A COMPARATIVE ANALYSIS OF DEEP LEARNING AND TRADITIONAL METHODS FOR IMPUTATION OF MISSING BODY WEIGHT DATA IN HOSPITAL RECORDS: SIMULATION IN IPD-ICU SETTING

Metas MUANGNAK[1*] and Vitara PUNGPAPONG[1]

1 Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Thailand; 6680309226@cbs.chula.ac.th (Corresponding Author)

ARTICLE HISTORY

Received: 7 April 2025 **Revised:** 21 April 2025 **Published:** 6 May 2025

ABSTRACT

Missing data in hospital records especially in Intensive Care Units (ICU) and Inpatient Departments (IPD) is a common problem that can affect patient care and research accuracy. This study compares ten imputation methods, including traditional methods (mean, median, k-NN, MICE, MissForest), a hybrid method (HyperImpute), and deep learning (DL) methods (MLPRegressor, AEImputer, MIWAE, GAIN), to evaluate their performance in handling missing body weight data. A simulated dataset with 63 features was created under controlled conditions, varying across three sample sizes (5,000, 25,000, 50,000), three missingness mechanisms (MCAR, MAR, MNAR), and six missingness rates (10% to 60%). Performance was assessed using RMSE, MAPE, runtime, and memory usage. The results show that simple methods like mean and median still perform well, offering solid baseline performance with minimal resource usage. MissForest and HyperImpute offer a good trade-off between accuracy and computational efficiency, making them suitable for moderate missingness scenarios. Although widely used, the MICE method showed limited adaptability to non-parametric or complex data structures, leading to suboptimal results in several conditions. In contrast, deep learning models gave mixed results. DL sometimes performed well but often required intensive hyperparameter tuning and used more runtime and memory usage. While AEs method showed stable performance, GAIN was sensitive to both missing data patterns and sample sizes, leading to inconsistent outcomes. Overall, while deep learning has potential, it comes with challenges such as sensitivity to hyperparameters and high computational demands. In many practical cases, traditional or hybrid methods may be more effective and easier to implement.

Keywords: Missing Value, Imputation, Artificial Neural Networks, Deep Learning, Healthcare

CITATION INFORMATION: Muangnak, M., & Pungpapong, V. (2025). A Comparative Analysis of Deep Learning and Traditional Methods for Imputation of Missing Body Weight Data in Hospital Records: Simulation in IPD-ICU Set. *Procedia of Multidisciplinary Research, 3*(5), 12

การศึกษาเปรียบเทียบการแทนค่าน้ำหนักสูญหายด้วยวิธีการเรียนรู้เชิงลึกกับวิธีดั้งเดิมด้วยวิธีการจำลองข้อมูลในบริบทของผู้ป่วยในและหอผู้ป่วยหนักภายในโรงพยาบาล

เมธัส ม่วงนาค[1*] และ วิฐรา พึ่งพาพงศ์[1]

1 ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย; 6680309226@cbs.chula.ac.th (ผู้ประพันธ์บรรณกิจ)

บทคัดย่อ

การสูญหายของข้อมูลในเวชระเบียนโรงพยาบาล โดยเฉพาะในหอผู้ป่วยหนัก (ICU) และหอผู้ป่วยใน (IPD) เป็นปัญหาที่พบบ่อยและส่งผลต่อการดูแลผู้ป่วยและความถูกต้องของงานวิจัย การศึกษานี้เปรียบเทียบวิธีการแทนค่าสูญหาย 10 วิธี ได้แก่ วิธีแบบดั้งเดิม (Mean, Median, k-NN, MICE, MissForest), วิธีแบบผสม (HyperImpute) และวิธีแบบการเรียนรู้เชิงลึกหรือ DL (MLPRegressor, AEImputer, MIWAE, GAIN) โดยใช้ข้อมูลจำลอง 63 ตัวแปร ภายใต้เงื่อนไขควบคุม ได้แก่ ขนาดตัวอย่าง 3 ระดับ (5,000, 25,000, 50,000), กลไกการสูญหาย 3 รูปแบบ (MCAR, MAR, MNAR) และอัตราการสูญหาย 6 ระดับ (10% ถึง 60%) โดยประเมินผลด้วย RMSE, MAPE, เวลาในการประมวลผล และการใช้หน่วยความจำ ผลการศึกษาพบว่า Mean และ Median ยังให้ผลลัพธ์ที่ดี พร้อมความเร็วและใช้ทรัพยากรต่ำ MissForest และ HyperImpute ให้ความแม่นยำที่สมดุลกับประสิทธิภาพ เหมาะกับกรณีที่ข้อมูลสูญหายในระดับปานกลาง วิธีที่นิยมอย่าง MICE กลับมีข้อจำกัดกับชุดข้อมูลที่เป็น Non-Parametric ทำให้ผลลัพธ์ด้อยกว่าในหลายเงื่อนไข ด้าน DL แม้บางวิธีให้ผลลัพธ์ดี แต่ต้องอาศัยการปรับแต่งพารามิเตอร์อย่างละเอียด และใช้ทรัพยากรมาก โดยวิธีกลุ่ม AEs มีความเสถียรที่สุด ส่วน GAIN มีความไวต่อรูปแบบข้อมูลและขนาดตัวอย่าง ให้ผลลัพธ์ไม่สม่ำเสมอ โดยสรุป แม้ DL จะมีศักยภาพ แต่ในหลายกรณี วิธีดั้งเดิมหรือแบบผสมยังคงเป็นทางเลือกที่ใช้งานได้จริงและคุ้มค่า

คำสำคัญ: ค่าสูญหาย, การแทนค่าสูญหาย, โครงข่ายประสาทเทียม, การเรียนรู้เชิงลึก, การแพทย์

ข้อมูลการอ้างอิง: เมธัส ม่วงนาค และ วิฐรา พึ่งพาพงศ์. (2568). การศึกษาเปรียบเทียบการแทนค่าน้ำหนักสูญหายด้วยวิธีการเรียนรู้เชิงลึกกับวิธีดั้งเดิมด้วยวิธีการจำลองข้อมูลในบริบทของผู้ป่วยในและหอผู้ป่วยหนักภายในโรงพยาบาล. *Procedia of Multidisciplinary Research, 3*(5), 12

บทนำ

ปัญหาข้อมูลสูญหาย (Missing Data) ในการวิเคราะห์ข้อมูล เป็นปัญหาที่พบได้ทั่วไปในหลายบริบท ซึ่งปัญหานี้อาจนำไปสู่การประมาณค่าที่มีความผิดพลาด ลดความสามารถในการคำนวณทางสถิติ เพื่อลดอคติในการวิเคราะห์และการอนุมาน จึงควรเลือกใช้วิธีการแทนค่าสูญหายอย่างเหมาะสม ในทางการแพทย์ ปัญหานี้อาจส่งผลต่อคุณภาพการดูแลผู้ป่วยและการรักษาเฝ้าติดตามที่แม่นยำ ข้อมูลอย่างน้ำหนักตัว (Body Weight) เป็นข้อมูลพื้นฐานที่มักจะสูญหายจากการเก็บข้อมูลภายในโรงพยาบาล โดยเฉพาะในบริบทของหอผู้ป่วยใน (In-Patient Department; IPD) รวมไปถึงหน่วยที่ต้องใช้ความรวดเร็ว เร่งด่วนในการช่วยเหลือดูแลผู้ป่วยอย่างหน่วยดูแลผู้ป่วยที่อยู่ในภาวะวิกฤต (Intensive Care Unit; ICU) โดยข้อมูลน้ำหนักนั้นสูญหายได้จากหลากหลายสาเหตุ เช่น ผู้ป่วยไม่สามารถชั่งได้ ไม่ได้ทำการชั่งเนื่องจากเป็นกรณีเร่งด่วน มีจำนวนปริมาณผู้ป่วยในหน่วยปริมาณมาก ทำให้เกิดการตกหล่น ทั้งนี้ข้อมูลน้ำหนักมีความจำเป็นที่จะต้องนำมาใช้งานต่อในแง่ของการคำนวณดัชนีมวลกาย การคำนวณการให้สารน้ำ ปริมาณยา ความต้องการพลังงานและโปรตีน รวมถึงการนำปัจจัยนี้ไปใช้ในการพัฒนาแบบทำนายพยากรณ์ความเสี่ยงอื่นๆ

วิธีการแทนค่าสูญหาย (Imputation Methods) ถูกพัฒนาและใช้อย่างแพร่หลายในบริบทหลากหลาย ตั้งแต่วิธีการแบบง่าย เช่น การแทนค่าด้วยค่าเฉลี่ย (Mean) มัธยฐาน (Median) หรือฐานนิยม (Mode) ไปจนถึงวิธีการที่ซับซ้อนมากขึ้น เช่น k-Nearest Neighbors (k-NN), การถดถอย (Regression), Multiple Imputation by Chained Equations (MICE) แต่ละวิธีมีจุดแข็งและข้อจำกัดแตกต่างกัน เช่น ความง่ายในการใช้งาน ความสามารถในการรักษาโครงสร้างของข้อมูล ความยืดหยุ่นในการจัดการข้อมูลประเภทต่างๆ โดยในปัจจุบัน การเรียนรู้เชิงลึก (Deep Learning; DL) ได้รับความสนใจเพิ่มขึ้นในบริบทของการแทนค่าสูญหาย เนื่องจากสามารถเรียนรู้ความสัมพันธ์ที่ซับซ้อนและไม่เป็นเชิงเส้นในข้อมูลได้อย่างมีประสิทธิภาพ ตัวอย่างเช่น Multilayer Perceptron (MLP) ใช้โครงข่ายประสาทเทียมแบบป้อนหน้าเพื่อเรียนรู้การประมาณค่าจากข้อมูลที่สังเกตได้ และวิธีในรูปแบบที่มีโครงสร้างซับซ้อนเพิ่มมากขึ้น อย่าง Autoencoders (AEs) และ Generative Adversarial Networks (GANs)

ในภาพรวม การแทนค่าสูญหายยังคงเป็นหัวข้อที่ได้รับความสนใจอย่างต่อเนื่อง การศึกษานี้มุ่งเน้นการเปรียบเทียบระหว่างวิธีการแบบดั้งเดิมและวิธีการที่ใช้การเรียนรู้เชิงลึก รวมไปถึงการผสมผสานระหว่างวิธีต่างๆ เพื่อเสนอแนวทางที่เหมาะสมที่สุดต่อการใช้งานในบริบทจริงของโรงพยาบาล ซึ่งปัญหาข้อมูลสูญหาย เช่น น้ำหนักตัว เกิดขึ้นบ่อยและมีผลกระทบต่อการติดตามการรักษาและการพยากรณ์ทางคลินิก การศึกษานี้มีมีวัตถุประสงค์เพื่อ 1) ศึกษาเปรียบเทียบประสิทธิภาพของวิธีการแทนค่าสูญหายแบบดั้งเดิม (Mean, Median, k-NN, MICE, MissForest) และวิธีแบบผสมผสาน (HyperImpute) กับวิธีการแทนค่าสูญหายที่ใช้การเรียนรู้เชิงลึก (MLP, AEs, GAIN) 2) เพื่อให้คำแนะนำเกี่ยวกับกลยุทธ์การแทนค่าสูญหายที่มีประสิทธิภาพมากที่สุดสำหรับข้อมูลที่สูญหายในบริบทของโรงพยาบาล

การทบทวนวรรณกรรม

ข้อมูลที่สูญหาย (Missing Data) ข้อมูลที่สูญหายเป็นปัญหาทั่วไปในงานวิจัยด้านวิทยาศาสตร์สุขภาพ การแพทย์ และการวิเคราะห์ข้อมูลทางคลินิก ซึ่งเกิดขึ้นได้จากหลากหลายสาเหตุ เช่น ความผิดพลาดในการบันทึก การไม่ตอบแบบสอบถาม หรือการไม่สามารถตรวจวัดค่าได้ในบางกรณี การสูญหายของข้อมูลอาจมีผลต่อความน่าเชื่อถือและความถูกต้องของการวิเคราะห์เชิงสถิติและการตัดสินใจทางคลินิก กลไกของข้อมูลที่สูญหายสามารถจำแนกได้เป็น 3 ประเภทหลัก ได้แก่ 1) Missing Completely at Random (MCAR) คือ การที่ข้อมูลสูญหายโดยไม่ขึ้นกับข้อมูลใดๆ ทั้งที่สังเกตได้และไม่ได้สังเกต 2) Missing at Random (MAR) ข้อมูลสูญหายขึ้นอยู่กับข้อมูลที่สังเกตได้เท่านั้น และ 3) Missing Not at Random (MNAR) คือ การสูญหายที่ขึ้นอยู่กับค่าของตัวแปรที่หายไปเอง การทำความเข้าใจกลไกของการสูญหายเป็นสิ่งสำคัญในการเลือกวิธีการจัดการที่เหมาะสม (Little & Rubin, 2019)

แม้ว่าน้ำหนักตัวจะเป็นข้อมูลพื้นฐานที่มีความสำคัญอย่างยิ่งต่อการดูแลรักษาผู้ป่วย ทั้งในด้านการคำนวณขนาดยา การประเมินภาวะโภชนาการ และการวางแผนการรักษาโดยรวม แต่วรรณกรรมวิชาการในช่วง 10 ปีที่ผ่านมาจนถึง

ปัจจุบันชี้ให้เห็นอย่างชัดเจนว่า ข้อมูลน้ำหนักของผู้ป่วยมักสูญหายเป็นจำนวนมาก โดยเฉพาะในผู้ป่วยใน (IPD) และผู้ป่วยวิกฤต (ICU) หลักฐานจากงานวิจัยหลายฉบับ เช่น Charani et al. (2015), Willson et al. (2016), McFall et al. (2019) และการวิเคราะห์ฐานข้อมูลขนาดใหญ่ (เช่น MIMIC-III) ล้วนรายงานสัดส่วนของการสูญหายของข้อมูลน้ำหนักอยู่ในช่วงประมาณ 30-60% โดยเฉลี่ย และในบางบริบทที่มีข้อจำกัด เช่น ผู้ป่วยฉุกเฉินหรือการรับเข้าระบบแบบเร่งด่วน อัตราการสูญหายข้อมูลอาจสูงถึง 70-80% ได้ (McFall et al., 2019) ด้วยเหตุนี้ การแก้ปัญหาการสูญหายของข้อมูลน้ำหนักจึงเป็นประเด็นสำคัญที่ปรากฏชัดในวรรณกรรม ทั้งในแง่ของการพัฒนาแนวปฏิบัติการบันทึกที่ดีขึ้น และการประยุกต์ใช้เทคนิคการแทนค่าข้อมูลสูญหาย (Missing Data Imputation) เพื่อฟื้นคืนคุณภาพของชุดข้อมูลในระดับการวิจัยและการดูแลผู้ป่วยจริง วิธีการแทนค่าสูญหาย (Imputation Methods) การจัดการข้อมูลที่สูญหายสามารถทำได้หลายวิธี โดยจำแนกออกเป็นสองกลุ่มใหญ่ ได้แก่ การลบข้อมูล (Deletion Methods) และการแทนค่าสูญหาย (Imputation Methods) วิธีการลบข้อมูล เช่น Listwise Deletion และ Pairwise Deletion แม้จะมีความง่าย แต่ทำให้ข้อมูลสูญหายเพิ่มเติมและเกิดความเอนเอียงได้

การศึกษาที่เกี่ยวข้องกับการเปรียบเทียบระหว่าง DL และวิธีดั้งเดิม อย่างการศึกษาโดย Liu et al. (2023) ได้ทำการทบทวนอย่างเป็นระบบ (Systematic Review) เกี่ยวกับการแทนค่าสูญหายในข้อมูลสุขภาพ พบว่า DL มักมีประสิทธิภาพเหนือกว่าวิธีแบบดั้งเดิมในข้อมูลที่ซับซ้อน แต่มีข้อจำกัดด้านความซับซ้อนของโมเดลและการตีความ (Liu et al., 2023) และการศึกษาโดย Sun et al. (2023) เปรียบเทียบวิธี DL เช่น GAIN และ VAE กับวิธีดั้งเดิมอย่าง MICE และ MissForest ในชุดข้อมูลจำลองและชุดข้อมูลจริง พบว่าวิธีดั้งเดิมมักทำงานได้ดีกว่าในข้อมูลขนาดเล็กถึงปานกลาง (น้อยกว่า 30,000 รายการ) ส่วนของวิธีที่ใช้การเรียนรู้เชิงลึกอย่าง VAE และ GAIN นั้นต้องการการปรับแต่ง Hyperparameters อย่างระมัดระวังและมีความไวต่อประเภทของกลไกการสูญหายของข้อมูล นอกจากนี้ GAIN มักประสบปัญหา Mode Collapse โดยเฉพาะในสถานการณ์ที่ข้อมูลสูญหายมีความซับซ้อนมากขึ้น (Sun et al., 2023)

การศึกษาที่เกี่ยวข้องกับวิธีการแทนค่าสูญหายที่เป็นที่นิยมและวิธีการแทนค่าสูญหายที่น่าสนใจ

- MICE: ใช้โมเดลแบบมีเงื่อนไขสำหรับแต่ละตัวแปร ทำการแทนค่าสูญหายโดยการสร้างชุดข้อมูลที่สมบูรณ์หลายชุดขึ้นมาแล้วรวมผลลัพธ์ (Van Buuren & Groothuis-Oudshoorn, 2011)
- MissForest: ใช้อัลกอริทึม Random Forest แบบวนซ้ำ เหมาะสำหรับข้อมูลรูปแบบผสมและไม่จำเป็นต้องปรับพารามิเตอร์หรือใช้สมมติฐานเกี่ยวกับการแจกแจงข้อมูล (Stekhoven & Buhlmann, 2012)
- HyperImpute: รวมความยืดหยุ่นของวิธีการประมาณค่าแบบวนซ้ำ (Iterative Imputation) เข้ากับประสิทธิภาพของการเลือกแบบจำลองอัตโนมัติ (AutoML) เลือกโมเดลที่เหมาะสมสำหรับแต่ละตัวแปร และสามารถจัดการการแทนค่าสูญหายแบบปรับตัวได้ (Jarrett et al., 2022)
- MIWAE: โมเดลในกลุ่ม AEs ที่ประยุกต์ใช้ Deep Latent Variable Models และ Importance Sampling เพื่อเพิ่มความแม่นยำและสามารถประเมินความไม่แน่นอนได้ (Mattei & Frellsen, 2019)
- GAIN: โมเดลในกลุ่ม GANs ที่ประกอบด้วย Generator และ Discriminator ที่จะพยายามแยกแยะระหว่างค่าที่แท้จริง (สังเกตได้) และค่าที่ประมาณ โดย GAIN ทำการเพิ่ม Hint Mechanism ที่จะให้ข้อมูลบางส่วนกับ Discriminator ทำให้มั่นใจว่า Generator ได้รับข้อมูลย้อนกลับที่มีเป้าหมาย นำไปสู่การแทนค่าสูญหายที่มีประสิทธิภาพดีมากขึ้น (Yoon et al., 2018)

สมมติฐานการวิจัย

1) วิธีการที่ใช้การเรียนรู้เชิงลึกจะมีประสิทธิภาพดีกว่าวิธีการแบบดั้งเดิมและวิธีแบบผสมผสานในการแทนค่าน้ำหนักสูญหาย โดยเฉพาะในกรณีที่มีจำนวนข้อมูลผู้ป่วยปริมาณมาก

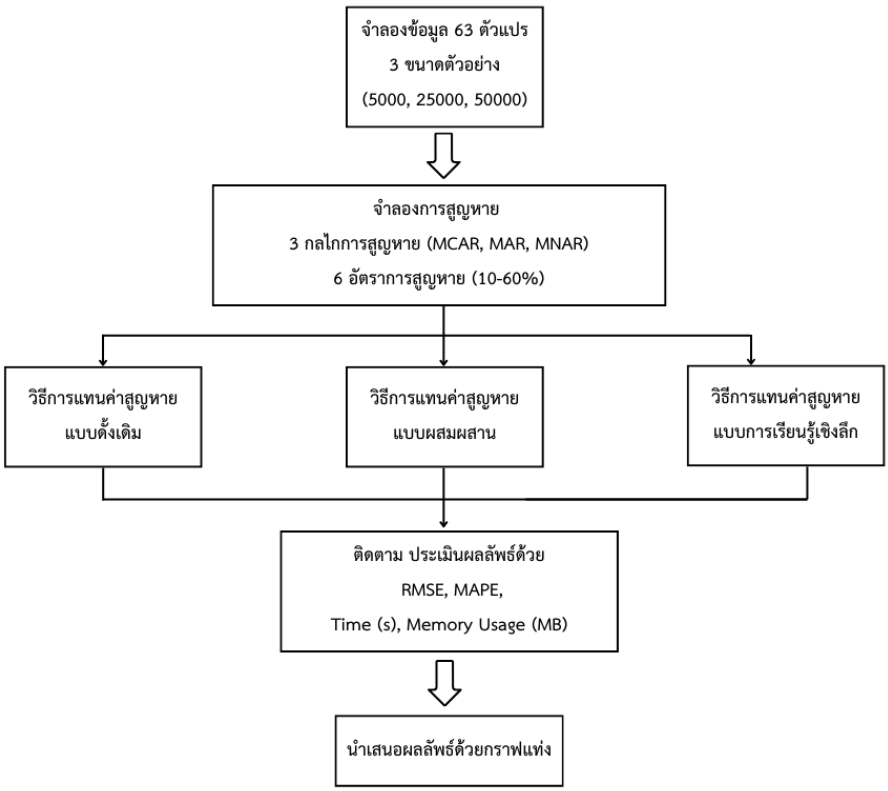
2) ประสิทธิภาพของวิธีการทั้งหมดจะลดลงเมื่อมีอัตราข้อมูลสูญหายที่เพิ่มขึ้น แต่วิธีการที่ใช้การเรียนรู้เชิงลึกจะมีประสิทธิภาพที่คงทนมากกว่าเมื่อมีอัตราข้อมูลสูญหายมาก

3) วิธีการแบบดั้งเดิมจะมีประสิทธิภาพในการประหยัดทรัพยากรการคำนวณมากกว่าวิธีแบบผสมผสานและวิธีการที่ใช้การเรียนรู้เชิงลึก เมื่อเปรียบเทียบที่ระดับจำนวนข้อมูลผู้ป่วย กลไกการสูญหายเดียวกันและอัตราข้อมูลสูญหายที่เท่ากัน

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงเปรียบเทียบ (Comparative Study) โดยใช้การจำลองข้อมูล (Simulation Data) ภายใต้บริบทของโรงพยาบาล (IPD-ICU) ทำการจำลองข้อมูลต่างๆ โดยอ้างอิงจากการศึกษาที่เกี่ยวข้องกับฐานข้อมูล MIMIC-III (Johnson et al., 2016) ซึ่งเป็นฐานข้อมูลสาธารณะที่ได้รับการอ้างอิงสูง ประกอบด้วยข้อมูลผู้ป่วย ICU ที่ครอบคลุมหลากหลายด้าน และใช้ในงานวิจัยหลากหลายสาขา เช่น การทำนายพยากรณ์โรค การวิเคราะห์คลินิก (Dai et al., 2020) และทำการอ้างอิงตัวแปรทางห้องปฏิบัติการเพิ่มเติมจากการศึกษาของ Sharafoddini et al. (2019) ซึ่งเป็นความเห็นจากจากแพทย์และผู้เชี่ยวชาญ และเสริมตัวแปรที่อาจมีอิทธิพลบางตัวตามความเหมาะสม (Sharafoddini et al., 2019) โดยปัจจัยทั้งหมด 63 ปัจจัย แบ่งเป็น 4 กลุ่มหลัก ได้แก่ ข้อมูลประชากรศาสตร์ (เพศ อายุ น้ำหนัก ส่วนสูง) ข้อมูลทางคลินิก (จำนวนวันใน ICU จำนวนวันใน IPD สถานะงดน้ำงดอาหาร สถานะผู้ป่วยติดเตียง กลุ่มโรค) สัญญาณชีพ (ค่าความดันโลหิต อุณหภูมิ ชีพจร การหายใจ) และข้อมูลทางห้องปฏิบัติการต่างๆ

แผนผังภาพรวมการทำงาน



ภาพที่ 1 แผนผังภาพรวมการทำงาน

ใช้การจำลองการสูญหายด้วยโปรแกรมภาษา R โดยใช้แพ็กเกจ “produce_NA” (Albu et al., 2024) ทำการจำลองการสูญหาย ในชุดข้อมูลตามกลไกการสูญหายของข้อมูล 3 รูปแบบที่แตกต่างกัน ได้แก่ MCAR, MAR และ MNAR โดยอัตราการสูญหายถูกปรับระหว่าง 10% ถึง 60% เพื่อจำลองสถานการณ์การสูญหายของข้อมูลในโลกความเป็นจริงที่มักพบในสภาพแวดล้อมทางคลินิก โดยมีรายละเอียดการใช้งานและการปรับแต่งดังแสดงในตารางที่ 1

ตารางที่ 1 การจำลองกลไกการสูญหายและอัตราการสูญหาย

กลไกการสูญหาย	การปรับแต่ง		
	ปรับกลไกการสูญหาย	ปรับอัตราการสูญหาย	เพิ่มเติม
MCAR	mechanism = "MCAR"	perc.missing = 0.1, 0.2,	ไม่มี
MAR	mechanism = "MAR"	0.3, 0.4, 0.5, 0.6	ใช้ idx.covariates
MNAR	mechanism = "MNAR"		ใช้ self.mask = "sym"

ในการศึกษาครั้งนี้ผู้ศึกษาวิจัยได้ใช้ทั้งวิธีการดั้งเดิม วิธีแบบผสมผสานและวิธีการที่ใช้การเรียนรู้เชิงลึกในการแทนค่าสูญหาย โดยเลือกวิธีเหล่านี้จากการทบทวนวรรณกรรม แต่ละวิธีมีความสามารถในการจัดการรูปแบบข้อมูลที่สูญหายไปที่แตกต่างกัน ดำเนินการแทนค่าสูญหายด้วยโปรแกรมภาษา Python โดยวิธีการแทนค่าสูญหายที่ใช้การศึกษาทั้งหมดดังแสดงในตารางที่ 2

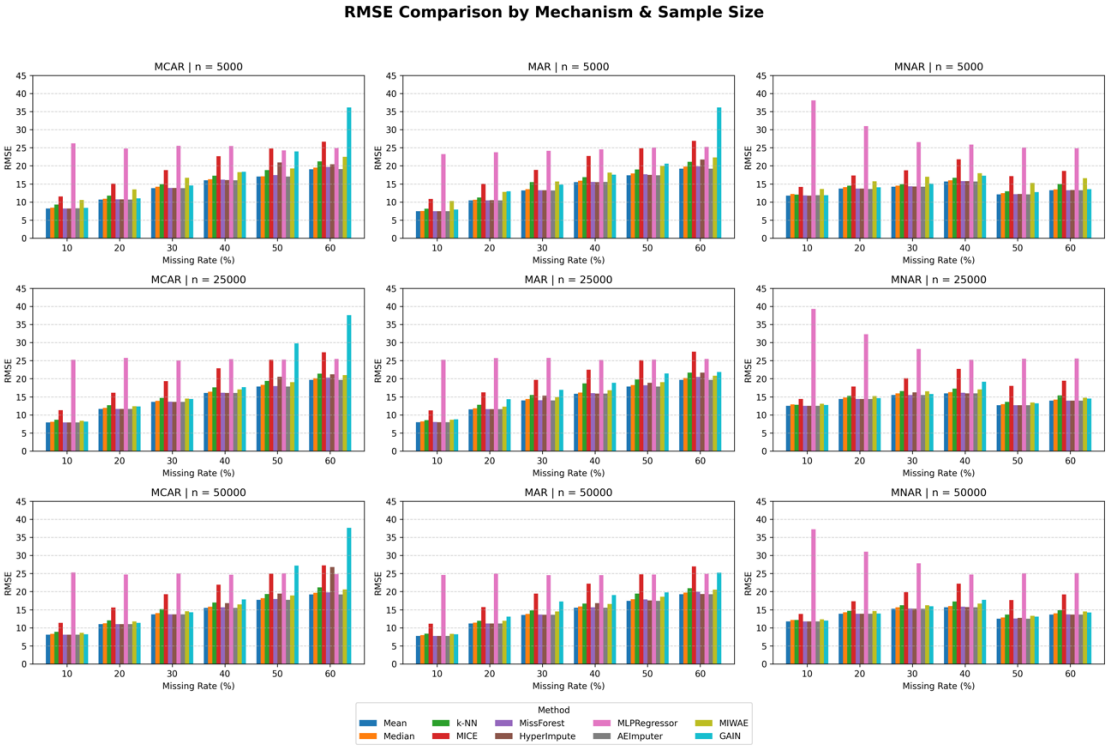
ตารางที่ 2 แสดงภาพรวมของวิธีการแทนค่าสูญหายที่ใช้ในการศึกษา

วิธีการแทนค่าสูญหาย	Python Library	การปรับแต่ง
Mean	Scikit-Learn	SimpleImputer: strategy='mean'
Median	Scikit-Learn	SimpleImputer: strategy='median'
k-NN	Scikit-Learn	KNNImputer: n_neighbors = 5
MICE	HyperImpute	Imputers().get(mice)
MissForest	HyperImpute	Imputers().get(missforest)
HyperImpute	HyperImpute	Imputers().get(hyperimpute)
MLPRegressor	Scikit-Learn	ใช้ MLPRegressor
AEImputer	AEImputer	ใช้ AEImputer
MIWAE	HyperImpute	Imputers().get(miwae)
GAIN	HyperImpute	Imputers().get(gain)

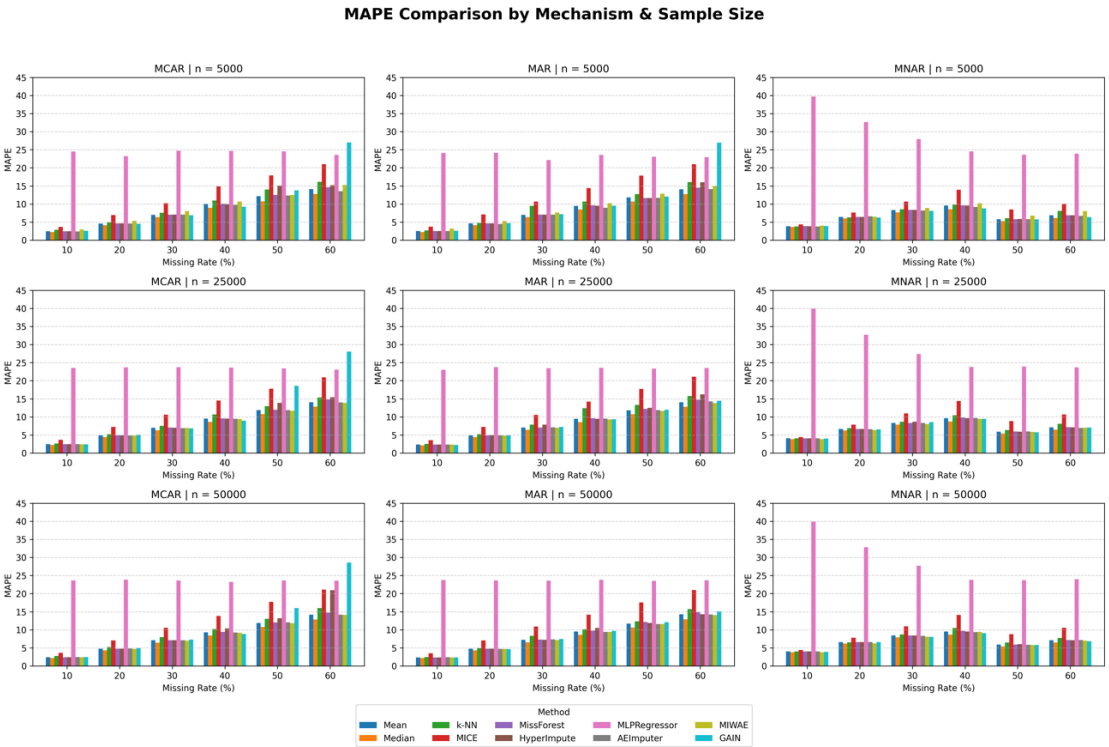
สำหรับตัวชี้วัดการประเมินผล ประสิทธิภาพของวิธีการแทนค่าสูญหายถูกประเมินโดยใช้ตัวชี้วัดต่อไปนี้ 1) ค่า Root Mean Squared Error (RMSE) วัดรากที่สองของค่าเฉลี่ยของความแตกต่างยกกำลังสองระหว่างค่าที่สังเกตได้และค่าที่คาดการณ์ RMSE ที่ต่ำกว่าหมายถึงประสิทธิภาพการแทนค่าสูญหายที่ดีกว่า เนื่องจากสะท้อนถึงความเบี่ยงเบนระหว่างค่าจริงและค่าที่คาดการณ์ที่น้อยลง 2) ค่า Mean Absolute Percentage Error (MAPE) แสดงถึงค่าเฉลี่ยของความแตกต่างสัมบูรณ์ในเชิงเปอร์เซ็นต์ระหว่างค่าที่สังเกตได้และค่าที่คาดการณ์ โดยเป็นค่าที่ปกติแล้วใช้ในการวัดข้อผิดพลาดของการแทนค่าสูญหาย ค่า MAPE ที่ต่ำกว่าหมายถึงประสิทธิภาพที่ดีกว่าในการรักษาสัดส่วนโดยรวมของข้อมูล ในแง่ของทรัพยากรการคำนวณ ทำการติดตาม 3) เวลาที่ใช้ในการฝึกฝนและแทนค่าสูญหายและ 4) หน่วยความจำที่ใช้ของวิธีการแทนค่าสูญหายทุกวิธี โดยใช้ไลบรารี 'time' ในการติดตามระยะเวลา และไลบรารี 'tracemalloc' ในการติดตามปริมาณหน่วยความจำที่ใช้งาน สำหรับการสรุปผลการวิเคราะห์ข้อมูล ใช้ไลบรารี 'Matplotlib' ในการสร้างภาพข้อมูล (Data Visualization) เพื่อเปรียบเทียบประสิทธิภาพของแต่ละวิธีการแทนค่าสูญหาย และจัดทำตารางเพื่อทำการสรุปและเปรียบเทียบข้อดีและข้อจำกัดของวิธีการแทนค่าสูญหายทั้งแบบดั้งเดิม แบบผสมผสาน และแบบวิธีการเรียนรู้เชิงลึก ทั้งนี้เพื่อเป็นข้อเสนอแนะและแนวทางสำหรับนักวิจัยและผู้ปฏิบัติงานทางคลินิกเกี่ยวกับการเลือกใช้วิธีการแทนค่าสูญหายในบริบทของโรงพยาบาล

ผลการวิจัย

การศึกษานี้ประเมินประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหาย โดยวัดจากค่าความแม่นยำของการประมาณค่า (RMSE และ MAPE) รวมทั้งเวลาและหน่วยความจำที่ใช้ในการประมวลผล ภายใต้สถานการณ์ที่แตกต่างกันทั้งในแง่ของกลไกข้อมูลสูญหาย (MCAR, MAR, MNAR) ขนาดตัวอย่าง (5,000; 25,000; 50,000) และระดับการสูญหายของข้อมูลที่แตกต่างกัน (ตั้งแต่ 10% ถึง 60%) โดยผลการวิเคราะห์แสดงดังภาพที่ 2, 3, 4, และ 5

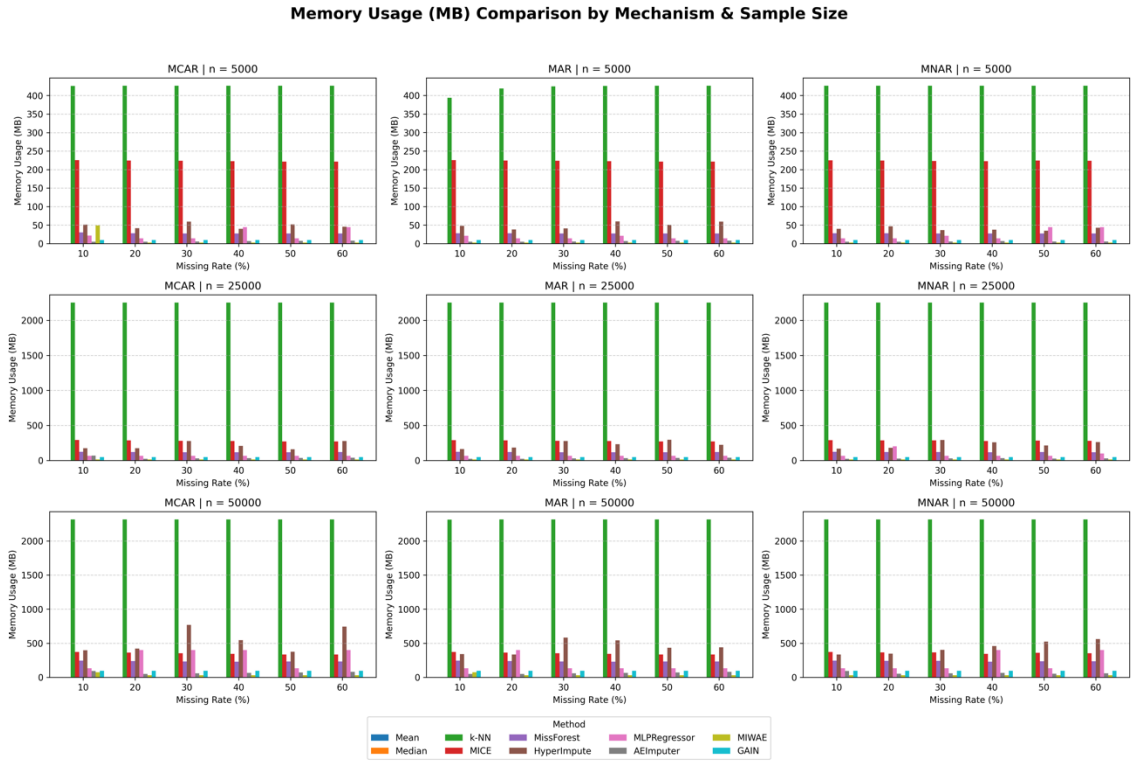


ภาพที่ 2 แสดงผลลัพธ์ RMSE ทั้ง 3 กลไกการสูญหายที่ชุดข้อมูล 5,000, 25,000 และ 50,000



ภาพที่ 3 แสดงผลลัพธ์ MAPE ทั้ง 3 กลไกการสูญหายที่ชุดข้อมูล 5,000, 25,000 และ 50,000

ภาพที่ 4 แสดงผลลัพธ์ Time (s) ทั้ง 3 กลไกการสูญหายที่ชุดข้อมูล 5,000, 25,000 และ 50,000



ภาพที่ 5 แสดงผลลัพธ์ Memory Usage (MB) ทั้ง 3 กลไกการสูญหายที่ชุดข้อมูล 5,000, 25,000 และ 50,000

ภาพรวมผลลัพธ์ด้านความแม่นยำ

จากภาพที่ 2 และ 3 พบว่า ค่าของ RMSE และ MAPE มีแนวโน้มเพิ่มสูงขึ้นอย่างต่อเนื่องเมื่อระดับการสูญหายของข้อมูลเพิ่มสูงขึ้น แสดงถึงความยากที่มากขึ้นในการประมาณค่าข้อมูลที่แม่นยำเมื่อข้อมูลสูญหายไปในสัดส่วนที่สูงขึ้น โดยวิธีการแทนค่าสูญหายแบบดั้งเดิม (Mean, Median) และวิธีที่ซับซ้อนมากขึ้นอย่าง MissForest และ HyperImpute รวมไปถึงวิธีการเรียนรู้เชิงลึกอย่าง AEImputer สามารถแสดงประสิทธิภาพที่ดีและมีความเสถียรอย่างต่อเนื่อง โดยมีค่า RMSE ค่อนข้างต่ำในทุกกลไกข้อมูลสูญหายที่ทดสอบ ในทางตรงกันข้าม วิธีที่มีความนิยมอย่าง MICE ให้ผลลัพธ์ที่ไม่ดีเท่าที่ควรเมื่อเทียบเท่ากับวิธีอื่นๆ ในกลุ่มวิธีดั้งเดิม ในขณะที่วิธีการเรียนรู้เชิงลึกบางวิธีแม้จะมีศักยภาพในเชิงทฤษฎีที่ดี แต่ในทางปฏิบัติมีความไวต่อการเปลี่ยนแปลงของลักษณะข้อมูล ทำให้ประสิทธิภาพที่ได้มีความผันผวนและไม่แสดงถึงข้อได้เปรียบที่ชัดเจนเมื่อเทียบกับวิธีที่เรียบง่ายกว่า

ภาพรวมผลลัพธ์ด้านการติดตามทรัพยากรการคำนวณ

จากภาพที่ 4 และ 5 วิธีการทางสถิติแบบดั้งเดิม (Mean, Median) มีประสิทธิภาพสูงในด้านการใช้ทรัพยากร ได้แก่ ใช้เวลาน้อยมาก (ระดับมิลลิวินาที) และใช้หน่วยความจำน้อยที่สุด (<1 MB) ทำให้เหมาะสมกับการใช้งานในสถานพยาบาลจริงที่มีข้อจำกัดด้านทรัพยากร ในทางตรงข้าม วิธีการที่มีความซับซ้อนมากขึ้น เช่น k-NN, MICE, MissForest และ HyperImpute ต้องใช้เวลาประมวลผลและหน่วยความจำสูง โดยเฉพาะในข้อมูลขนาดใหญ่ วิธีการเรียนรู้เชิงลึก เช่น AEImputer และ MIWAE มีการใช้เวลาหลายร้อยถึงพันวินาที และหน่วยความจำหลายสิบถึงหลายร้อยเมกะไบต์ ในขณะที่ GAIN แม้จะมีประสิทธิภาพในแง่การประหยัดทรัพยากรการคำนวณค่อนข้างดีกว่า แต่ก็ยังใช้ทรัพยากรสูงกว่าแบบดั้งเดิม

สรุปและอภิปรายผลการวิจัย

จากผลการศึกษาเปรียบเทียบการแทนค่าน้ำหนักสูญหายด้วยวิธีการเรียนรู้เชิงลึกกับวิธีดั้งเดิมด้วยวิธีการจำลองข้อมูลในบริบทของผู้ป่วยในและหอผู้ป่วยหนักภายในโรงพยาบาลนี้สามารถอภิปรายผลได้ ดังนี้

วิธีแทนค่าสูญหายแบบดั้งเดิม เช่น วิธีการแทนค่าด้วยค่าเฉลี่ย (Mean), ค่ามัธยฐาน (Median) และวิธี k-NN มีความแม่นยำที่คงที่และสม่ำเสมอในทุกสถานการณ์ของข้อมูลที่สูญหาย โดยเฉพาะอย่างยิ่ง วิธี Mean และ Median สามารถรักษาระดับ RMSE และ MAPE ที่ค่อนข้างต่ำได้ แม้ในสภาวะที่มีข้อมูลสูญหายระดับสูง (40%-60%) อีกทั้งยังมีความเร็วในการประมวลผลที่สูงมาก (ระดับมิลลิวินาที) และมีการใช้หน่วยความจำต่ำมาก (น้อยกว่า 1 MB) ทำให้วิธีเหล่านี้เหมาะสมกับการใช้งานในสภาพแวดล้อมทางคลินิกที่ต้องการผลลัพธ์อย่างรวดเร็ว ส่วนวิธี k-NN แม้ดูเรียบง่าย แต่กลับใช้หน่วยความจำสูงมาก เมื่อขนาดของตัวอย่างข้อมูลเพิ่มขึ้น ซึ่งอาจเป็นปัญหาในการใช้งานกับชุดข้อมูลขนาดใหญ่ในชีวิตจริง เช่นเดียวกับที่ Psychogyios และคณะ (2022) ได้กล่าวถึงข้อจำกัดในการขยายขนาด (Scalability) ของวิธีการเรียนรู้แบบง่าย ทำให้ต้องคำนึงถึงการแลกเปลี่ยนระหว่างความแม่นยำและการใช้ทรัพยากรในแต่ละวิธี (Psychogyios et al., 2022)

ในทางตรงกันข้าม MICE และ MissForest แม้จะมีความแม่นยำโดยทั่วไปค่อนข้างสูง แต่กลับมีการใช้ทรัพยากรในการประมวลผลสูงมาก ทั้งในด้านของเวลาและหน่วยความจำ ตัวอย่างเช่น วิธี MICE ใช้หน่วยความจำสูงถึง 150 MB หรือมากกว่า ในตัวอย่างขนาดเล็ก และใช้เวลาคำนวณที่นานอย่างชัดเจนในกรณีที่ตัวอย่างใหญ่ขึ้นหรือข้อมูลมีอัตราการสูญหายสูง ซึ่งสอดคล้องกับการศึกษา CALIBER ก่อนหน้านี้ โดย Shah et al. (2014) ว่า MissForest ที่ประมาณค่าด้วยอัลกอริทึม Random Forest มีความสามารถจับความไม่เป็นเชิงเส้นและปฏิสัมพันธ์ได้ดีมากกว่าวิธีการที่ใช้การถดถอยที่ง่ายกว่า (MICE) จึงมีความยืดหยุ่นกว่า ทำงานได้ดีกว่าวิธีแบบพารามิเตอร์ล้วน (Shah et al., 2014) และจากผลลัพธ์จากการศึกษานี้ยืนยันว่า เมื่อพิจารณา 2 วิธีนี้ แลกเปลี่ยนในแง่ความแม่นยำและการใช้ทรัพยากรแล้ว MissForest เป็นวิธีการแทนค่าสูญหายที่ดีกว่า MICE ชัดเจน นอกจากนี้ยังมีความน่าสนใจในการนำไปใช้งานต่อยอด

ในการพัฒนาแบบทำนายพยากรณ์ อย่างการศึกษา Framingham Heart Study โดย Psychogyios et al. (2022) ใช้ชุดข้อมูล Framingham เพื่อเปรียบเทียบกลยุทธ์การประมาณค่าที่สูญหายต่างๆ สำหรับทำนายความเสี่ยงโรคหัวใจ พบว่าวิธีที่ MissForest มีประสิทธิภาพ F1-scores ที่ดีกว่า k-NN และ GAIN (Psychogyios et al., 2022) และอีกการศึกษาโดย Li และคณะ (2024) ทำการประเมินอัลกอริทึมการประมาณค่าที่สูญหายแปดวิธี ในชุดข้อมูลโรคหัวใจในโลกแห่งความเป็นจริง โดยสรุปว่า Random Forest และ k-NN ให้ประสิทธิภาพที่เหนือกว่าวิธีอื่นๆ ทั้งในแง่ของ RMSE และ AUC (Li et al., 2024) สำหรับวิธีผสมผสานอย่าง HyperImpute ที่มีทฤษฎีรองรับที่น่าสนใจ ให้ผลลัพธ์ในภาพรวมที่ค่อนข้างดีเช่นกัน สามารถทำงานได้ดีในช่วงอัตราการสูญหาย 10-30% แต่เมื่ออัตราการสูญหายสูงขึ้นก็เริ่มมีความแปรปรวนมากขึ้น และมีข้อจำกัดในแง่ของการใช้ทรัพยากรทั้งเวลาและหน่วยความจำที่ค่อนข้างสูง

ในส่วนของวิธีการเรียนรู้เชิงลึก เช่น AEImputer, MIWAE และ GAIN พบผลลัพธ์ที่หลากหลาย แม้วิธีเหล่านี้จะมีข้อได้เปรียบในเชิงทฤษฎีจากความสามารถในการเรียนรู้โครงสร้างข้อมูลที่ซับซ้อน แต่ในทางปฏิบัติกลับพบว่าประสิทธิภาพมีความแปรปรวน ไม่สม่ำเสมอและไม่ได้เหนือกว่าวิธีแบบดั้งเดิมอย่างชัดเจน โดยเฉพาะวิธี MLPRegressor ที่ให้ผลลัพธ์ไม่ดีเท่าที่ควร โดยเฉพาะในข้อมูลที่มีอัตราสูญหายสูงและตัวอย่างขนาดเล็ก ซึ่งอาจมีสาเหตุมาจากการปรับแต่งค่า Hyperparameters ได้ไม่ดีพอ นอกจากนี้วิธีกลุ่มนี้ยังต้องใช้ทรัพยากรในการคำนวณสูงมาก ซึ่งทำให้การประยุกต์ใช้งานจริงในสภาพแวดล้อมทางคลินิกที่จำกัดด้วยทรัพยากรไม่เหมาะสม ผลลัพธ์นี้สอดคล้องกับบททบทวนจาก Liu et al. (2023) และ Sun et al. (2023) ที่ชี้ให้เห็นว่าวิธีการเรียนรู้เชิงลึกไม่ได้ให้ผลลัพธ์ที่ดีกว่าวิธีทั่วไปอย่างสม่ำเสมอ โดยเฉพาะเมื่อจำนวนตัวอย่างน้อยหรือมีข้อมูลสูญหายสูง

ข้อเสนอแนะที่ได้รับจากการวิจัย

จากผลการศึกษาเปรียบเทียบการแทนค่าน้ำหนักสูญหายด้วยวิธีการเรียนรู้เชิงลึกกับวิธีดั้งเดิมด้วยวิธีการจำลองข้อมูลในบริบทของผู้ป่วยในและหอผู้ป่วยหนักภายในโรงพยาบาล ผู้วิจัยมีข้อเสนอแนะ ดังนี้

1) วิธีการประมาณค่าข้อมูลสูญหายแบบดั้งเดิม อย่าง Mean และ Median แสดงให้เห็นถึงประสิทธิภาพที่มีความเสถียร แม่นยำ และประหยัดทรัพยากรในการจัดการข้อมูลน้ำหนักตัวที่หายไปในข้อมูลบริบทของผู้ป่วยในและหอผู้ป่วยหนักภายในโรงพยาบาลที่จำลองขึ้น

2) วิธีอย่าง MissForest ซึ่งสามารถแลกเปลี่ยนระหว่างประสิทธิภาพและทรัพยากรการคำนวณได้ดี และจากการทบทวนวรรณกรรมพบว่าวิธีนี้นั้นส่งผลถึงประสิทธิภาพของแบบทำนายพยากรณ์ความเสี่ยงต่างๆ ด้วย จึงถือเป็นอีกวิธีที่มีความน่าสนใจในการใช้งานและพัฒนาต่อยอด นอกจากนี้ความก้าวหน้าของวิธีการแบบผสมผสาน ซึ่งรวมจุดเด่นด้านความสามารถในการตีความและประสิทธิภาพของวิธีการแบบดั้งเดิมเข้ากับโมเดลของเทคนิคการเรียนรู้เชิงลึกอย่าง HyperImpute, RF-GAIN ถือเป็นโอกาสสำคัญในการปรับปรุงความแม่นยำและความน่าเชื่อถือในการประมาณค่าข้อมูลสูญหายต่อไปในอนาคต

3) วิธีการแทนค่าสูญหายแบบใดมักใช้ได้ดีที่สุดนั้นขึ้นอยู่กับขนาดตัวอย่าง กลไกการสูญหายของข้อมูล และอัตราการสูญหาย โดยภาพรวมข้อสรุปและคำแนะนำทั้งหมดดังแสดงในตารางที่ 3

ตารางที่ 3 แสดงข้อสรุปและคำแนะนำสำหรับวิธีการแทนค่าสูญหายที่เหมาะสมกับกลไกและอัตราการสูญหาย

วิธีการประมาณค่า	ขนาดตัวอย่าง	กลไกการสูญหาย	อัตราการสูญหาย
Mean	ทุกขนาด	MCAR, MAR, MNAR	ทุกอัตราการสูญหาย
Median	ทุกขนาด	MCAR, MAR, MNAR	ทุกอัตราการสูญหาย
k-NN	เล็ก	MCAR, MAR, MNAR	ต่ำ (10-20%)
MICE	เล็กถึงกลาง	MAR	ต่ำ-กลาง (10-40%)
MissForest	ทุกขนาด	MCAR, MAR, MNAR	ทุกอัตราการสูญหาย
HyperImpute	ทุกขนาด	MAR	ต่ำ-กลาง (10-40%)

วิธีการประมาณค่า	ขนาดตัวอย่าง	กลไกการสูญหาย	อัตราการสูญหาย
MLPRegressor	กลางถึงใหญ่	MAR	ต่ำ-กลาง (10-40%)
AEImputer	กลางถึงใหญ่	MCAR, MAR, MNAR	กลาง-สูง (30-60%)
MIWAE	กลางถึงใหญ่	MCAR, MAR, MNAR	กลาง-สูง (30-60%)
GAIN	กลางถึงใหญ่	MAR, MNAR	ต่ำ-กลาง (10-40%)

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1) ทิศทางสำหรับการวิจัยต่อไปควรมุ่งเน้นไปที่การตรวจสอบความถูกต้องและความน่าเชื่อถือของผลการศึกษานี้โดยการใช้ข้อมูลจากโรงพยาบาลจริง ซึ่งจะทำให้สามารถประเมินประสิทธิภาพและความเหมาะสมของแต่ละวิธีการได้อย่างแม่นยำมากขึ้น นอกจากนี้ ควรศึกษาเพิ่มเติมเกี่ยวกับผลกระทบของวิธีการประมาณค่าที่มีต่อการวิเคราะห์ต่อยอด เช่น การสร้างแบบทำนายพยากรณ์ หรือการสนับสนุนการตัดสินใจทางคลินิก เพื่อให้สามารถเข้าใจถึงประโยชน์และข้อจำกัดของวิธีการแต่ละประเภทเมื่อประยุกต์ใช้งานจริง

2) ยิ่งไปกว่านั้น จากแนวทางการศึกษาล่าสุดที่แนะนำวิธีการแบบผสมผสานซึ่งผนวกข้อดีของวิธีทางสถิติแบบดั้งเดิมกับวิธีการเรียนรู้เชิงลึก เช่น HyperImpute (Jarrett et al., 2022) ที่สามารถเลือกโมเดลที่เหมาะสมอัตโนมัติ RF-GAIN ที่ทำการรวม Random Forest เข้ากับ GANs (Ou et al., 2024) และ MIVAE ที่นำหลักการ Variational Autoencoders มาใช้ร่วมกับการประมาณค่าหลายครั้ง (Ma et al., 2023) ถือเป็นแนวทางที่ควรศึกษาและประเมินประสิทธิภาพอย่างละเอียด เพื่อเพิ่มประสิทธิภาพในการประมาณค่าข้อมูลสูญหายในอนาคต

3) ท้ายสุดนี้ การศึกษาเพิ่มเติมเกี่ยวกับการจัดการข้อมูลสูญหายที่เกิดจากกลไกที่ผสมผสานหลายรูปแบบ (Mixed Missingness Mechanisms) จะช่วยเสริมสร้างความแข็งแกร่งและเพิ่มความน่าเชื่อถือของวิธีการประมาณค่าเมื่อถูกนำไปใช้กับข้อมูลทางคลินิกที่มีความซับซ้อนสูง

เอกสารอ้างอิง

Albu, E., Gao, S., Wynants, L., & Van Calster, B. (2024). missForestPredict--Missing data imputation for prediction settings. *arXiv preprint arXiv:2407.03379*.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software, 45*, 1-67.

Dai, Z., Liu, S., Wu, J., Li, M., Liu, J., & Li, K. (2020). Analysis of adult disease characteristics and mortality on MIMIC-III. *PLoS One, 15*(4), e0232176.

Jarrett, D., Cebere, B. C., Liu, T., Curth, A., & van der Schaar, M. (2022, June). Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning* (pp. 9916-9937). PMLR.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database, Sci. *Data, 3*(1), 1-9.

Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., Ding, Y., Li, Y., Jian, L., Cheng, J., & Guo, H. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Med Res Methodol, 24*(1), 41.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. Wiley.

Liu, M., Li, S., Yuan, H., Ong, M. E. H., Ning, Y., Xie, F., ... & Liu, N. (2023). Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial intelligence in medicine, 142*, 102587.

Ma, Q., Li, X., Bai, M., Wang, X., Ning, B., & Li, G. (2023). MIVAE: Multiple imputation based on variational auto-encoder. *Engineering Applications of Artificial Intelligence, 123*, 106270.

Mattei, P. A., & Frellsen, J. (2019, May). MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning* (pp. 4413-4423). PMLR.

McCoy, J. T., Kroon, S., & Auret, L. (2018). Variational autoencoders for missing data imputation with application to a simulated milling circuit. IFAC-PapersOnLine 51 (21), 141-146 (2018). ISSN 2405-8963. In *5th IFAC Workshop on Mining, Mineral and Metal Processing MMM*.

McFall, A., Peake, S. L., & Williams, P. J. (2019). Weight and height documentation: Does ICU measure up?. *Australian Critical Care, 32*(4), 314-318.

Ou, H., Yao, Y., & He, Y. (2024). Missing data imputation method combining random forest and generative adversarial imputation network. *Sensors, 24*(4), 1112.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Psychogyios, K., Ilias, L., & Askounis, D. (2022, September). Comparison of missing data imputation methods using the Framingham heart study dataset. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 1-5). IEEE.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology, 179*(6), 764-774.

Sharafoddini, A., Dubin, J. A., Maslove, D. M., & Lee, J. (2019). A new insight into missing data in intensive care unit patient profiles: observational study. *JMIR medical informatics, 7*(1), e11605.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112-118.

Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications, 227*, 120201.

Yoon, J., Jordon, J., & Schaar, M. (2018, July). Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning* (pp. 5689-5698). PMLR.

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.