด

B. การเตรียมข้อมูล (Preparing Data)

- จัดการกับข้อมูลที่ขาดหายไป (Missing Values) พบข้อมูลที่ขาดหายไปจำนวนมาก ได้จัดการกับข้อที่ขาดหายไปด้วยหลักการเติมศูนย์ (Fill missing with zero) เพราะข้อมูลคนไข้ไม่สามารถแทนที่กันได้
- จัดการกับชนิดข้อมูลที่เป็นตัวอักษรให้อยู่ในรูปแบบตัวเลขด้วยวิธี Label encoder เพื่อให้เหมาะกับนำไปสร้างแบบจำลอง
- แบ่งข้อมูลสำหรับฝึกสอน (Train data) และสำหรับทดสอบ (Test data) ด้วยอัตราส่วน 70:30 จะได้ 903:387 ของข้อมูลที่นำมาวิเคราะห์
- จัดการกับข้อมูลที่ไม่สมดุลกัน เนื่องจากปัญหาของข้อมูลที่นำมาวิเคราะห์นี้เป็นปัญหาที่ข้อมูลไม่สมดุลกัน ผู้วิจัยจึงได้จัดการกับปัญหาดังกล่าวด้วย 5 วิธี ดังต่อไปนี้
    - วิธีสุ่มเกิน หรือ Oversampling เป็นวิธีเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนที่อยู่ในกลุ่มส่วนมาก
    - วิธีสุ่มลด หรือ Undersampling เป็นการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนที่อยู่ในกลุ่มส่วนน้อย
    - วิธีสังเคราะห์ข้อมูลเพิ่ม หรือ Synthetic Minority Oversampling TEchnique : SMOTE เป็นเทคนิคการสุ่มตัวอย่างแบบสุ่มเพิ่มแทนที่จะสุ่มเพิ่มจากข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิม หลักการคือการสุ่มจากข้อมูลเพื่อนบ้านที่อยู่ใกล้ที่สุด
    - วิธี Class weight คือ การให้น้ำหนักในแต่ละ Class ซึ่งจะให้ความสำคัญกับข้อมูลกลุ่มส่วนน้อยมากกว่า เพื่อจะให้ผลลัพธ์ที่ได้นั้นมาจากการเรียนรู้จากทุกกลุ่มที่เท่าเทียมกัน
    - วิธี Ensemble sampling โดยใช้ Balanced Bagging Classifier คือ ความน่าจะเป็นของการสุ่มตัวอย่างของ Training data ที่จะถูกทำให้กระจายไปตามกลุ่มที่อยู่ใกล้เคียงกันทำให้ข้อมูลสมดุลขึ้น

C. การสร้างโมเดล (Modeling)

เนื่องจากงานวิจัยนี้เป็นปัญหาแบบ classification ผู้วิจัยได้เลือกแบบจำลองการทำนายทั้งหมด 3 เทคนิค เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง ดังต่อไปนี้

- เทคนิค Decision Tree เป็นเทคนิคต้นไม้ตัดสินใจแบบต้นเดียวโดดๆ ที่เรียนรู้จากคุณลักษณะของข้อมูลแล้วสร้างผังการตัดสินใจคล้ายๆกับต้นไม้ซึ่ งจะพยายามแบ่งข้อมูลออกมาเป็นคลาสๆให้ชัดเจนได้มากที่สุด หลักการสร้างโมเดล คือจะเลือกตัวแปรที่สามารถแบ่งแยกคลาสคำตอบได้ชัดเจนมากที่สุด
- เทคนิค Logistic Regression นิยมใช้กับปัญหา Binary classification ทำนายคลาสที่มี 2 คลาส มี Sigmoid Function เพื่อมา normalize $Sigmoid\ (Z) = \frac{e^z}{1+e^z}$ ผลลัพธ์ที่ได้จาก Sigmoid Function คือความน่าจะเป็นที่สามารถทำนายคลาสได้
- เทคนิค Neural Network เป็นเทคนิคที่จำลองการทำงานของสมองมนุษย์ มี 3 layers ได้แก่ Input layer, Hidden layer และ Output layer

D. การประเมินผล (Evaluate)

งานวิจัยนี้เป็นการทำนายผลลัพธ์ที่เป็น classification ในทางการแพทย์ จึงจำเป็นต้องมีการประเมินผลลัพธ์ให้แม่นยำที่สุด โดยใช้ตาราง Confusion matrix ในประเมินผลของแบบจำลอง ดังต่อไปนี้

ตารางที่ 1 แสดงตาราง Confusion matrix สำหรับประเมินผลของแบบจำลอง

		ทำนาย (Prediction)	
		Negative (0)	Positive (1)
ค่าจริง (Actual)	Negative (0)	True Negative (TN)	False Positive (FP)
	Positive (1)	False Negative (FN)	True Positive (TP)

จากตารางที่ 1

TN คือ จำนวนข้อมูลที่เครื่องทำนายไม่จริง และถูกต้องตามข้อมูลจริง

TP คือ จำนวนข้อมูลที่เครื่องทำนายจริง และถูกต้องตามข้อมูลจริง

FN คือ จำนวนข้อมูลที่เครื่องทำนายไม่จริง และไม่ถูกต้องตามข้อมูลจริง

FP คือ จำนวนข้อมูลที่เครื่องทำนายจริง และไม่ถูกต้องตามข้อมูลจริง

- Accuracy คือเปอร์เซ็นความถูกต้อง คือจำนวนที่ทำนายถูก/จำนวนทั้งหมด

$$Accuracy = \frac{TP + TN}{All}$$

- Recall คือความแม่นยำที่สนใจผลลัพธ์เทียบกับที่เป็นของจริง (Actual)

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

- Precision คือความแม่นยำที่สนใจผลการทำนาย (Prediction)

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

V. การทดลองและผลลัพธ์

การทดลองของงานวิจัยนี้ได้สร้างแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง 3 เทคนิค ได้แก่ Decision Tree, Logistic Regression และ Neural Network ผลลัพธ์ของการสร้างแบบจำลอง ได้เปรียบเทียบผลของการประเมินแบบจำลองแต่ละแบบจำลองกับวิธีปรับความสมดุลของข้อมูลแต่ละวิธี แสดงดังตารางที่ 2

ตารางที่ 2 แสดงผลลัพธ์ของแบบจำลองการทำนาย 3 เทคนิคพร้อมเปรียบเทียบวิธีปรับความสมดุลของข้อมูลแต่ละวิธี

Balance method / Classifier		DecisionTree	LogisticRegression	NeuralNetwork
Over_sampling	% accuracy	96.1	76.0	76.5
	% recall	18.2	6.2	5.4
	% precision	25.0	75.0	62.5
Under_sampling	% accuracy	81.1	56.8	38.8
	% recall	7.8	3.5	2.5
	% precision	75.0	75.0	75.0
SMOTE	% accuracy	96.1	80.4	93.0
	% recall	18.2	6.4	12.0
	% precision	25.0	62.5	37.5
Class_weight	% accuracy	96.6	76.2	-
	% recall	22.2	6.3	-
	% precision	25.0	75.0	-
Ensemble_resampling	% accuracy	76.7	73.1	69.8
	% recall	5.4	5.6	5.0
	% precision	62.5	75.0	75.0

หมายเหตุ : วิธี Class Weight จะไม่สามารถวัดได้กับเทคนิค Neural Network

จากผลการทดลองพบว่าแบบจำลอง Decision Tree ให้ค่าความถูกต้องสูงสุด 96.6% ด้วยการปรับความสมดุลของข้อมูลด้วยวิธี Class Weight และในทางการแพทย์จะดูค่า Recall เป็นหลักเพราะผลลัพธ์ต้องเปรียบเทียบกับค่าความเป็นจริง ซึ่งจากตารางที่ 2 ค่า Recall ของแต่ละแบบจำลองให้ผลได้ไม่ดี จึงทำให้ประสิทธิภาพของแต่ละแบบจำลองไม่เหมาะสมที่จะเป็นแบบจำลองที่ดีได้

VI. สรุปผลและอภิปรายผลการทดลอง

งานวิจัยนี้ได้สร้างแบบจำลองการวินิจฉัยแบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง จากการเก็บรวบรวมข้อมูลผู้ป่วยในหอผู้ป่วยอายุรศาสตร์ โรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย งานวิจัยนี้ได้เตรียมข้อมูลทั้งหมด 1290 แถว 65 คอลัมน์และตรวจสอบค่าที่ขาดหายไป พร้อมทั้งแปลงข้อมูลให้พร้อมสำหรับนำไปสร้างแบบจำลองการทำนายในลำดับต่อไป จากนั้นแบ่งข้อมูลสำหรับการฝึกสอนและข้อมูลสำหรับการทดสอบในอัตราส่วน 70:30 จัดการกับปัญหาข้อมูลที่ไม่สมดุลทั้งหมด 5 วิธี ได้แก่ Oversampling, Undersampling, Synthetic Minority Oversampling TEchnique : SMOTE, Class weight และ Ensemble sampling แล้วทำการประเมินผลด้วยตาราง Confusion matrix ผลการทดลองของงานวิจัยได้เปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง 3 เทคนิค ได้แก่ Decision Tree, Logistic Regression และ Neural Network แบบจำลอง Decision Tree ให้ค่าความถูกต้องสูงสุด 96.6% ด้วยการปรับความสมดุลของข้อมูลด้วยวิธี Class Weight

จากตารางที่ 2 ค่า Accuracy จะบอกถึงสัดส่วนเปอร์เซ็นความถูกต้องของการทำนาย Precision จะสนใจเฉพาะผลการทำนายคือแบบจำลองทำนายมานั้นถูกต้องกี่เปอร์เซ็น และในส่วนของค่า Recall ปัญหาทางการแพทย์จะดูค่า Recall เป็นหลักเพราะผลลัพธ์ต้องเปรียบเทียบกับค่าความเป็นจริง ซึ่งจากตารางที่ 2 ค่า Recall ของแต่ละแบบจำลองให้ผลได้ไม่ดี ซึ่งปัญหาอาจเกิดจากการมีข้อมูลที่นำมาวิเคราะห์มีน้อยเกินไป อนาคตอาจจะต้องมีข้อมูลที่มากขึ้นกว่านี้เพื่อผลการทำนายจะมีประสิทธิภาพมากขึ้นและเหมาะที่จะนำไปพัฒนาแบบจำลองเพื่อช่วยให้แพทย์วินิจฉัยการเกิดโรคลิ่มเลือดอุดตันได้มีประสิทธิภาพและแม่นยำมากขึ้น

กิตติกรรมประกาศ

งานวิจัยนี้ขอขอบพระคุณโรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย ที่ให้การสนับสนุนข้อมูลที่ดีและมีประสิทธิภาพ ในงานวิจัยครั้งนี้ และขอขอบพระคุณสาขาวิทยาการข้อมูล ภาควิชาคอมพิวเตอร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ให้การสนับนุนเป็นอย่างดี

อ้างอิง

[1] Rojnuckarin, P. et al. “Risk factors for symptomatic venous thromboembolism in Thai hospitalised medical patients.” Thrombosis and haemostasis vol. 106, no. 6, 2011, 1103-1108.

[2] Cohen, A. T. et al. “Venous thromboembolism risk and prophylaxis in the acute hospital care setting (ENDORSE study): a multinational cross-sectional study.” Lancet (London, England) vol. 371, no. 9610, 2008, pp. 387-394.

[3] Motwani M., et al. “Advanced cardiovascular magnetic resonance myocardial perfusion imaging: high-spatial resolution versus 3-dimensional whole-heart coverage,” Circ Cardiovasc Imaging. Vol. 6, no. 2, 2013, pp. 339–348.

[4] Agharezaei, L., Agharezaei, Z., Nemati, A., Bahaadinbeigy, K., Keynia, F., Baneshi, M., Agharezaei, a. “The Prediction of the Risk Level of Pulmonary Embolism and Deep Vein Thrombosis through Artificial Neural Network,” Acta Informatica Medica, vol. 24, no. 5, 2016, pp. 354-359.

[5] Fei, Y., Hu, J., Gao, K., Tu, J., Li, W. Q., & Wang, W., “Predicting risk for portal vein thrombosis in acute pancreatitis patients: A comparison of radical basis function artificial neural network and logistic regression models,” J Crit Care, vol. 39, 2017, pp. 115-123.

[6] Fei, Y., Hu, J., Li, W. Q., Wang, W., & Zong, G. Q., “Artificial neural networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis,” J Thromb Haemost, vol. 15, no. 3, pp. 439-445.

[7] Qatawneh, Z., Alshraideh, M., Almasri, N., Tahat, L., & Awidi, A. , “Clinical decision support system for venous thromboembolism risk classification,” Applied Computing and Informatics, vol. 15, no. 1, 2019, pp. 12-18.

[8] Ferroni, P., Zanzotto, F. M., Scarpato, N., Riondino, S., Guadagni, F., & Roselli, M., “Validation of a Machine Learning Approach for Venous Thromboembolism Risk Prediction in Oncology,” Dis Markers, 2017, pp. 1-7.

[9] Ferroni, P., Zanzotto, F. M., Scarpato, N., Riondino, S., Nanni, U., Roselli, M., & Guadagni, F., “Risk Assessment for Venous Thromboembolism in Chemotherapy-Treated Ambulatory Cancer Patients,” Med Decis Making, vol. 37, no. 2, 2017, pp. 234-242.

[10] Liu, S., Zhang, F., Xie, L., Wang, Y., Xiang, Q., Yue, Z., Yu, C., “Machine learning approaches for risk assessment of peripherally inserted Central catheter-related vein thrombosis in hospitalized patients with cancer,” Int J Med Inform, vol. 129, 2019, pp. 175-183.

[11] Nafee, T., Gibson, C. M., Travis, R., Yee, M. K., Kerneis, M., Chi, G., Goldhaber, S. Z., “Machine learning to predict venous thrombosis in acutely ill medical patients,” Res Pract Thromb Haemost, vol. 4, no. 2, 2020, pp. 230-237.