EFFICIENCY IN PREDICTING THE RISK OF DIABETES JOINTLY WITH THE RISK OF HYPERTENSION USING DEEP LEARNING FOR MULTI-LABEL CLASSIFICATION

Watsa SUDJAI[1]

1 Faculty of Commerce and Accountancy, Chulalongkorn University, Thailand;
6580283426@cbs.chula.ac.th (Corresponding Author)

ARTICLE HISTORY

Received: 18 October 2024 **Revised:** 1 November 2024 **Published:** 15 November 2024

ABSTRACT

In the medical field, deep learning is widely used to create predictive models, which outperform traditional models. However, patients can suffer from multiple diseases simultaneously. To address this, deep learning has been developed to predict multiple diseases at once, known as multi-label classification neural networks. Although neural networks have shown great potential for prediction, they still face challenges with limited data. This research aims to improve the overall performance of the model by studying the relationship between labels using data on diabetes and hypertension, which often co-occur. The experiments were divided into two parts: simulated data and real-world data to compare multi-label and single-label feedforward neural networks. The results showed that multi-label neural networks performed well theoretically when tested on simulated data. However, in real-world data, using related labels did not significantly reduce the loss function but had the advantage of mitigating overfitting and maintaining comparable predictive performance to using a single label.

Keywords: Deep Learning, Single-Label Neural Network, Multi-Label Neural Network

CITATION INFORMATION: Sudjai, W. (2024). Efficiency in Predicting the Risk of Diabetes Jointly with the Risk of Hypertension Using Deep Learning for Multi-Label Classification. *Procedia of Multidisciplinary Research, 2*(11), 18

ประสิทธิภาพของการทำนายความเสี่ยงการเกิดโรคเบาหวานร่วมกับความเสี่ยงการเกิดโรคความดันโลหิตสูงด้วยวิธีการเรียนรู้เชิงลึกสำหรับการจำแนกประเภทหลายเลเบล

วรรษา สุดใจ[1]

1 คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย; 6580283426@cbs.chula.ac.th (ผู้ประพันธ์บรรณกิจ)

บทคัดย่อ

ในทางการแพทย์ การเรียนรู้เชิงลึกนิยมนำมาใช้ในการสร้างตัวแบบทำนายซึ่งค่อนข้างให้ผลที่ดีกว่าเมื่อเทียบกับตัวแบบดั้งเดิมแต่บางครั้งผู้ป่วยสามารถเป็นโรคพร้อมกันได้มากกว่าหนึ่งโรค การเรียนรู้เชิงลึกจึงถูกพัฒนาให้สามารถทำนายพร้อมกันได้หลายโรค เรียกว่าโครงข่ายประสาทเทียมประเภทจำแนกหลายเลเบล ถึงแม้โครงข่ายประสาทเทียมจะมีความสามารถที่ดีสำหรับการทำนายแต่ยังมีความท้าทายในข้อมูลบางกลุ่มที่ข้อมูลมีจำกัด งานวิจัยนี้จึงมีจุดประสงค์ที่จะทำการศึกษาโดยต้องการเพิ่มประสิทธิภาพของตัวแบบโดยรวมด้วยการใช้เลเบลผลลัพธ์ที่เกี่ยวข้องกันมาศึกษาผ่านข้อมูลโรคเบาหวานและโรคความดันโลหิตสูงซึ่งเป็นโรคที่มักเกิดร่วมกัน แบ่งการทดลองเป็นสองส่วนคือส่วนข้อมูลจำลองและข้อมูลจริงเพื่อเปรียบเทียบระหว่างโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าหลายเลเบลกับหนึ่งเลเบล ผลการศึกษาพบว่าในโครงข่ายประสาทเทียมหลายเลเบลให้ผลที่ดีในทางทฤษฎีที่ทดสอบกับข้อมูลจำลอง แต่ในข้อมูลจริงผลลัพธ์ของการใช้เลเบลที่มีความเกี่ยวข้องกันไม่สามารถลดค่าฟังชันการสูญเสียได้อย่างมีนัยสำคัญ แต่มีข้อดีคือช่วยลดความรุนแรงของปัญหา overfit ได้และสามารถให้ประสิทธิภาพการทำนายยังคงเทียบเท่าการใช้หนึ่งเลเบล

คำสำคัญ: การเรียนรู้เชิงลึก, โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าหลายเลเบล, โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าหนึ่งเลเบล

ข้อมูลการอ้างอิง: วรรษา สุดใจ. (2567). ประสิทธิภาพของการทำนายความเสี่ยงการเกิดโรคเบาหวานร่วมกับความเสี่ยงการเกิดโรคความดันโลหิตสูงด้วยวิธีการเรียนรู้เชิงลึกสำหรับการจำแนกประเภทหลายเลเบล. *Procedia of Multidisciplinary Research, 2*(11), 18

บทนำ

จากข้อมูลสำรวจสุขภาพประชาชนไทย ปี 2019 จากกรมควบคุมโรคกองโรคไม่ติดต่อ พบว่า ประเทศไทยมีผู้ป่วยเป็นโรคเบาหวาน 5 ล้านคน และ 1 ใน3 ของผู้ป่วยที่โรคเบาหวาน ไม่เคยทราบมาก่อนว่าเป็นโรค อีกทั้งพบว่า14 ล้านคนเป็นโรคความดันโลหิตสูง ซึ่งร้อยละ 48.8 ของผู้ที่เป็นโรคความดันโลหิตสูง ไม่ทราบว่าตนเองเป็นโรคมาก่อน โดยโรคทั้งสองเป็นโรคไม่ติดต่อชนิดเรื้อรัง สามารถทำให้เกิดโรคแทรกซ้อนที่อันตราย เช่น โรคหลอดเลือดหัวใจ โรคหลอดเลือดสมอง อาจนำไปสู่การเกิดภาวะอัมพฤกษ์ อัมพาต หรือการเสียชีวิต และมีค่าใช้จ่ายในการรักษาที่สูง เมื่อเป็นแล้วไม่สามารถรักษาให้หายขาดได้ การนำตัวแบบมาช่วยในการระบุผู้ที่มีความเสี่ยงในการเป็นโรค เป็นการช่วยให้แพทย์สามารถเริ่มการรักษาและป้องกันโรคแทรกซ้อนที่เกิดจากโรคเบาหวานและโรคความดันโลหิตสูงได้เร็ว และผู้ป่วยสามารถปรับเปลี่ยนพฤติกรรมเพื่อป้องกันภาวะแทรกซ้อนจากโรคได้ ปัจจุบันโดยเฉพาะในทางการแพทย์ การเรียนรู้เชิงลึกถูกนำมาใช้อย่างแพร่หลายในการสร้างตัวแบบทำนาย ซึ่งให้ผลที่ดีเมื่อเทียบกับตัวแบบดั้งเดิม ซึ่งโดยทั่วไปข้อมูลทางการแพทย์มักมีลักษณะเป็นข้อมูลแบบ Person-period คือ ข้อมูลจะมีหลายแถวต่อหนึ่งคนโดยแต่ละแถวจะเป็นข้อมูลในช่วงเวลาที่ทำการสังเกต ในการทำนายมักจะสนใจว่านานเท่าใด ถึงจะมีเหตุการณ์ที่สนใจเกิดขึ้น เช่น การเกิดโรคหรือ ตาย เป็นต้น ซึ่งข้อมูลจะมีลักษณะเป็นเวลาไม่ต่อเนื่องมักนิยมใช้การวิเคราะห์การรอดชีพแบบไม่ต่อเนื่องในการทำนายเพราะสามารถนำอัลกอริทึ่มในการเรียนรู้ของเครื่องที่สามารถจำแนกประเภททวิภาคมาใช้ได้ จากงานวิจัยของ Suresh et al. (2022) ทำการศึกษาเกี่ยวกับตัวแบบการทำนายการรอดชีพแบบเวลาไม่ต่อเนื่อง พบว่า โครงข่ายประสาทเทียมมีประสิทธิภาพในการทำนายดีที่อยู่ในลำดับต้นท่ามกลางตัวแบบอื่น ซึ่งโครงข่ายประสาทเทียมได้ถูกนำมาใช้ในการทำนายโรคเพิ่มขึ้นแต่มีข้อจำกัดคือ สามารถทำนายได้เพียงหนึ่งโรคเท่านั้น บางครั้งผู้ป่วยสามารถเป็นโรคพร้อมกันได้หลายโรค ภายหลังจึงมีการพัฒนาโครงข่ายประสาทเทียมให้สามารถทำนายพร้อมกันได้หลายค่าเรียกว่าโครงข่ายประสาทเทียมประเภทจำแนกหลายเลเบล จากการศึกษาของ Maxwell et al. (2017) ที่ศึกษาการทำนายแบบการจำแนกหลายเลเบลด้วย การเรียนรู้เชิงลึกสำหรับความเสี่ยงต่อสุขภาพ นำมาเปรียบเทียบกับการจำแนกแบบหลายเลเบลวิธีการทั่วไป ผลลัพธ์แสดงให้เห็นว่า โครงข่ายประสาทเทียมเชิงลึก (Deep Neural Network) ให้ประสิทธิภาพที่ดีกว่าเมื่อเทียบกับตัวแบบอื่น และแนะนำว่าควรค่าแก่การนำมาศึกษาต่อ ถึงแม้โครงข่ายประสาทเทียมจะมีความสามารถที่ดีสำหรับการทำนายแต่ยังคงมีความท้าทายในข้อมูลบางกลุ่มที่ข้อมูลมีจำกัด การนำแนวคิดใหม่มาใช้สำหรับการเพิ่มประสิทธิภาพภายใต้ข้อมูลและทรัพยากรที่มีอยู่อย่างจำกัดจึงมีความสำคัญ ซึ่งข้อมูลทางการแพทย์มักจะพบข้อมูลที่มีหลายเลเบลที่แสดงถึงผลลัพธ์ที่ได้มากจากหน่วยสังเกตเดียวกัน โดยทั่วไปผลลัพธ์เหล่านี้มักจะได้รับอิทธิพลมาจากตัวแปรอิสระที่ใช้ร่วมกัน และข้อมูลที่ถูกใช้ร่วมกันระหว่างเลเบลอาจมีความสามารถในการนำมาใช้ประโยชน์สำหรับพัฒนาแบบจำลองเพื่อเพิ่มประสิทธิภาพได้มากขึ้น แนวคิดนี้ถูกเสนอและทำการทดลองโดย Kiatsupaibul and Chantarasiripas (2024) ได้เสนอแนวคิด Single-response Analogy เป็นการมองว่าโครงข่ายประสาทเทียมแบบหลายเลเบลมีโครงสร้างเหมือนกับโครงข่ายประสาทเทียมแบบหนึ่งเลเบลแต่มีการใช้ข้อมูลเพิ่มขึ้นเนื่องจากต้องใช้ข้อมูลเดิมในการเรียนรู้เลเบลเพิ่ม ซึ่งทำการทดลองด้วยการสร้าง Constrain bi-response ที่กำหนดให้น้ำหนักและความเอนเอียงที่เลเบลผลลัพธ์มีขนาดเท่ากัน และทำการเปรียบเทียบประสิทธิภาพตัวแบบโดยวัดผลจากหนึ่งเลเบลที่สนใจเท่านั้นเทียบกับตัวแบบ Single-response ซึ่งพบว่า ในทางทฤษฎีตัวแบบโครงข่ายประสาทเทียมแบบหลายเลเบลสามารถเพิ่มประสิทธิภาพได้จากการที่ฟังก์ชันการสูญเสียลดลงเร็วขึ้น งานวิจัยนี้จึงมีจุดมุ่งหมายที่จะทำการเพิ่มประสิทธิภาพของตัวแบบด้วยเลเบลผลลัพธ์ที่เกี่ยวข้องกันซึ่งได้มาจากการวัดผลมาจากคนเดียวกัน โดยทำการทดลองในข้อมูลโรคเบาหวานและโรคความดันโลหิตสูง เพราะโรคเบาหวานและโรคความดันโลหิตสูงเป็นโรคที่มักพบร่วมกัน มีงานวิจัยที่ศึกษาการทำนายโรคเบาหวานและความดันโลหิตสูงร่วมกัน พบว่าการพัฒนาของความดันโลหิตสูงและโรคเบาหวานประเภท 2 มีแนวโน้มที่จะเกิดขึ้นควบคู่กันไปตามกาลเวลา (Tsimihodimos et al., 2018) และจากงานของ Zhou et al. (2021) ทำการทดลองนำความสัมพันธ์ของโรคในตัวแบบหลายเลเบลมาปรับปรุงประสิทธิภาพ ซึ่งพบว่า ค่าสหสัมพันธ์จากการเรียนรู้ของเครื่อง(Machine Learning) มีความสอดคล้องกับค่าสหสัมพันธ์

ที่ได้จากข้อมูลจริง และประสิทธิภาพการทำนายที่ได้จากตัวแบบการเรียนรู้ของเครื่องแบบหลายเลเบล ให้ค่าที่เหนือกว่าโมเดลการจำแนกประเภทแบบไบนารีแบบดั้งเดิมในการทำนายภาวะแทรกซ้อนจากโรคเบาหวาน แสดงว่าการเรียนรู้ของเครื่องมีความสามารถในการใช้ความสัมพันธ์ของโรคแทรกซ้อนจากเบาหวานในการส่งเสริมการทำนายให้มีประสิทธิภาพดีขึ้นได้ ซึ่งวิธีการเรียนรู้เชิงลึกสำหรับการจำแนกประเภทหลายเลเบลก็เป็นส่วนหนึ่งของการเรียนรู้ของเครื่อง จึงมีความเป็นไปได้ที่จะสามารถใช้ความสัมพันธ์ของโรคเบาหวานและโรคความดันโลหิตสูงมาเพิ่มประสิทธิภาพได้ งานวิจัยนี้จึงมุ่งเน้นไปที่การศึกษาศักยภาพของเลเบลผลลัพธ์ในการใช้ข้อมูลที่เกี่ยวข้องกันเพื่อลดฟังก์ชันความสูญเสียภายใต้การการทดลองในการทำนายโรคเบาหวานเมื่อใช้ฝึกร่วมกับโรคความดันโลหิตสูงด้วยวิธีโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าประเภทหลายเลเบล เทียบกับโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าหนึ่งเลเบล รวมถึงศึกษาประสิทธิภาพการทำนายความเสี่ยงในการเกิดโรค

วิธีดำเนินการวิจัย

ในการทำการทดลองแบ่งการทดลองเป็นสองแบบคือการทดลองในข้อมูลจำลองและข้อมูลจริงโรคเบาหวานและโรคความดันโลหิตสูง มีขั้นตอนการดำเนินการทดลอง ดังนี้

การจำลองข้อมูล

ทำการจำลองข้อมูล เพื่อเปรียบเทียบประสิทธิภาพของตัวแบบ โดยจะทำการจำลองข้อมูล ดังนี้

1) กำหนดข้อมูลที่มีขนาดตัวอย่าง 2000 ตัวอย่างและกำหนดให้ตัวแปรอิสระ 5 ตัว คือ $x_1, x_2, \dots, x_5 \overset{iid}{\sim} N(0,1)$

2) จำลองข้อมูลตัวแปรตาม 2 ตัวแปร คือ y_1, y_2 ด้วยโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าหลายเลเบล ซึ่งมีโครงสร้างชั้นที่ซ่อนอยู่ 2 ชั้น จำนวนโหนด 10 และ 5 ตามลำดับ และมีชั้นผลลัพธ์ 2 โหนด

3) กำหนดให้สุ่มน้ำหนักของแต่ละโหนด คือ $w_i \overset{iid}{\sim} N(0,1)$ และกำหนดให้ชั้นสุดท้าย $w_i \overset{iid}{\sim} N(0,0.1)$ เพื่อให้ผลทำนายที่ออกมามีค่าในแต่ละคลาสมีจำนวนข้อมูลที่มีความสมดุล และให้ความเอนเอียงในแต่ละชั้นมีค่าเป็น 0

4) ฝึกฝนตัวแบบจนได้ค่าทำนายจากตัวแบบ คือ y_1^*, y_2^* ซึ่งมีลักษณะเป็นข้อมูลต่อเนื่อง

5) ตัดให้ค่าทำนายให้มีลักษณะเป็นทวิภาค โดยสุ่มเลขจาก $U_\ell \overset{iid}{\sim} Unif(0,1)$ และแบ่งค่าพยากรที่ได้ดังนี้

$$y_\ell = \begin{cases} 1 \text{ เมื่อ } y_\ell^* > U_\ell \\ 0 \text{ เมื่อ } y_\ell^* < U_\ell \end{cases}$$

เมื่อ

y_ℓ เป็นตัวแปรตามไบนารี เมื่อ $\ell = 1,2$

y_ℓ^*เป็นตัวแปรตามที่ได้จากการฝึกฝนตัวแบบ เมื่อ $\ell = 1,2$

U_ℓ เป็นค่าที่ได้จากการสุ่ม เมื่อ $i = 1,2$

6) สร้างชุดข้อมูลสำหรับทำการทดลองทั้งหมด 50 ชุดข้อมูล โดยแต่ละชุดข้อมูลจะมีการสุ่มข้อมูลที่ต่างกัน

การเตรียมข้อมูลจริง

ในงานวิจัยนี้ใช้ชุดข้อมูลจริง โรคเบาหวานและโรคความดันโลหิตสูง จากฐานข้อมูลคัดกรองของประชาชนจากสำนักงานหลักประกันสุขภาพแห่งชาติ จำนวน 869,340 ราย รวมมีข้อมูล 3,335,431 ตัวอย่าง รวบรวมตั้งต่ปี 2017 ถึง 2022 ทำการสุ่มข้อมูลเพื่อให้ได้ชุดข้อมูลที่เล็กลงซึ่งเป็นการทดลองที่ช่วยให้ประหยัดทรัพยากรมากขึ้น สำหรับการทดลองในชุดข้อมูลจริงผู้วิจัยมีขั้นตอนการทดลองดังนี้

1) แบ่งการทดลองเป็นสองกรณีคือ กรณีใช้ข้อมูลจริงที่มีส่วนประกอบโครงสร้างเสมือนข้อมูลจำลอง และกรณีข้อมูลจริงที่ใช้จำนวนตัวแปรอิสระครบทุกตัวแปร

2) เฉพาะกรณีที่มีส่วนประกอบโครงสร้างเสมือนข้อมูลจำลอง ต้องใช้ตัวแปรอิสระ 5 ตัวแปร และลักษณะข้อมูลต้องเป็นข้อมูลแบบต่อเนื่องเท่านั้น ทำการคัดเลือกตัวแปรอิสระที่มีลักษณะเป็นข้อมูลแบบต่อเนื่อง (Continuous Data) จำนวน

5 ตัวแปร โดยคัดเลือกจากตัวแปรอิสระชนิดต่อเนื่องทุกตัวในชุดข้อมูลจริง มาทำการหาความสำคัญของตัวแปร โดยใช้ตัวแบบการวิเคราะห์ข้อมูลการถดถอยโลจิสติกแบบทวิภาคหา 5 ลำดับที่มีค่าสัมประสิทธิ์การถดถอยที่มากที่สุด

3) สำหรับการทดลองทุกกรณี จะทำการสุ่มจำนวนข้อมูลที่ใช้สำหรับการทดลอง 1 ครั้ง ให้มีขนาด มากกว่าหรือเท่ากับ 20 เท่า ของจำนวนพารามิเตอร์ที่ได้จากตัวแบบโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าสองเลเบล และเนื่องด้วยข้อมูลมีลักษณะเป็น Person-peroid ผู้ป่วยหนึ่งหน่วยตัวอย่างจะมีข้อมูลได้หลายปี จึงทำการสุ่มตามเลขไอดีของผู้ป่วย เพื่อให้ข้อมูลของแต่ละคนมีข้อมูลทุกปีอยู่ในชุดข้อมูลเดียวกัน และผลจากการเก็บข้อมูลที่เก็บจนกว่าจะแสดงว่าเป็นโรค ทำให้มีผลลัพธ์ที่แสดงว่าเป็นโรค มีอยู่ 1% จากข้อมูลที่มีทั้งหมด จึงทำการสุ่มข้อมูลที่ไอดีของผู้ป่วยมีการแสดงว่าเป็นโรคเบาหวานในปีสุดท้ายที่ทำการสำรวจ ให้มีจำนวนมากกว่า ไอดีของผู้ป่วยที่แสดงว่าไม่เป็นโรคเบาหวานในปีสุดท้ายอยู่ 5 เท่า เพราะต้องการให้มีค่าที่แสดงว่าเป็นโรคมีมากเพียงพอที่ไม่ทำให้ผลการทำนายไม่เที่ยงตรง

4) ทำการสุ่มข้อมูล 50 ชุด เพื่อใช้ในการทดลอง จำนวน 50 รอบ

เงื่อนไขที่ทำการศึกษา

เพื่อให้เปรียบเทียบผลที่เกิดขึ้นในแต่ละการทดลองโดยไม่เกิดความลำเอียง จึงกำหนดค่าพารามิเตอร์ที่สำคัญให้เป็นค่าเดียวกัน ดังนี้ อัตราการเรียนรู้ (learning rate) คือ 0.001 ,วิธีการหาค่าที่ดีที่สุด (optimizer) ใช้ Adam optimizer กำหนดให้ จำนวนรอบในการทำ backpropagation ต่อ 1 ตัวแบบ (epoch) คือ 100, ฟังก์ชันกระตุ้นที่ใช้ในตัวแบบ คือ Sigmoid function และ ฟังก์ชันการสูญเสียที่ใช้ในกระบวนการ backpropagation คือ Binary-crossentropy

แนวคิดในการออกแบบตัวแบบ

สำหรับทุกกรณีในงานวิจัย การออกแบบโครงสร้างของตัวแบบจะถูกกำหนด ดังนี้ กำหนดให้ตัวแบบมีชั้นที่ซ่อน (hidden layer) 2 ชั้น โดย input layer มีโหนดเท่ากับจำนวนตัวแปรอิสระ, hidden layer แรก มีจำนวนโหนดเป็น 2 เท่าของจำนวนตัวแปรอิสระ, hidden layer ชั้นที่ 2 มีจำนวนโหนดเท่ากับจำนวนตัวแปรอิสระ และ output layer จะมีโหนดเป็นจำนวนเท่ากับตัวแปรตาม

ขั้นตอนดำเนินงานวิจัย

การวิเคราะห์ข้อมูลในข้อมูลขั้นตอนการดำเนินงานวิจัยดังนี้

1) เตรียมการจำลองข้อมูลดังที่กำหนดไว้ โดยแต่ละชุดข้อมูลแบ่งข้อมูลออกเป็นสองส่วน คือ ชุดข้อมูลฝึก 50% และชุดข้อมูลทดสอบ 50% ออกแบบโครงสร้างของตัวแบบโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าตามแนวคิดการออกแบบตัวแบบและกำหนดเงื่อนไขให้ใกล้เคียงกันตามเงื่อนไขที่ทำการศึกษา

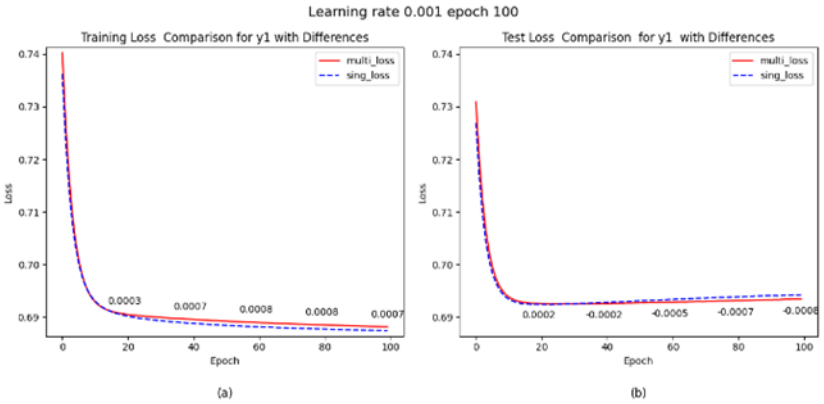
2) เมื่อทำการฝึกสอนและทำการฝึกในชุดทดสอบแล้ว นำค่าฟังก์ชันการสูญเสียแต่ละ epoch หาค่าเฉลี่ยและพล็อตกราฟเพื่อเปรียบเทียบประสิทธิภาพในการลดค่าฟังก์ชันการสูญเสียของตัวแบบและหาค่าทางสถิติเพื่อทดสอบความมีนัยสำคัญทางสถิติ โดยใช้ Paired sample t-test และหาความแตกต่างระหว่างสองตัวแบบในแต่ละ epoch ด้วยการใช้ค่าเฉลี่ยแต่ละ epoch ของ multi-label ลบกับ ค่าเฉลี่ยแต่ละ epoch ของ single-label

3) ประเมินประสิทธิภาพของตัวแบบแต่ละตัวด้วย Area Under ROC Curve เปรียบเทียบผลการทำนายกับค่าจริง

ผลการวิจัย

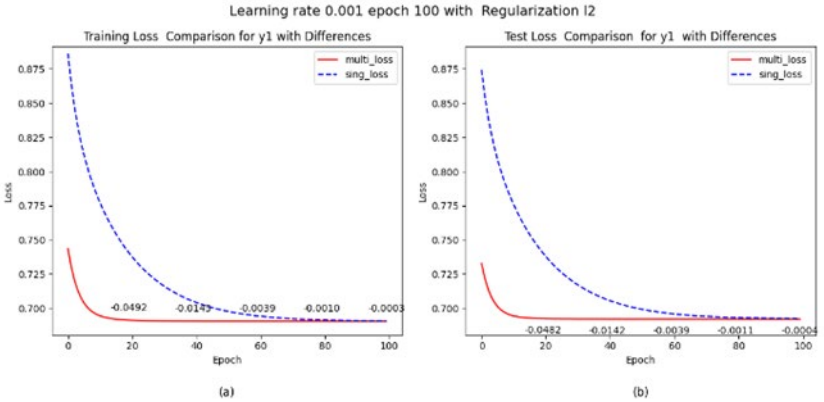
ผลการทดลองกรณีข้อมูลจำลอง

จากการทดลองจำนวน 50 รอบ แต่ละรอบสร้างข้อมูล 2000 ตัวอย่าง ค่าเฉลี่ยของค่าฟังก์ชันการสูญเสียในแต่ละ epoch ที่ได้จากข้อมูลชุดฝึกและข้อมูลชุดทดสอบ แสดงดังภาพที่ 1



ภาพที่ 1 กราฟแสดงการลดลงของค่าเฉลี่ยฟังก์ชันการสูญเสียในแต่ละ epoch ของข้อมูลชุดฝึกและข้อมูลชุดทดสอบในข้อมูลจำลอง

ผลการทดลองพบว่าตัวแบบเกิด overfit ขึ้นเพราะกราฟมีลักษณะเชิดขึ้นในทั้งสองตัวแบบ แต่ multi-label มีความรุนแรงของ overfit ที่น้อยกว่า ผู้วิจัยจึงทำการใช้วิธี dropout และ regularizers L2 สำหรับแก้ไขปัญหา overfit พบว่า การใช้ dropout ไม่ช่วยลดปัญหาได้ จึงรายงานผลการทดลองเฉพาะวิธีการ regularizers L2 เท่านั้น



ภาพที่ 2 กราฟแสดงการลดลงของค่าเฉลี่ยฟังก์ชันการสูญเสียในแต่ละ epoch ของข้อมูลชุดฝึกและข้อมูลชุดทดสอบในข้อมูลจำลองเมื่อทำการแก้ไขปัญหา overfit

จากภาพที่ 2 กราฟของ multi-label ในช่วงเริ่มต้นมีการลดลงของฟังก์ชันการสูญเสียที่ไวกว่า และเมื่อทำการเรียนรู้ไประยะหนึ่งค่าฟังก์ชันการสูญเสียทั้งสองตัวแบบมีลักษณะที่เริ่มลู่เข้าหากัน ถึงแม้การลดลงของค่าฟังก์ชันการสูญเสียใน multi-label จะลดลงได้ดีกว่าแต่ผลการทำนาย ยังคงให้ผลลัพธ์ของค่าเฉลี่ย AUC และ ค่าเฉลี่ย accuracy ที่ไม่มีความแตกต่างจาก single-label อย่างมีนัยสำคัญทางสถิติที่ 0.05

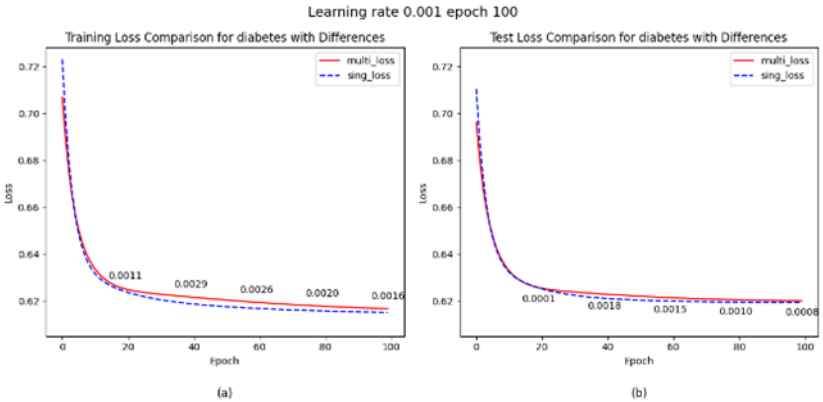
ตารางที่ 1 แสดงจำนวนพารามิเตอร์และประสิทธิภาพในข้อมูลจำลอง

ตัวแบบ	Multi-label	Single-label	t	Sig.
จำนวนพารามิเตอร์	127	121		
เวลาที่ใช้ในการเรียนรู้ตัวแบบเฉลี่ย (วินาที)	11.1597	8.2570		
ค่าเฉลี่ย AUC	0.4977	0.4990	-0.8324	.4092
ค่า accuracy	0.5188	0.5189	-0.1546	.8777

* $p < .05$

ผลการทดลองในข้อมูลจริงโรคเบาหวานเมื่อฝึกร่วมกับโรคความดันโลหิตสูงที่มีโครงสร้างการทดลองเสมือนข้อมูลจำลอง

การทดลองส่วนนี้สร้างขึ้นเพื่อใช้เปรียบเทียบกับกรณีข้อมูลจำลอง ต้องการข้อมูลชนิดต่อเนื่องจำนวน 5 ตัวแปร ผลการคัดเลือกความสำคัญของตัวแปรอิสระด้วยสัมประสิทธิ์การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาค ในข้อมูลโรคเบาหวาน ซึ่งผู้วิจัยเลือกตัวแปร BMI ,t.bslevel ,age ,SBP และ DBP มาทำการทดลอง และสุ่มข้อมูล 50 ชุดจากชุดข้อมูลใหญ่ แต่เนื่องจากข้อมูลเป็นลักษณะ Person-period จึงต้องทำการสุ่มให้ทุกแถวของผู้ป่วยหนึ่งไอดีอยู่ในชุดข้อมูลเดียวกัน แต่ละชุดเลือกจำนวนผู้ป่วยที่ปีสุดท้ายที่มีข้อมูลเป็นโรคเบาหวาน 825 ราย และผู้ป่วยที่ปีสุดท้ายที่มีข้อมูลไม่เป็นโรคเบาหวาน 165 ราย รวมสุ่มมา 990 ราย ทำให้มีข้อมูลเฉลี่ยรวม 2590 ตัวอย่าง ต่อหนึ่งชุดข้อมูล



ภาพที่ 3 กราฟแสดงการลดลงของค่าเฉลี่ยฟังก์ชันการสูญเสียในแต่ละ epoch ของข้อมูลชุดฝึกและข้อมูลชุดทดสอบในข้อมูลจริงที่มีโครงสร้างเสมือนข้อมูลจำลอง

จากภาพที่ 3 การลดลงของค่าฟังก์ชันการสูญเสียในทั้งสองตัวแบบมีการลดลงที่ไวใกล้เคียงกัน ในช่วง epochที่ 20 ในช่วงต่อมากราฟของ single-label ลดลงไวกว่าเล็กน้อยและเมื่อเรียนรู้ไปจนสิ้นสุด ทั้ง multi-label และ single-label มีค่าฟังก์ชันการสูญเสียที่เข้าใกล้กันมากขึ้น โดยเฉพาะในชุดข้อมูลทดสอบ และจากตารางที่ 2 ค่าเฉลี่ยของ AUC ของตัวแบบใน single-label แตกต่างจาก multi-label อย่างมีนัยสำคัญทางสถิติ โดย single-label ให้ค่า AUC ที่ดีกว่าแต่มีความแตกต่างเพียงร้อยละ 0.6 ของ AUC ของ multi-label แต่ค่าเฉลี่ย accuracy ของทั้งสองตัวแบบมีค่าไม่ต่างกัน

ตารางที่ 2 แสดงจำนวนพารามิเตอร์และประสิทธิภาพของตัวแบบในข้อมูลจริงโรคเบาหวานเมื่อฝึกร่วมกับโรคความดันโลหิตสูงที่มีโครงสร้างการทดลองเสมือนข้อมูลจำลอง

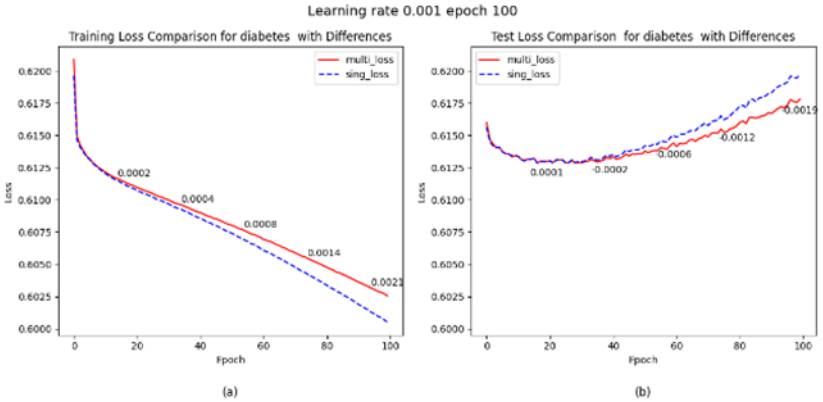
ตัวแบบ	Multi-label	Single-label	t	Sig.
จำนวนพารามิเตอร์	127	121		
เวลาที่ใช้ในการเรียนรู้ตัวแบบเฉลี่ย (วินาที)	11.8440	9.2841		
ค่าเฉลี่ย AUC	0.5730	0.5764	-2.4553	.0177
ค่า accuracy	0.6804	0.6804	0.2858	.7763

* $p < .05$

ผลการทดลองในข้อมูลจริงโรคเบาหวานเมื่อฝึกร่วมกับโรคความดันโลหิตสูง

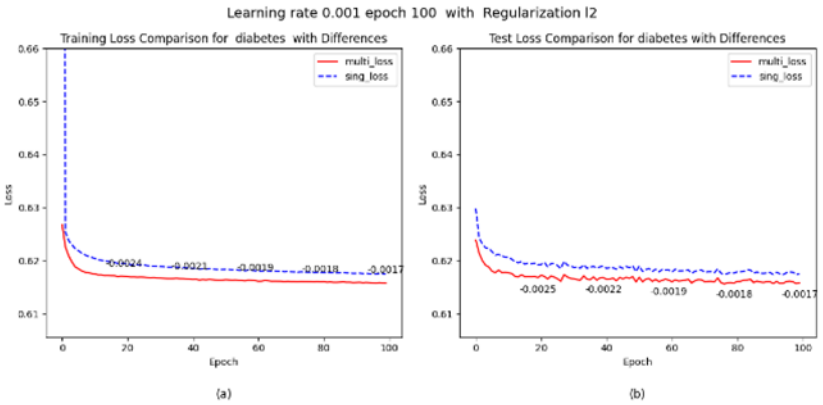
ทดลองในข้อมูลเชิงปริมาณและเชิงคุณภาพ จำนวนตัวแปรอิสระมี 19 ตัวแปร เมื่อทำการแปลงเพื่อใช้สำหรับฝึกฝนตัวแบบทำให้มีจำนวนพารามิเตอร์ในตัวแบบ 3752 ตัว ดังนั้นต้องการจำนวนข้อมูลอย่างน้อย 75040 หน่วยตัวอย่าง สุ่ม

ให้ข้อมูลทุกแถวของผู้ป่วยอยู่ในชุดข้อมูลเดียวกัน สุ่มผู้ป่วยที่ปีสุดท้ายแสดงว่าเป็นโรคเบาหวาน 25000 ราย และผู้ป่วยที่ปีสุดท้ายแสดงว่าไม่เป็นโรคเบาหวาน 5000 ราย ทำให้มีข้อมูลเฉลี่ยรวม 77975 ตัวอย่าง ต่อหนึ่งชุดข้อมูล



ภาพที่ 4 กราฟแสดงการลดลงของค่าเฉลี่ยฟังก์ชันการสูญเสียในแต่ละ epoch ของข้อมูลชุดฝึกและข้อมูลชุดทดสอบในข้อมูลจริงโรคเบาหวานและโรคความดันโลหิตสูง

จากภาพที่ 4 ชุดข้อมูลจริงโรคเบาหวานและโรคความดันโลหิตสูงมีลักษณะที่เกิด overfit มาก ตัวแบบเรียนรู้ได้ดีจนถึงช่วง 20 ถึง 30 epoch เท่านั้น หลังจากนั้นค่าฟังก์ชันการสูญเสียจะเริ่มเพิ่มขึ้นตามภาพที่ 4 (b) สังเกตว่าหลังจากตัวแบบเริ่มเกิด overfit ตัวแบบ multi-label จะช่วยลดความรุนแรงของการเกิด overfit ได้ แต่ยังไม่สามารถแก้ปัญหาได้ ทำการปรับปรุงตัวแบบให้ไม่เกิด overfit ด้วยการใช้ regularizers l2 ผลการทดลองเป็นดังภาพที่ 5



ภาพที่ 5 กราฟแสดงการลดลงของค่าเฉลี่ยฟังก์ชันการสูญเสียในแต่ละ epoch ของข้อมูลชุดฝึกและข้อมูลชุดทดสอบในข้อมูลจริงโรคเบาหวานและโรคความดันโลหิตสูงหลังจากแก้ปัญหา overfit

จากภาพที่ 5 แสดงให้เห็นว่าค่าฟังก์ชันการสูญเสียของ multi-label มีการลดลงที่ไวกว่าทั้งในชุดฝึกและชุดทดสอบ multi-label และค่อยๆลดลงจนมีลักษณะลู่เข้าหากัน แต่ถึงแม้ค่าฟังก์ชันการสูญเสียของตัวแบบ multi-label จะลดลงไวกว่าและมีค่าน้อยกว่าแต่ค่าเฉลี่ยของ AUC และ accuracy ทั้งสองตัวแบบไม่มีความแตกต่างกันที่ระดับนัยสำคัญ 0.05

ตารางที่ 3 แสดงจำนวนพารามิเตอร์และประสิทธิภาพของตัวแบบในข้อมูลจริงโรคเบาหวานและโรคความดันโลหิตสูง

ตัวแบบ	Multi-label	Single-label	t	Sig.
จำนวนพารามิเตอร์	3752	3721		
เวลาที่ใช้ในการเรียนรู้ตัวแบบเฉลี่ย(วินาที)	176.1968	155.4337		
ค่าเฉลี่ย AUC	0.5943	0.5943	-0.0566	.9551
ค่าเฉลี่ย accuracy	0.6813	0.6814	-0.0524	.9584

* $p < .05$

สรุปและอภิปรายผลการวิจัย

การวิจัยนี้มีจุดประสงค์เพื่อศึกษาศักยภาพในการใช้เลเบลที่มีความเกี่ยวข้องกันอย่างอย่างโรคเบาหวานและโรคความดันโลหิตสูงมาใช้ประโยชน์สำหรับการเพิ่มประสิทธิภาพข้อมูลเพื่อช่วยลดค่าฟังก์ชันการสูญเสียและเพิ่มประสิทธิภาพการทำนายด้วยตัวแบบโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าสองเลเบลเปรียบเทียบกับตัวแบบโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าหนึ่งเลเบล จากผลการวิจัย สามารถสรุปผลลัพธ์ได้ดังนี้

เมื่อการเรียนรู้เกิด overfit ตัวแบบ multi-label มีค่าฟังก์ชันการสูญเสียเพิ่มขึ้นช้ากว่าตัวแบบ single-label เสมอ และจะมีความแตกต่างของต่างของฟังก์ชันการสูญเสียมากขึ้นเมื่อการฝึกฝนของตัวแบบถูกดำเนินไป เพราะ multi-label มีข้อมูลของอีกเลเบลหนึ่งที่ช่วยไม่ให้การประมาณค่าพารามิเตอร์มีช่วงที่แคบเกินไป จึงทำให้เมื่อเกิด overfit ขึ้นจึงสามารถควบคุมความรุนแรงได้ และ ผลการทดลองในทุกกรณี เมื่อข้อมูลถูกเรียนรู้ไปอย่างต่อเนื่อง ค่าฟังก์ชันการสูญเสียจะลดลงเข้าหากันจนใกล้เคียงกันในที่สุด ซึ่งการลดของฟังก์ชันการสูญเสียในการทดลองนี้ไม่มีผลต่อประสิทธิภาพการทำนายของตัวแบบ ผลลัพธ์ที่ได้ทั้งสองตัวแบบไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ สามารถใช้ตัวแบบ multi-label ในการใช้ประโยชน์สำหรับการทำนายได้ไม่ต่างจาก single-label ซึ่งมีข้อดีคือใช้เวลาน้อยกว่าการสร้างการทำนายจากสองตัวแบบ

โดยสรุป ความสามารถในการลดค่าฟังก์ชันการสูญเสียและเพิ่มประสิทธิภาพการทำนายด้วยการใช้ความเสี่ยงข้อมูลโรคเบาหวานฝึกร่วมกับความดันโลหิตสูงในชุดข้อมูลโรคเบาหวานและโรคความดันโลหิตสูง ในทางทฤษฎีสามารถใช้ข้อมูลลดค่าฟังก์ชันการสูญเสียได้เร็วกว่าแต่ในแง่ประสิทธิภาพการทำนายไม่ได้แสดงผลที่เป็นที่ประจักษ์ และในทางปฏิบัติกับข้อมูลจริงโรคเบาหวานและโรคความดันโลหิตสูงยังไม่มีความแตกต่างที่ความจัดเจนมากนักในแง่ของการใช้ข้อมูลในการเพิ่มประสิทธิภาพการใช้ข้อมูลและเพิ่มประสิทธิภาพการทำนาย

ข้อเสนอแนะในการวิจัยครั้งต่อไป

เนื่องจากในการทดลองครั้งนี้ผลลัพธ์ในการทดลองกับข้อมูลจริงยังไม่เป็นที่ประจักษ์มากนัก ซึ่งอาจเป็นผลจากจำนวนของเลเบลที่มีน้อยอาจไม่เพียงพอที่สำหรับใช้ประโยชน์ได้ในการทดลองครั้งต่อไปอาจเพิ่มจำนวนเลเบลผลลัพธ์ที่มีความเกี่ยวข้องกันจากการมีตัวแปรอิสระร่วมกันเข้ามาและผู้วิจัยมีข้อเสนอแนะสำหรับผู้อ่านที่มีความสนใจงานวิจัยเพิ่มเติม ดังนี้ งานวิจัยนี้ มีการกำหนดเงื่อนไขโดยไม่ได้มีการเปลี่ยนค่าที่หลากหลาย ซึ่งในแต่ละตัวแบบ และการจำลองข้อมูล อาจสามารถปรับให้แตกต่างไปจากที่ใช้ในงานวิจัยฉบับนี้ได้ และงานวิจัยนี้ได้ทำการศึกษาในข้อมูลที่เป็นลักษณะผลลัพธ์เป็นทวิภาค อาจจะมีการนำไปใช้กับข้อมูลที่ตัวแปรตามมีลักษณะต่อเนื่องหรือทำการทดลองในชุดข้อมูลจริงโรคเบาหวานและความดันโลหิตสูงชุดอื่นๆ หรือนำไปประยุกต์ใช้กับข้อมูลโรคอื่นๆที่เกี่ยวข้องกัน

เอกสารอ้างอิง

Elujide, I., Fashoto, S. G., Fashoto, B., Mbunge, E., Folorunso, S. O., & Olamijuwon, J. O. (2021). Application of deep and machine learning techniques for multi-label classification performance on psychotic disorder diseases. *Informatics in Medicine Unlocked*, *23*, 100545.

Kiatsupaibul, S., & Chantarasiripas, P. (2024). Data Efficiency in Multi-response Neural Networks. *Unpublished work*.

Klein, J. P., & Moeschberger, M. L. (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer New York.

Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis* (Vol. 476). John Wiley & Sons.

Marubini, E., & Valsecchi, M. G. (2004). *Analysing survival data from clinical trials and observational studies* (Vol. 15). John Wiley & Sons.

Maxwell, A., Li, R., Yang, B., Weng, H., Ou, A., Hong, H., Zhou, Z., Gong, P., & Zhang, C. (2017). Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC Bioinformatics*, *18*(14), 523.

Menard, S. (2002). *Applied logistic regression analysis*. Sage.

Suresh, K., Severn, C., & Ghosh, D. (2022). Survival prediction models: an introduction to discrete-time modeling. *BMC Medical Research Methodology*, *22*(1), 207.

Szandała, T. (2021). Review and comparison of commonly used activation functions for deep neural networks. *Bio-inspired neurocomputing*, 203-224.

Tsimihodimos, V., Gonzalez-Villalpando, C., Meigs, J. B., & Ferrannini, E. (2018). Hypertension and Diabetes Mellitus: Coprediction and Time Trajectories. *Hypertension*, *71*(3), 422-428.

Vieira, S., Pinaya, W. H., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neurosci Biobehav Rev*, *74*(Pt A), 58-75

Zhou, L., Zheng, X., Yang, D., Wang, Y., Bai, X., & Ye, X. (2021). Application of multi-label classification models for the diagnosis of diabetic complications. *BMC Medical Informatics and Decision Making*, *21*(1), 182.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.