

Analysis and Comparison of the Impact of Statistical Feature Reduction for Classification using Machine Learning

Yotsapat Ruangpaisarn¹, Sakulchai Saramat², Auttanit Wongjak³ and
Sukanya Chansamorn^{4,*}

¹Program in Computer Science, Rajamangala University of Technology Tawan-ok, Chonburi 20110,
Thailand; Email: yotsapat_ru@rmutto.ac.th

²Program in Computer Science, Rajamangala University of Technology Tawan-ok, Chonburi 20110,
Thailand; Email: sakulchai_sa@rmutto.ac.th

³Program in Information and Communication Technology, Rajamangala University of Technology Tawan-ok,
Chonburi 20110, Thailand; Email: auttaniti_wo@rmutto.ac.th

^{4,*}Program in Information and Communication Technology, Rajamangala University of Technology
Tawan-ok, Chonburi 20110, Thailand; Email: sukanya_ch@rmutto.ac.th

*Corresponding author: Sukanya Chansamorn (sukanya_ch@rmutto.ac.th)

Received: 30 October 2025; **Revised:** 5 March 2026;

Accepted: 8 March 2026; **Published:** 29 April 2026

Abstract

This study aims to analyze the impact of statistical feature reduction on classification performance and to identify the optimal hyperparameter configurations for predicting medical conditions. Three machine learning algorithms—Support Vector Machine (SVM), Decision Tree (DT) and K-Nearest Neighbors (KNN) were compared using the Healthcare Risk Factors Dataset. Two filter-based feature selection methods, the Kruskal–Wallis test (KW) and Mutual Information (MI), were employed to generate a full feature set and feature subsets at the 25th, 50th, and 75th percentiles. Model performance was evaluated using 5 repeats of 5-fold Stratified K-Fold Cross-Validation, integrated with systematic hyperparameter tuning. The experimental results demonstrated that SVM achieved the highest performance using the full feature set ($C = 300$), yielding an accuracy of 0.9214 ± 0.0005 , precision of 0.9212 ± 0.0005 , recall of 0.9214 ± 0.0005 , and an F1-Score of 0.9210 ± 0.0005 . This accuracy significantly outperformed DT (0.8752 ± 0.0028) and KNN (0.7916 ± 0.0015). Furthermore, reducing the feature set to the 75th percentile using the KW method maintained a high accuracy of 0.9165 ± 0.0006 , representing a marginal decline of only 0.49%. These findings indicate that SVM is highly suitable for disease diagnosis within this dataset. Moreover, statistical feature reduction effectively minimizes computational complexity without a significant loss in accuracy.

Keywords: Disease Classification, Feature Selection, Hyperparameter, Machine Learning



Copyright © 2026 Journal of Computer and Creative Technology. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC-ND).

Citation:

Ruangpaisarn, Y., Saramat, S., Wongjak, A., & Chansamorn, S. (2026). Analysis and Comparison of the Impact of Statistical Feature Reduction for Classification using Machine Learning. *Journal of Computer and Creative Technology*, 4(1), e3002. <https://doi.org/10.65205/jcct.2026.e3002>

การวิเคราะห์และเปรียบเทียบผลกระทบของการลดขนาดคุณลักษณะเชิงสถิติ สำหรับการจำแนกด้วยการเรียนรู้ของเครื่อง

ยศภัทร เรืองไพศาล¹ สกุลชาย สารมาศ² อรรถนิติ วงศ์จักร³ และ สุกัลยา ชาญสมร^{4,*}

¹สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก จังหวัดชลบุรี 20110 ประเทศไทย;
Email: yotsapat_ru@rmutto.ac.th

²สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก จังหวัดชลบุรี 20110 ประเทศไทย;
Email: sakulchai_sa@rmutto.ac.th

³สาขาวิชาเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก จังหวัดชลบุรี 20110
ประเทศไทย; Email: auttaniti_wo@rmutto.ac.th

^{4,*}สาขาวิชาเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก จังหวัดชลบุรี 20110
ประเทศไทย; Email: sukanya_ch@rmutto.ac.th

*ผู้ประพันธ์บทความ: สุกัลยา ชาญสมร (sukanya_ch@rmutto.ac.th)

รับบทความ: 30 ตุลาคม 2568; แก้ไขบทความ: 5 มีนาคม 2569;

ตอบรับบทความ: 8 มีนาคม 2569; เผยแพร่บทความ: 29 เมษายน 2569

บทคัดย่อ

การศึกษานี้มีวัตถุประสงค์เพื่อวิเคราะห์ผลกระทบของการลดขนาดคุณลักษณะเชิงสถิติต่อประสิทธิภาพการจำแนกประเภทและค้นหาชุดค่าพารามิเตอร์ที่เหมาะสมที่สุดในการทำนายภาวะทางการแพทย์ โดยเปรียบเทียบอัลกอริทึมการเรียนรู้ของเครื่อง 3 รูปแบบ ได้แก่ Support Vector Machine (SVM) Decision Tree (DT) และ K-Nearest Neighbors (KNN) และใช้ชุดข้อมูล Healthcare Risk Factors ร่วมกับวิธีการคัดเลือกคุณลักษณะแบบตัวกรอง 2 วิธี คือ Kruskal-Wallis test (KW) และ Mutual Information (MI) เพื่อสร้างชุดข้อมูลที่มีคุณลักษณะครบถ้วนและชุดข้อมูลย่อยที่ระดับเปอร์เซ็นต์ไคล์ที่ 25, 50 และ 75 ประเมินประสิทธิภาพด้วยวิธี 5 repeats of 5-fold Stratified K-Fold Cross-Validation ที่ร่วมกับการปรับจูนไฮเปอร์พารามิเตอร์อย่างเป็นระบบ ผลการทดลอง พบว่า SVM ให้ประสิทธิภาพสูงสุดเมื่อใช้คุณลักษณะครบถ้วน ($C = 300$) โดยมีค่าความถูกต้อง (Accuracy) 0.9214 ± 0.0005 ค่าความเที่ยงตรง (Precision) 0.9212 ± 0.0005 ค่าความระลึก (Recall) 0.9214 ± 0.0005 และค่าเอฟวันสกอร์ (F1-Score) 0.9210 ± 0.0005 ซึ่งค่าความถูกต้องสูงกว่า DT (0.8752 ± 0.0028) และ KNN (0.7916 ± 0.0015) อย่างมีนัยสำคัญ นอกจากนี้การลดจำนวนคุณลักษณะลงเหลือร้อยละ 75 ด้วยวิธี KW ยังคงรักษาความถูกต้องไว้ได้สูงถึง 0.9165 ± 0.0006 ซึ่งลดลงจากเดิมเพียงร้อยละ 0.49 ผลการวิจัยชี้ให้เห็นว่า SVM มีความเหมาะสมอย่างยิ่งสำหรับการวินิจฉัยโรคในชุดข้อมูลนี้ และการลดขนาดคุณลักษณะเชิงสถิติช่วยลดความซับซ้อนในการประมวลผลได้ โดยไม่สูญเสียความแม่นยำอย่างมีนัยสำคัญ

คำสำคัญ: การจำแนกภาวะทางการแพทย์, การคัดเลือกคุณลักษณะ, ไฮเปอร์พารามิเตอร์, การเรียนรู้ของเครื่อง



Copyright © 2026 Journal of Computer and Creative Technology. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC-ND).

การอ้างอิง:

ยศภัทร เรืองไพศาล, สกุลชาย สารมาศ, อรรถนิติ วงศ์จักร, และ สุกัลยา ชาญสมร. (2569). การวิเคราะห์และเปรียบเทียบผลกระทบของการลดขนาดคุณลักษณะเชิงสถิติสำหรับการจำแนกด้วยการเรียนรู้ของเครื่อง. วารสารคอมพิวเตอร์และเทคโนโลยีสร้างสรรค์, 4(1), e3002. <https://doi.org/10.65205/jcct.2026.e3002>

1. บทนำ (Introduction)

ในปัจจุบันเทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) ได้เข้ามามีบทบาทสำคัญอย่างยิ่งในการยกระดับบริการทางการแพทย์ โดยเฉพาะในด้านการสนับสนุนการวินิจฉัยและการพยากรณ์โรค (Uddin et al., 2019) อัลกอริทึมการเรียนรู้ของเครื่องหลายประเภทได้รับการพัฒนาและนำมาใช้ในการจำแนกข้อมูลทางการแพทย์ โดยอัลกอริทึมที่ได้รับความนิยมและมีการใช้งานอย่างแพร่หลาย ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) ต้นไม้ตัดสินใจ (Decision Tree: DT) เพื่อนบ้านใกล้สุด (K-Nearest Neighbors: KNN) นาอิวเบย์ (Naive Bayes: NB) แรנדอมเฟอเรสต์ (Random Forest : RF) เป็นต้น (Kavakiotis et al., 2017)

อย่างไรก็ตามข้อมูลทางการแพทย์มักประสบกับปัญหาความซับซ้อนของมิติข้อมูล (Curse of Dimensionality) (Clarke et al., 2008) ซึ่งเกิดจากการมีจำนวนคุณลักษณะ (Features) ที่มากเกินไปจนทำให้ประสิทธิภาพในการประมวลผลลดลงและอาจก่อให้เกิดปัญหาภาวะการเรียนรู้เกิน (Overfitting) (Guyon & Elisseeff, 2003) การคัดเลือกคุณลักษณะ (Feature Selection) จึงเป็นกระบวนการสำคัญในการลดจำนวนมิติของข้อมูล เพื่อให้ได้มาซึ่งชุดข้อมูลที่เหมาะสมที่สุดสำหรับการสร้างโมเดล (Remeseiro & Bolon-Canedo, 2019)

จากการทบทวนวรรณกรรมที่ผ่านมา พบว่างานวิจัยส่วนใหญ่มุ่งเน้นการเปรียบเทียบอัลกอริทึมในโจทย์การจำแนกแบบทวิภาค (Binary Classification) เช่น งานวิจัยของ Saha et al. (2020) ใช้ชุดข้อมูลโรคพาร์กินสัน งานวิจัยของ Narwane & Sawarkar (2021) ใช้ชุดข้อมูลโรคเบาหวาน เป็นต้น หรือโจทย์การสกัดคุณลักษณะของข้อมูล (Feature Extraction) เช่น งานวิจัยของ Narwane & Sawarkar (2021) ได้นำเสนอการใช้เทคนิค PCA Feature Extraction ในการจัดการกับข้อมูลที่ไม่สมดุล ซึ่งแม้ว่าจะช่วยลดความซับซ้อนของข้อมูลได้ดี แต่ทำให้เกิดความสูญเสียความหมายทางกายภาพของข้อมูลเดิม (Loss of Interpretability) ทำให้ยากต่อการอธิบายผลลัพธ์ในทางการแพทย์

เพื่อเป็นการตอบโต้ข้อจำกัดของงานวิจัยดังกล่าว งานวิจัยนี้จึงมุ่งเน้นการศึกษาเปรียบเทียบประสิทธิภาพของอัลกอริทึม SVM, DT และ KNN บนชุดข้อมูล Healthcare Risk Factors Dataset ซึ่งเป็นการขยายขอบเขตการศึกษาจากการจำแนกแบบทวิภาค (Binary) ไปสู่การจำแนกที่มีความซับซ้อนหลายคลาส (Multi-class)

นอกจากนี้ เพื่อแก้ไขปัญหาการสูญเสียความหมายของข้อมูลจากการสกัดคุณลักษณะ (Feature Extraction) งานวิจัยนี้จึงเลือกใช้วิธีการคัดเลือกคุณลักษณะ (Feature Selection) แบบ KW และ MI ในสัดส่วนที่แตกต่างกัน (25%, 50%, 75%) ร่วมกับการปรับแต่งไฮเปอร์พารามิเตอร์อย่างเป็นระบบ ซึ่งจะช่วยให้ได้ผลลัพธ์ที่แม่นยำและยังคงสามารถอธิบายความหมายทางการแพทย์ได้ เพื่อหาแนวทางที่เหมาะสมที่สุดในการเพิ่มประสิทธิภาพการวินิจฉัยทางการแพทย์

2. วัตถุประสงค์งานวิจัย (Research Objectives)

1. เพื่อวิเคราะห์ผลกระทบของการลดขนาดคุณลักษณะด้วยวิธี KW และ MI ส่งผลต่อประสิทธิภาพการจำแนกภาวะทางการแพทย์ด้วยการเรียนรู้ของเครื่อง
2. เพื่อเปรียบเทียบชุดพารามิเตอร์ที่เหมาะสมสำหรับการจำแนกภาวะทางการแพทย์ด้วยการเรียนรู้ของเครื่อง
3. เพื่อเปรียบเทียบประสิทธิภาพการเรียนรู้ของอัลกอริทึม SVM Decision Tree และ KNN

3. การทบทวนวรรณกรรมและทฤษฎีที่เกี่ยวข้อง (Literature Review)

3.1 การทบทวนวรรณกรรม

จากการรวบรวมงานวิจัยที่เกี่ยวข้องระหว่างปี ค.ศ. 2022 ถึง 2025 ดังแสดงใน Table 1.

Table 1. Summary of Related Works on Feature Selection Techniques and Classification Models

Author (Year)	Dataset	Feature Selection	Classifier	Results
Siburian et al. (2025)	Water Potability	Hyperparameter Optimization	SVM (RBF Kernel), KNN	SVM (RBF Kernel)

Table 1. Summary of Related Works on Feature Selection Techniques and Classification Models (Cont.)

Author (Year)	Dataset	Feature Selection	Classifier	Results
Salehin et al. (2024)	1. Heart disease	Wrapper (SFS,	SVM, NB, DT, KNN,	1. 85.45%
	2. Framingham	SBS), MI, Chi-Square	RF, AdaBoost	(SFS + RF) 2. 86.14% (SFS + RF)
Majed et al. (2023)	Parkinson's Disease	SMOTE	RF, KNN, SVM	94.11% (KNN)
Jothi et al. (2022)	1. Breast Cancer	Hybrid	SVM, RF, DT	1. 86% (SVM)
	Wisconsin (Diagnostic)	(ReliefF + Shapley		2. 84.96% (DT)
	2. Parkinsons	Value)		3. 96.17% (DT)
	3. Heart Disease			4. 95.78% (DT)
	4. Statlog (Heart)			5. 92.3% (SVM)
Bashir et al. (2022)	1. Microarray	Hybrid	SVM	1. 94.45%
	2. Cleveland Heart	(Filter + Wrapper + Embedded)		2. 91.00%
Prusty et al. (2022)	Cervical Cancer	N/A	SVM, RF, KNN, XGBoost	98.10% (RF)

จาก Table 1. พบประเด็นสำคัญที่เกี่ยวข้องกับการจำแนกข้อมูลทางการแพทย์และข้อมูลวิทยาศาสตร์ ดังนี้

1. ลักษณะของชุดข้อมูลและจำนวนคลาส (Datasets & Number of Classes)

งานวิจัยส่วนใหญ่ยังคงมุ่งเน้นไปที่การแก้ปัญหาแบบ ทวิภาค (Binary Classification) หรือการจำแนกข้อมูลเพียง 2 คลาส เช่น การจำแนกผู้ป่วยโรคหัวใจ (Salehin et al., 2024; Jothi et al., 2022) โรคพาร์กินสัน (Majed et al., 2023) และคุณภาพน้ำ (Siburian et al., 2025) ซึ่งสะท้อนให้เห็นว่าการจำแนกแยกแยะระหว่าง “กลุ่มปกติ” และ “กลุ่มผิดปกติ” ยังคงเป็นโจทย์หลักในปัจจุบัน แต่ยังคงขาดการศึกษาเปรียบเทียบในข้อมูลที่มีความซับซ้อนหลายกลุ่ม (Multi-Class) อย่างแพร่หลาย

2. แนวทางการวิจัยและเทคนิคการคัดเลือก (Research Approaches) โดยแนวโน้มของงานวิจัยแบ่งออกเป็น 3 กลุ่มหลัก ได้แก่

1) กลุ่มเทคนิคผสมผสาน (Hybrid Methods) เป็นแนวทางที่ได้รับความนิยมสูงในช่วงปี 2022 ดังที่พบในงานของ Jothi et al. (2022) และ Bashir et al. (2022) ซึ่งนำวิธี Filter มาผสมกับ Wrapper หรือเทคนิคขั้นสูงอย่าง Shapley Value เพื่อเพิ่มความแม่นยำ

2) กลุ่มการเปรียบเทียบและปรับสมดุล (Comparative & Data Handling) งานวิจัยปี 2023-2024 เริ่มหันมาสนใจการจัดการคุณภาพข้อมูล เช่น Majed et al. (2023) ใช้เทคนิค SMOTE แก้ปัญหาข้อมูลไม่สมดุล (Imbalanced Data) ก่อนเข้าสู่กระบวนการคัดเลือก

3) กลุ่มการปรับจูนพารามิเตอร์ (Optimization) งานวิจัยนี้ของ (Siburian et al., 2025) ชี้ให้เห็นว่าการคัดเลือก Feature อาจไม่เพียงพอ ควรควบคู่ไปกับการทำ Hyperparameter Optimization เพื่อรีดประสิทธิภาพสูงสุดของโมเดล

3.2 ทฤษฎีที่เกี่ยวข้อง

3.2.1 Kruskal-Wallis Test (KW)

Kruskal-Wallis Test (KW) เป็นเทคนิคทางสถิตินอนพารามेटริกซ์ (Non-Parametric) ที่ถูกนำมาประยุกต์ใช้เป็นเครื่องมือในการกรองเบื้องต้น (Filter Method) เพื่อประเมินความสามารถในการจำแนกของแต่ละคุณลักษณะที่เป็นตัวเลข

(Numeric Features) หรืออันดับ (Ordinal Features) เมื่อเทียบกับตัวแปรเป้าหมายที่เป็นกลุ่ม (Categorical Target Variable) เช่น สถานะ “เป็นโรคร” หรือ “ไม่เป็นโรคร” หลักการทำงานอยู่บนพื้นฐานของการเปรียบเทียบค่ามัธยฐาน (Median) ของการแจกแจงข้อมูลระหว่างกลุ่มต่าง ๆ โดยไม่ต้องอาศัยสมมติฐานว่าข้อมูลมีการแจกแจงแบบปกติ (Normal Distribution) ซึ่งทำให้เหมาะสมอย่างยิ่งสำหรับข้อมูลทางการแพทย์ที่มักมีการแจกแจงที่ไม่สมมาตรหรือมีค่าผิดปกติ (Outliers) กระบวนการทำงานเริ่มต้นด้วยการแปลงค่าข้อมูลดิบให้เป็นอันดับ (Rank) จากนั้นคำนวณสถิติทดสอบ H เพื่อประเมินความแตกต่างระหว่างกลุ่ม คุณลักษณะที่มีค่า H สูงและค่า p-value ต่ำ (โดยทั่วไปน้อยกว่า 0.05) จะถูกพิจารณาว่ามีความสัมพันธ์กับตัวแปรเป้าหมายอย่างมีนัยสำคัญ (Kruskal & Wallis, 1952)

$$H = \frac{12}{N(N+1)} \cdot \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \quad (1)$$

3.2.2 Mutual Information (MI)

Mutual Information (MI) ใช้แนวคิดจากทฤษฎีสารสนเทศ (Information Theory) ที่ถูกนำมาประยุกต์ใช้เป็นเครื่องมือในการคัดเลือกคุณลักษณะแบบการกรองเบื้องต้น โดยมีความสามารถในการวัดปริมาณข้อมูลร่วมกันระหว่างคุณลักษณะแต่ละตัวกับตัวแปรเป้าหมาย ซึ่งเป็นวิธีการที่ยืดหยุ่นกว่าการใช้สัมประสิทธิ์สหสัมพันธ์แบบดั้งเดิม เนื่องจากสามารถวัดความสัมพันธ์ได้ทั้งแบบเชิงเส้น (Linear) และไม่เป็นเชิงเส้น (Non-Linear) หลักการพื้นฐานของ Mutual Information คือการวัดว่าการรู้ค่าของคุณลักษณะหนึ่งช่วยลดความไม่แน่นอน (Uncertainty) หรือเอนโทรปี (Entropy) ของตัวแปรเป้าหมายได้มากน้อยเพียงใด หากคุณลักษณะมีค่า MI สูง แสดงว่าคุณลักษณะนั้นมีข้อมูลที่เป็นประโยชน์ต่อการทำนายตัวแปรเป้าหมายมาก ในทางกลับกัน หากค่า MI เท่ากับศูนย์ หมายความว่าคุณลักษณะนั้นเป็นอิสระจากตัวแปรเป้าหมายและไม่มีประโยชน์ในการทำนาย (Shannon, 1948)

$$MI(X; Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \cdot \log \frac{P(x, y)}{P(x) \cdot P(y)} \quad (2)$$

3.2.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) เป็นอัลกอริทึมสำหรับการจำแนกข้อมูลที่มีพื้นฐานจากการค้นหาเส้นแบ่งหรือระนาบตัดสินใจ (Hyperplane) ที่เหมาะสมที่สุดในการแยกข้อมูลออกเป็นกลุ่มต่าง ๆ โดยมีวัตถุประสงค์หลัก คือ การเพิ่มระยะขอบเขต (Margin) ระหว่างเส้นแบ่งกับจุดข้อมูลที่อยู่ใกล้ที่สุดของแต่ละกลุ่มให้มากที่สุด ซึ่งช่วยเพิ่มความสามารถของแบบจำลองในการทำนายข้อมูลใหม่ได้อย่างมีประสิทธิภาพ ในกรณีที่ข้อมูลไม่สามารถแยกออกจากกันได้ด้วยเส้นตรงในปริภูมิมิติเดิม SVM จะใช้เทคนิค Kernel Trick เพื่อแปลงข้อมูลไปยังปริภูมิมิติที่สูงขึ้น ส่งผลให้สามารถกำหนดเส้นแบ่งที่เหมาะสมสำหรับข้อมูลที่มีความซับซ้อนได้ดียิ่งขึ้น ดัง Figure 1.

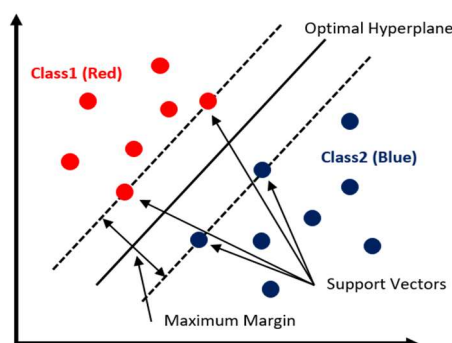


Figure 1. Support Vector Machine Classification Diagram

3.2.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) เป็นอัลกอริทึมการจำแนกข้อมูลที่ใช้หลักการของความคล้ายคลึงกัน โดยสมมติว่าข้อมูลที่มีลักษณะใกล้เคียงกันมักจะอยู่ใกล้กันหรือเกาะกลุ่มกัน อัลกอริทึมนี้จัดอยู่ในกลุ่มวิธีการเรียนรู้แบบไม่สร้างแบบจำลองล่วงหน้า (Lazy Learning) โดยที่เมื่อมีข้อมูลใหม่เข้ามาจะทำการคำนวณระยะทางระหว่างข้อมูลใหม่กับข้อมูลตัวอย่างทั้งหมด จากนั้นเลือกข้อมูลที่อยู่ใกล้ที่สุดจำนวน K ตัวอย่าง ซึ่งค่า K ส่วนใหญ่เป็นเลขคี่ การพิจารณาผลการจำแนกจะอาศัยการลงคะแนนเสียงจากข้อมูลเพื่อนบ้านเหล่านี้ โดยข้อมูลใหม่จะถูกจัดให้อยู่ในกลุ่มที่มีจำนวนเสียงมากที่สุด ดัง Figure 2.

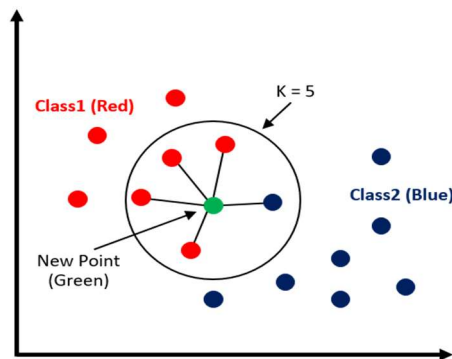


Figure 2. K-Nearest Neighbors Classification Diagram

3.2.5 Decision Tree (DT)

Decision Tree (DT) เป็นอัลกอริทึมที่สร้างแบบจำลองการจำแนกข้อมูลในรูปแบบโครงสร้างต้นไม้ โดยเริ่มต้นจากโหนดราก (Root Node) ซึ่งทำหน้าที่แบ่งข้อมูลตามเงื่อนไขหรือคุณลักษณะที่มีความสำคัญสูงสุด จากนั้นกระบวนการแบ่งข้อมูลจะดำเนินไปอย่างเป็นลำดับ ผ่านกิ่งก้านของต้นไม้จนถึงโหนดปลายทาง (Leaf Node) ซึ่งใช้สำหรับกำหนดคลาสของข้อมูล จุดเด่นของ Decision Tree คือ มีความเข้าใจง่ายและสามารถนำเสนอในรูปแบบแผนผังได้อย่างเป็นระบบ ดัง Figure 3.

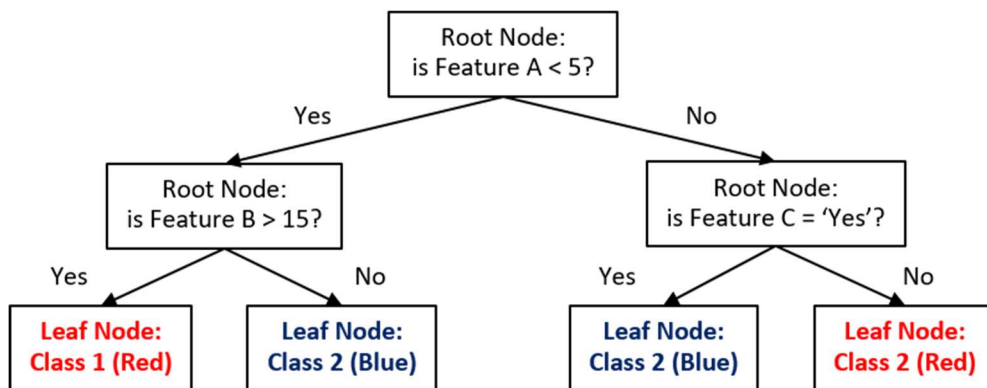


Figure 3. Decision Tree Classification Diagram

4. กรอบแนวคิดงานวิจัย (Conceptual Framework)

งานวิจัยนี้เป็นงานวิจัยแบบกรอบแนวคิดเชิงเปรียบเทียบ (Comparative Framework) โดยใช้ระบบ IPO Model กับชุดข้อมูลสุขภาพ โดยมีวัตถุประสงค์หลักคือ การเปรียบเทียบชุดคุณลักษณะ (Feature Sets) ที่ถูกคัดเลือกด้วยวิธีการที่แตกต่างกันสองแบบต่อประสิทธิภาพการจำแนกข้อมูลของการเรียนรู้ของเครื่อง (Machine Learning) ที่มีการทำซ้ำการตรวจสอบไขว้แบบแบ่งชั้นร่วมกับการปรับจูนไฮเพอร์พารามิเตอร์ เพื่อเพิ่มความเชื่อมั่นของผลลัพธ์ของการจำแนกข้อมูล

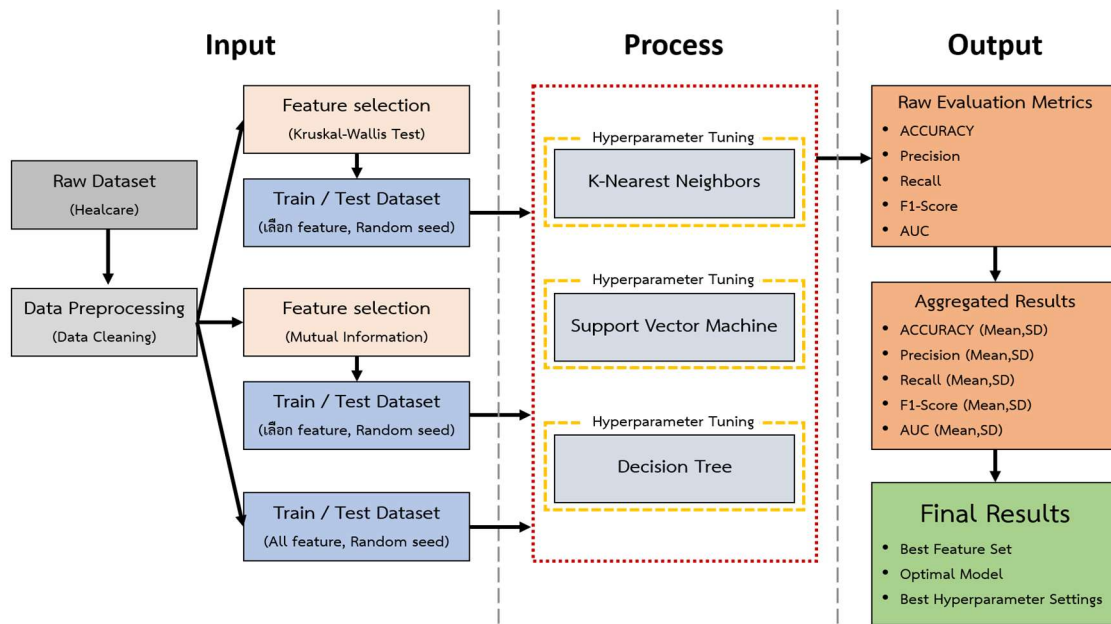


Figure 4. Conceptual Framework

จาก Figure 4. กรอบแนวคิดการวิจัย (Conceptual Framework) สามารถแบ่งออกเป็น 3 ส่วนหลัก ดังนี้

1. ส่วนนำเข้าข้อมูล (Input) เริ่มต้นจากการนำข้อมูลดิบ (Raw Dataset) มาผ่านกระบวนการเตรียมข้อมูล (Data Preprocessing) เพื่อทำความสะอาดและจัดการค่าที่ผิดปกติ จากนั้นทำการแบ่งชุดข้อมูลออกเป็น 3 กลุ่มการทดลอง ได้แก่ 1) การคัดเลือกคุณลักษณะด้วยวิธี KW 2) การคัดเลือกคุณลักษณะด้วยวิธี MI และ 3) การใช้คุณลักษณะทั้งหมด (All features) โดยมีการกำหนดค่า Random seed เพื่อควบคุมความคงที่ของการแบ่งข้อมูลสำหรับฝึกฝนและทดสอบ (Train/Test Dataset)

2. ส่วนการประมวลผล (Process) เป็นขั้นตอนการสร้างแบบจำลองการเรียนรู้ของเครื่อง 3 อัลกอริทึม ได้แก่ K-Nearest Neighbors (KNN), Support Vector Machine (SVM) และ Decision Tree (DT) โดยในขั้นตอนนี้จะมีการปรับแต่งค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Tuning) อย่างเป็นระบบสำหรับแต่ละโมเดลเพื่อให้ได้ประสิทธิภาพสูงสุด

3. ส่วนผลลัพธ์ (Output): ดำเนินการวัดผลประสิทธิภาพของโมเดลด้วยตัวชี้วัด Accuracy, Precision, Recall, F1-Score, และ AUC จากนั้นทำการรวบรวมผลลัพธ์ในรูปแบบค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน (Mean, S.D.) เพื่อนำไปสู่ข้อสรุปสุดท้ายในการระบุชุดข้อมูล โมเดล และค่าพารามิเตอร์ที่ดีที่สุด (Best Feature Set & Optimal Model)

5. วิธีดำเนินงานวิจัย (Research Methodology)

5.1 เครื่องมือและอุปกรณ์ที่ใช้ในการวิจัย (Tools and Equipment)

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ได้ใช้ระบบ Cloud Computing ของ Google Colab เป็นเครื่องมือหลักในการประมวลผลข้อมูล เนื่องจากรองรับการทำงานของภาษา Python และมีหน่วยประมวลผลกราฟิก (GPU) ที่ช่วยเร่งความเร็วในการฝึกสอนโมเดล

5.2 ข้อมูลดิบ (Raw Data)

งานวิจัยนี้ใช้ชุดข้อมูลปัจจัยเสี่ยงด้านการดูแลสุขภาพ (Healthcare Risk Factors Dataset) ที่มาจากฐานข้อมูลออนไลน์ Kaggle ซึ่งเป็นชุดข้อมูลสำหรับการวิเคราะห์ปัจจัยเสี่ยงด้านการดูแลสุขภาพ 7 ภาวะทางการแพทย์ ได้แก่ โรคข้อเสื่อม (Arthritis) หอบหืด (Asthma) โรคมะเร็ง (Cancer) โรคเบาหวาน (Diabetes) สุขภาพดี (Healthy) ความดันโลหิตสูง (Hypertension) และโรคอ้วน (Obesity) โดยมี 20 คุณลักษณะ(รวมคลาส) และ 30000 ระเบียบข้อมูล (Ahmed, 2025) ดังแสดงรายละเอียดข้อมูลใน Table 2.

Table 2. Features and Descriptions of the Healthcare Risk Factors Dataset

Feature	Description
Age	Patient's age (in years)
Gender	Male / Female
Medical Condition	Reported health condition (e.g., Diabetes, Hypertension, Asthma, Obesity, Healthy)
Glucose	Blood glucose level
Blood Pressure	Blood pressure measurement
BMI	Body Mass Index
Oxygen Saturation	Blood oxygen saturation level
LengthOfStay	Hospital length of stay (days)
Cholesterol	Cholesterol level
Triglycerides	Triglyceride level
HbA1c	Hemoglobin A1c (glycated hemoglobin)
Smoking	Smoking status (0 = Non-smoker, 1 = Smoker)
Alcohol	Alcohol consumption (0 = No, 1 = Yes)
Physical Activity	Physical activity (approx. hours/week)
Diet Score	Diet quality score (numeric)
Family History	Family medical history (0 = No, 1 = Yes)
Stress Level	Stress level (numeric scale)
Sleep Hours	Average sleep hours per day
random_notes	Random notes (lorem / ipsum / ###)
noise_col	Noise column (unrelated random values)

5.3 การเตรียมข้อมูล (Data Preprocessing)

การเตรียมข้อมูลมีการกำจัดคอลัมน์ Random Notes และ Noise Col ซึ่งเป็นคอลัมน์ที่ไม่เกี่ยวข้อง เพื่อลดสัญญาณรบกวน กำจัดคอลัมน์ Gender เพื่อให้ใช้งานกับวิธีคัดเลือกคุณลักษณะและการจัดการกับค่าว่าง (Missing Values) โดยใช้วิธีการลบแถวที่มีค่าว่างทั้งหมด รวมถึงแถวในคอลัมน์ class ที่เป็นค่าว่างด้วยเพื่อให้มั่นใจว่าข้อมูลที่นำมาใช้เป็นข้อมูลที่สมบูรณ์เท่านั้น จำนวนข้อมูลก่อนและหลังการเตรียมข้อมูลได้แสดงรายละเอียดข้อมูลใน Table 3.

Table 3. Data Preparation Summary by Medical Condition

Class (Medical Condition)	Pre-Processed Count (rows)	Post-Processed Count (rows)	Post-Processed (%)
Arthritis	1796	921	51.28
Asthma	2037	1061	52.09
Cancer	1234	610	49.43
Diabetes	6417	3396	52.92
Healthy	3039	1602	52.71
Hypertension	7120	3748	52.64
Obesity	3857	2017	52.29
(blank)	4500	0	0.00
Total	30000	13355	44.52

5.4 การคัดเลือกคุณลักษณะ (Feature selection: FS)

การคัดเลือกคุณลักษณะในงานวิจัยนี้ได้ใช้ DEFAULT SEED ที่ 42 เพื่อความคงที่ของผลลัพธ์ร่วมกับวิธีการคัดเลือกคุณลักษณะที่แตกต่างกัน 2 วิธี คือ Kruskal-Wallis Test (KW) ซึ่งเป็นสถิติที่ทดสอบสมมติฐานความแตกต่างของมัธยฐานระหว่างกลุ่มของตัวแปรเป้าหมาย กับวิธี Mutual Information (MI) ที่สามารถวัดความสัมพันธ์เชิงข้อมูลที่สามารถจับความสัมพันธ์แบบไม่เชิงเส้น (Non-Linear Dependency) ระหว่างคุณลักษณะกับตัวแปรเป้าหมายได้ คุณลักษณะทั้งหมดจะถูกจัดเรียงตามค่าสถิติ (H-Statistic หรือ MI Score) ที่ได้จากแต่ละวิธีและบันทึกผลลัพธ์การจัดอันดับเพื่อใช้ในการสร้างชุดข้อมูลย่อยต่อไป

5.5 การสร้างชุดคุณลักษณะย่อย

ชุดคุณลักษณะที่ถูกจัดอันดับจาก KW และ MI จะถูกนำมาสร้างเป็นชุดคุณลักษณะย่อย (Feature Subsets) ของ KW ตามเกณฑ์สัดส่วนของเปอร์เซ็นต์ไทล์ (Percentile) ที่ 25% 50% 75% จำนวน 3 ชุด ชุดคุณลักษณะย่อยของ MI จำนวน 3 ชุด ซึ่งเป็นชุดข้อมูลที่สะท้อนถึงระดับความเข้มข้นของคุณภาพข้อมูล เมื่อนำมาเปรียบเทียบกับชุดคุณลักษณะทั้งหมดอีก 1 ชุด รวมเป็นจำนวนทั้งหมด 7 ชุด เพื่อศึกษาผลกระทบของรูปแบบคุณลักษณะ จำนวนคุณลักษณะมีผลต่อประสิทธิภาพการจำแนกข้อมูลของแบบจำลองอย่างไร

5.6 การตรวจสอบความถูกต้องด้วยการตรวจสอบไขว้

การตรวจสอบความถูกต้องของการเรียนรู้ของเครื่องในงานวิจัยนี้ เพื่อให้แบบจำลองมีความน่าเชื่อถือและลดความเอนเอียง (Bias) จากความไม่สมดุลของข้อมูล (Class Imbalance) งานวิจัยนี้ได้กำหนดขั้นตอนการตรวจสอบความถูกต้องด้วยวิธี Stratified 5-Fold Cross-Validation (SKCV) (Prusty et al., 2022) โดยมีลำดับขั้นตอนดังนี้

1. การกำหนดค่าเริ่มต้น (Initialization) กำหนดให้มีการทดสอบซ้ำจำนวน 5 รอบ (Runs) เพื่อตรวจสอบความเสถียรของแบบจำลอง โดยแต่ละรอบจะกำหนดค่า Random Seed ที่แตกต่างกัน คือ [12, 22, 32, 42, 52]
2. การแบ่งส่วนข้อมูลแบบแบ่งชั้น (Stratified Data Partitioning) ในแต่ละรอบการทดสอบ ข้อมูลจะถูกแบ่งออกเป็น 5 ส่วนย่อย (Folds) โดยใช้หลักการ Stratification เพื่อควบคุมให้สัดส่วนของแต่ละคลาส (Class Distribution) ในทุก Fold มีความสมดุลและสอดคล้องกับชุดข้อมูลตั้งต้น
3. กระบวนการเรียนรู้และทดสอบ (Training and Testing Loop) ดำเนินการวนซ้ำ 5 ครั้ง (Iterations) ตามจำนวน Fold โดยในแต่ละครั้งจะมีกระบวนการดังนี้
 - 1) ใช้ข้อมูลจำนวน 4 ส่วน เป็นชุดข้อมูลฝึกสอน (Training Set)
 - 2) ใช้ข้อมูลจำนวน 1 ส่วน ที่เหลือ เป็นชุดข้อมูลทดสอบ (Testing Set)
4. การประมวลผลผลลัพธ์ (Performance Aggregation) นำผลการจำแนกที่ได้จากการทดสอบทั้ง 5 ครั้ง ในแต่ละรอบ Seed มาคำนวณหาค่าเฉลี่ย เพื่อใช้เป็นตัวแทนประสิทธิภาพของแบบจำลองในการรายงานผลขั้นสุดท้าย

5.7 การฝึกและปรับจูนการเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่องที่ใช้ในงานวิจัยนี้ คือ เพื่อนบ้านใกล้สุด (K-Nearest Neighbors: KNN) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) และต้นไม้ตัดสินใจ (Decision Tree: DT) โดยชุดคุณลักษณะย่อยทั้ง 7 ชุดจะถูกฝึกด้วยแบบจำลองเหล่านี้ ภายใต้เงื่อนไข 5-Fold SKF ที่มีการปรับจูนไฮเพอร์พารามิเตอร์ (Hyperparameter) โดยมีการกำหนดกรอบการปรับค่าพารามิเตอร์ ดังต่อไปนี้ KNN ใช้ค่า K ที่ [3, 5, 7, 9, 11] ในส่วนของ SVM กำหนดค่า C เป็น [1, 10, 100, 200, 300] สำหรับ DT ใช้ค่า Max Depth ที่ [5, 10, 15, 20, None] ซึ่งมีวัตถุประสงค์เพื่อค้นหาชุดพารามิเตอร์ที่เหมาะสมที่สุด สำหรับการจำแนกสภาวะการแพทย์ของแต่ละแบบจำลอง

5.8 การประเมินประสิทธิภาพ

การวัดประสิทธิภาพของการเรียนรู้ของเครื่องที่ผ่านการฝึกและปรับจูนในงานวิจัยนี้จะถูกประเมิน โดยใช้ชุดของเมตริกวัดผล 3 รายการ ได้แก่ ความแม่นยำ (Accuracy) คือ ค่าสัดส่วนของข้อมูลทั้งหมดที่โมเดลทำนายได้ถูกต้อง ค่าเอฟ1-วันสกอร์ (F1-Score) เป็นค่าเฉลี่ยแบบฮาร์มอนิก (Harmonic Mean) ระหว่าง ความเที่ยงตรง (Precision) และความไว (Recall) ซึ่งเป็นเมตริกที่เหมาะสมสำหรับการประเมินแบบจำลองบนชุดข้อมูลที่ไม่สมดุล (Imbalance) นอกจากนี้ยังมีการคำนวณพื้นที่ใต้เส้นโค้ง (Area Under the Curve: AUC) ซึ่งเป็นค่าที่บอกประสิทธิภาพในการแยกแยะคลาสของแบบจำลอง

5.9 การสรุปและรายงานผลลัพธ์

ขั้นตอนนี้เปรียบเสมือนการวัดประสิทธิภาพของชุดคุณลักษณะร่วมกับชุดพารามิเตอร์ของแบบจำลองแต่ละชนิด คือ การรวมผลการทดลองทั้งหมดจาก 5 ค่าเริ่มต้นแบบสุ่ม (Random Seed) แล้วหาค่าเฉลี่ย (Mean) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: S.D.) ของเมตริกประสิทธิภาพทั้งหมด เพื่อบอกประสิทธิภาพโดยเฉลี่ยและความเสถียร (Robustness) ต่อความแปรปรวนของการแบ่งชุดคุณลักษณะข้อมูล

6. ผลการวิจัย (Results)

6.1 ผลการวิเคราะห์ผลกระทบของการลดขนาดคุณลักษณะด้วยวิธี KW และ MI ส่งผลต่อประสิทธิภาพการจำแนกภาวะทางการแพทย์ด้วยการเรียนรู้ของเครื่อง

6.1.1 ผลการจัดอันดับและสร้างชุดคุณลักษณะด้วย Kruskal-Wallis Test

ชุดข้อมูลหลังการจัดอันดับความสำคัญทั้ง 16 คุณลักษณะด้วยวิธี Kruskal-Wallis Test จะได้ H Statistic ซึ่งเป็นตัววัดความสำคัญของคุณลักษณะและผลการทดสอบยังยืนยันว่า คุณลักษณะทั้งหมดมีความสัมพันธ์อย่างมีนัยสำคัญทางสถิติกับตัวแปรเป้าหมาย ($p < 0.05$) อีกด้วย โดยที่คุณลักษณะที่มีความสำคัญสูงสุดคือ HbA1c (7202.26) Glucose (6950.96) และ LengthOfStay (6796.18) ตามลำดับ ทั้งนี้การจัดอันดับจะถูกนำไปใช้เพื่อสร้างชุดคุณลักษณะย่อยตามสัดส่วน 25% (4 คุณลักษณะ) 50% (8 คุณลักษณะ) และ 75% (12 คุณลักษณะ) ดังแสดงรายละเอียดข้อมูลใน Table 4.

Table 4. Kruskal-Wallis Test Feature Ranking and Subset Creation

Feature	H Statistic	Significant ($p < 0.05$)	Percentile		
			25%	50%	75%
HbA1c	7202.26	Yes	/	/	/
Glucose	6950.96	Yes	/	/	/
LengthOfStay	6796.18	Yes	/	/	/
Physical_Activity	6242.59	Yes	/	/	/
Blood_Pressure	5673.96	Yes		/	/
Age	5310.00	Yes		/	/
Diet_Score	5269.80	Yes		/	/
BMI	3932.52	Yes		/	/
Stress_Level	3166.27	Yes			/
Sleep_Hours	2997.73	Yes			/
Cholesterol	2995.98	Yes			/
Triglycerides	2739.62	Yes			/
Oxygen_Saturation	1870.24	Yes			
Family_History	913.78	Yes			
Smoking	851.15	Yes			
Alcohol	300.47	Yes			

6.1.2 ผลการจัดอันดับและสร้างชุดคุณลักษณะด้วย Mutual Information

ชุดข้อมูลหลังการจัดอันดับความสำคัญทั้ง 16 คุณลักษณะด้วยวิธี Mutual Information จะได้ MI Score ซึ่งเป็นตัววัดความสำคัญของคุณลักษณะ ซึ่งคุณลักษณะที่มีความสำคัญสูงสุดคือ LengthOfStay (0.5995) ตามมาด้วย

Glucose (0.4591) และ HbA1c (0.4464) ตามลำดับ ทั้งนี้การจัดอันดับนี้ถูกนำไปใช้เพื่อสร้างชุดคุณลักษณะย่อยตามสัดส่วน 25% (4 คุณลักษณะ) 50% (8 คุณลักษณะ) และ 75% (12 คุณลักษณะ) ดังแสดงรายละเอียดข้อมูลใน Table 5.

Table 5. Mutual Information Feature Ranking and Subset Creation

Feature	MI_Score	Percentile		
		25%	50%	75%
LengthOfStay	0.5995	/	/	/
Glucose	0.4591	/	/	/
HbA1c	0.4464	/	/	/
Age	0.4180	/	/	/
Physical_Activity	0.3534		/	/
Diet_Score	0.3109		/	/
Blood_Pressure	0.2918		/	/
BMI	0.2066		/	/
Triglycerides	0.1848			/
Cholesterol	0.1552			/
Oxygen_Saturation	0.1481			/
Sleep_Hours	0.1394			/
Stress_Level	0.1368			
Family_History	0.0341			
Smoking	0.0215			
Alcohol	0.0091			

6.1.3 ผลการประเมินประสิทธิภาพเปรียบเทียบความแม่นยำระหว่างชุดคุณลักษณะ KW และ MI สำหรับแบบจำลอง SVM และ DT

การเปรียบเทียบประสิทธิภาพความแม่นยำระหว่างชุดคุณลักษณะ KW และ MI สำหรับแบบจำลอง SVM และ DT โดยดูเฉพาะชุดคุณลักษณะที่ให้ความแม่นยำสูงที่สุด นอกเหนือจากชุดที่มีคุณลักษณะครบทั้งหมด

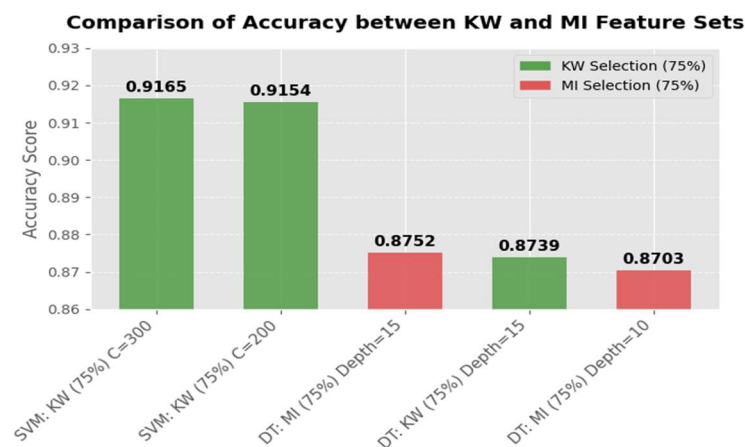


Figure 5. Comparison of Accuracy between KW and MI Feature Sets for SVM and Decision Tree

จาก Figure 5. แสดงให้เห็นว่าชุดคุณลักษณะที่ 75% ของ KW ในการจำแนกด้วย SVM แสดงประสิทธิภาพความแม่นยำที่สูง โดยมีประสิทธิภาพที่สูงเกินกว่า 90% ขึ้นไป ในขณะที่ชุดคุณลักษณะที่ 75% ของ MI ในการจำแนกด้วย DT ให้ค่าความแม่นยำที่ 87.52% และนอกจากนี้การไม่เห็นชุดคุณลักษณะที่ 25% 50% ของทั้ง KW และ MI ติดอันดับทำให้สามารถบอกได้ว่าจำนวนคุณลักษณะที่ลดลงส่งผลต่อประสิทธิภาพในการจำแนก

6.2 ผลการเปรียบเทียบชุดพารามิเตอร์ที่เหมาะสมสำหรับการจำแนกภาวะทางการแพทย์ด้วยการเรียนรู้ของเครื่อง

6.2.1 ผลการประเมินประสิทธิภาพของการจำแนกข้อมูลด้วย Support Vector Machine (SVM)

ผลการประเมินประสิทธิภาพของการจำแนกข้อมูลด้วย Support Vector Machine (SVM) พบว่าชุดคุณลักษณะทั้งหมด (100%) และค่า C ที่ 300 ให้ค่า Accuracy เฉลี่ยสูงสุดที่ 0.9214 และค่า AUC ที่ 0.9935 ตามลำดับ ประสิทธิภาพจะลดลงเล็กน้อยเมื่อค่า C มีค่า 200 จะให้ค่า Accuracy เฉลี่ยสูงสุดที่ 0.9203 และหากค่า C มีค่าที่ 100 ค่า Accuracy เฉลี่ยสูงสุดจะอยู่ที่ 0.9158 อย่างไรก็ตามผลลัพธ์ที่น่าสนใจคือ ชุดคุณลักษณะด้วยวิธี Kruskal-Wallis test ที่ 75% ใช้ค่า C ที่ 300 ให้ค่า Accuracy เฉลี่ยสูงสุดที่ 0.9165 แสดงให้เห็นถึงการลดขนาดข้อมูลยังคงมีประสิทธิภาพการจำแนกที่มีความแม่นยำสูงอยู่ ดังแสดงรายละเอียดข้อมูลใน Table 6.

Table 6. Top 5 Performance Evaluation Results of the Support Vector Machine (SVM) Classifier

FS	Feature Used (Percentile)	C	Accuracy Mean	Precision Mean	Recall Mean	F1_Score Mean	AUC Mean
	16 (100%)	300	0.9214	0.9212	0.9214	0.9210	0.9935
	16 (100%)	200	0.9203	0.9212	0.9214	0.9198	0.9933
KW	12 (75%)	300	0.9165	0.9202	0.9203	0.9159	0.9927
	16 (100%)	100	0.9158	0.9202	0.9203	0.9153	0.9930
KW	12 (75%)	200	0.9154	0.9160	0.9165	0.9148	0.9926

6.2.2 ผลการประเมินประสิทธิภาพของการจำแนกข้อมูลด้วย Decision Tree (DT)

ผลการประเมินประสิทธิภาพของการจำแนกข้อมูลด้วย Decision Tree (DT) พบว่าชุดคุณลักษณะทั้งหมด 16 คุณลักษณะและค่า Max depth ที่ 15 ให้ค่า Accuracy เฉลี่ยสูงสุดที่ 0.8752 และค่า AUC ที่ 0.9353 ตามลำดับ และจะเห็นว่าชุดคุณลักษณะที่ 75% ทั้งวิธี Kruskal-Wallis test และ Mutual Information มีประสิทธิภาพการจำแนกข้อมูลด้วยค่า Max depth ที่ 15 ให้ค่า Accuracy เฉลี่ยสูงสุดที่ 0.8715 และ Accuracy เฉลี่ยสูงสุดที่ 0.8739 ตามลำดับ ซึ่งจะเห็นว่า Mutual Information มีประสิทธิภาพดีกว่า Kruskal-Wallis test เล็กน้อยเมื่อใช้งานร่วมกับ Decision Tree ดังแสดงรายละเอียดข้อมูลใน Table 7.

Table 7. Top 5 Performance Evaluation Results of the Decision Tree (DT) Classifier

FS	Feature used (Percentile)	Max depth	Accuracy Mean	Precision Mean	Recall Mean	F1_Score Mean	AUC Mean
	16 (100%)	15	0.8752	0.8774	0.8752	0.8760	0.9353
MI	12 (75%)	15	0.8752	0.8775	0.8752	0.8761	0.9351
KW	12 (75%)	15	0.8739	0.8773	0.8739	0.8752	0.9378
	16 (100%)	10	0.8715	0.8737	0.8715	0.8723	0.9317
MI	12 (75%)	10	0.8703	0.8747	0.8703	0.8714	0.9598

6.2.3 ผลการประเมินประสิทธิภาพของการจำแนกข้อมูลด้วย K-Nearest Neighbors (KNN)

ผลการประเมินประสิทธิภาพของการจำแนกข้อมูลด้วย K-Nearest Neighbors (KNN) พบว่าชุดคุณลักษณะทั้งหมด 5 อันดับแรกเป็นชุดคุณลักษณะที่ 50% หรือใช้คุณลักษณะเพียง 8 คุณลักษณะจาก 16 คุณลักษณะและการคัดเลือก

คุณลักษณะทั้งวิธี Kruskal-Wallis test และ Mutual Information ให้ผลลัพธ์ที่เท่ากัน อย่างไรก็ตามค่า Accuracy เฉลี่ยสูงสุดที่ได้คือ 0.7916 ที่ค่า K มีค่าเป็น 9 ดังแสดงรายละเอียดข้อมูลใน Table 8.

Table 8. Top 5 Performance Evaluation Results of the K-Nearest Neighbors (KNN) Classifier

FS	Features used (Percentile)	K	Accuracy Mean	Precision Mean	Recall Mean	F1_Score Mean	AUC Mean
KW / MI	8 (50)	9	0.7916	0.7943	0.7916	0.7897	0.9470
KW / MI	8 (50)	7	0.7908	0.7949	0.7908	0.7901	0.9409
KW / MI	8 (50)	11	0.7903	0.7924	0.7903	0.7875	0.9507
KW / MI	8 (50)	5	0.7890	0.7944	0.7890	0.7895	0.9305
KW / MI	8 (50)	3	0.7803	0.7932	0.7803	0.7847	0.9098

6.3 ผลการเปรียบเทียบประสิทธิภาพการเรียนรู้ของอัลกอริทึม SVM, Decision Tree และ KNN

ในส่วนของงานวิจัยนี้เป็นการนำเสนอผลการประสิทธิภาพของการจำแนกข้อมูลด้วยการเรียนรู้ของเครื่อง (Machine Learning) จำนวน 3 ชนิด ได้แก่ Support Vector Machine (SVM) Decision Tree (DT) และ K-Nearest Neighbors (KNN) กับชุดคุณลักษณะที่ถูกลดทอนด้วยวิธี Kruskal-Wallis test และวิธี Mutual Information รวมชุดคุณลักษณะทั้งหมด 7 ชุดที่ไม่ซ้ำกันร่วมกับการปรับจูน Hyperparameter 5 ค่าต่อการเรียนรู้ของเครื่อง อีกทั้งยังเพิ่มการรันแบบกำหนดค่าเริ่มต้นแบบสุ่ม (Random Seed) จำนวน 5 รอบ

ผลลัพธ์ที่รายงานเป็นค่าเฉลี่ย (Mean) ของชุดพารามิเตอร์ในแต่ละค่าเริ่มต้นแบบสุ่มของแต่ละชุดข้อมูลคุณลักษณะเพื่อให้มั่นใจในความเสถียรและความน่าเชื่อถือ โดยตัวชี้วัดหลักที่ใช้ในการประเมินประสิทธิภาพคือ Accuracy F1-Score และ Area Under the ROC Curve (AUC)

ผลการประเมินประสิทธิภาพของแบบจำลอง (Model Performance Evaluation Results) ในภาพรวมผลการทดลองเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 3 ชนิด (SVM, DT และ KNN) พบว่า Support Vector Machine (SVM) เป็นแบบจำลองที่เหมาะสมที่สุดสำหรับชุดข้อมูลนี้ โดยให้ค่าความแม่นยำสูงสุดที่ 92.14% รองลงมาคือ Decision Tree (87.52%) และ K-Nearest Neighbors (79.16%) ตามลำดับ

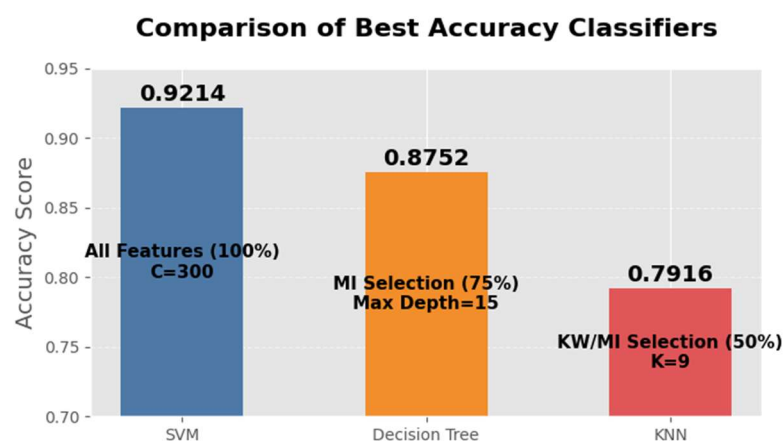


Figure 6. Comparison of Best Accuracy Classifiers

จาก Figure 6. แสดงให้เห็นความแตกต่างของประสิทธิภาพสูงสุด (Top Performance) ของแต่ละแบบจำลองอย่างชัดเจน ซึ่งสะท้อนให้เห็นว่าชุดข้อมูลนี้มีลักษณะการกระจายตัวที่เอื้อต่อการจำแนกด้วยการสร้างระนาบแบ่งกัน (Hyperplane) ของ SVM มากกว่าการพิจารณาความใกล้ชิดของข้อมูลแบบ KNN

7. สรุปผลการวิจัย (Conclusion)

การศึกษานี้ประสบความสำเร็จในการวิเคราะห์และเปรียบเทียบประสิทธิภาพของการลดขนาดคุณลักษณะเชิงสถิติต่อการจำแนกภาวะทางการแพทย์ โดยสามารถสรุปผลตามวัตถุประสงค์การวิจัยได้ดังนี้

1. ผลกระทบของการลดขนาดคุณลักษณะ การลดจำนวนคุณลักษณะด้วยวิธี Kruskal-Wallis (KW) ที่ระดับ 75% (12 คุณลักษณะ) ได้รับการพิสูจน์แล้วว่าเป็นจุดสมดุลที่ดีที่สุด โดยสามารถรักษาประสิทธิภาพความแม่นยำของแบบจำลอง SVM ได้สูงถึง 0.9165 ซึ่งลดลงจากชุดข้อมูลเต็มเพียง 0.49% แสดงให้เห็นว่าการกำจัดตัวแปรที่น้อยสำคัญต่ำช่วยลดความซับซ้อนของข้อมูลได้จริงโดยไม่ส่งผลกระทบต่อความถูกต้องในการวินิจฉัย

2. การเปรียบเทียบชุดพารามิเตอร์ การปรับจูนไฮเปอร์พารามิเตอร์อย่างเป็นระบบพบว่า ชุดค่าที่เหมาะสมที่สุดคือ การใช้ค่า C ที่ 300 ร่วมกับ RBF Kernel สำหรับ SVM และค่า Max Depth ที่ 15 สำหรับ Decision Tree โดยแบบจำลอง Decision Tree ตอบสนองต่อการคัดเลือกคุณลักษณะด้วยวิธี Mutual Information (MI) ได้ดีกว่าวิธี KW เล็กน้อย

3. การเปรียบเทียบประสิทธิภาพอัลกอริทึม แบบจำลอง SVM มีประสิทธิภาพเหนือกว่าอย่างมีนัยสำคัญในทุกมิติการประเมิน (Accuracy 0.9214, AUC 0.9935) ตามด้วย Decision Tree และ KNN ตามลำดับ ซึ่งบ่งชี้ว่า SVM เป็นเครื่องมือที่มีความทนทานและเหมาะสมที่สุดสำหรับโครงสร้างข้อมูลสุขภาพชุดนี้ที่มีความซับซ้อน

8. อภิปรายผลการวิจัย (Discussion)

8.1 ความสอดคล้องกับสมมติฐานและบริบทของข้อมูล

ผลการศึกษาเป็นไปตามสมมติฐานที่ตั้งไว้ว่า การลดมิติข้อมูลด้วยวิธีทางสถิติ (Statistical Feature Reduction) สามารถคงประสิทธิภาพการจำแนกได้ โดยในการทดลองชี้ให้เห็นว่าคุณลักษณะที่ถูกคัดออกในช่วง 25% สุดท้าย (Rank ต่ำสุด) เป็นสัญญาณรบกวนหรือข้อมูลซ้ำซ้อนที่ไม่จำเป็นต่อการเรียนรู้ของแบบจำลองมากนักเมื่อเทียบกับชุดข้อมูลที่มีคุณลักษณะครบ ซึ่งให้ค่าความแม่นยำของการจำแนกด้วยแบบจำลอง SVM มีความแตกต่างกันไม่เกิน 0.5% เท่านั้นเมื่อเทียบกับการเรียนรู้ข้อมูลที่เพิ่มขึ้น 25% นอกจากนี้ ผลการทดลองยังสอดคล้องกับทฤษฎีที่ว่า SVM มีความสามารถในการจัดการกับข้อมูลที่มีมิติสูงและซับซ้อนได้ดีกว่าอัลกอริทึมพื้นฐานอย่าง KNN และ Decision Tree โดยเฉพาะในบริบทของข้อมูลทางการแพทย์ที่มีความแปรปรวนสูง

8.2 การเปรียบเทียบกับงานวิจัยที่เกี่ยวข้อง

เมื่อเปรียบเทียบกับงานวิจัยในอดีต ผลลัพธ์ของ SVM ที่ได้ค่าความแม่นยำสูงสุด (92.14%) มีความสอดคล้องกับงานวิจัยของ Bashir et al. (2022) และ Jothi et al. (2022) ที่พบว่า SVM เป็นอัลกอริทึมที่ให้ผลลัพธ์โดดเด่นในการจำแนกโรค (92-94%) เมื่อเทียบกับอัลกอริทึมอื่น เนื่องจากการสร้าง Hyperplane ในมิติสูงช่วยแยกคลาสข้อมูลได้ดีกว่า

อย่างไรก็ตาม ผลการวิจัยนี้มีความแตกต่างจากงานของ Majed et al. (2023) ที่พบว่า KNN ให้ผลลัพธ์ดีที่สุด (94.11%) ในชุดข้อมูลโรคพาร์กินสัน สาเหตุของความขัดแย้งนี้คาดว่าเกิดจากลักษณะของชุดข้อมูล Healthcare Risk Factors ที่มีความไม่สมดุลของคลาส (Class Imbalance) สูง (ดังแสดงใน Table 2.) ซึ่ง KNN มีข้อจำกัดในการคำนวณระยะทางในกลุ่มข้อมูลข้างน้อย (Minority Class) ส่งผลให้ประสิทธิภาพตกลงเหลือเพียง 79.16% ในงานวิจัยนี้ นอกจากนี้ การที่งานวิจัยปัจจุบันพบว่า Decision Tree ตอบสนองต่อ Mutual Information ได้ดีกว่า KW ยังสอดคล้องกับแนวคิดของ Salehin et al. (2024) ที่สนับสนุนการใช้วิธีวัดความสัมพันธ์แบบไม่เชิงเส้น (Non-Linear) เพื่อเพิ่มประสิทธิภาพในโมเดลที่มีความซับซ้อน

9. ข้อเสนอแนะงานวิจัย (Recommendation)

1. การจัดการกับความไม่สมดุลของข้อมูล ควรพิจารณาประยุกต์ใช้เทคนิคขั้นสูงเพื่อแก้ปัญหาความไม่สมดุลของข้อมูล (Class Imbalance Problem) เพื่อเพิ่มประสิทธิภาพในการจำแนกกลุ่มข้อมูลที่มีจำนวนน้อย (Minority Class) ให้ดียิ่งขึ้น โดยเลือกใช้เทคนิคต่าง ๆ เช่น การสุ่มตัวอย่างส่วนน้อยเชิงสังเคราะห์แบบเพิ่มจำนวน (SMOTE) การสุ่มตัวอย่างเชิงสังเคราะห์แบบปรับตัว (ADASYN) หรือการเรียนรู้แบบคำนึงถึงต้นทุน (Cost-sensitive learning)

2. การขยายผลสู่บริบทการใช้งานจริง เนื่องจากชุดข้อมูลที่ใช้ในการศึกษานี้เป็นข้อมูลสาธารณะจากแหล่งข้อมูล Kaggle ซึ่งอาจมีบริบทที่แตกต่างจากลักษณะทางประชากรและระบบสาธารณสุขของประเทศไทย ดังนั้น เพื่อให้การประเมินประสิทธิภาพมีความสมบูรณ์และสะท้อนความสามารถในการใช้งานจริง (Practicality) เช่น นำแบบจำลองไปใช้กับระบบคัดกรองสุขภาพหรือนำแบบจำลองที่พัฒนาขึ้นไปทดสอบกับข้อมูลเวชระเบียนของผู้ป่วยจริงในโรงพยาบาลของประเทศไทยต่อไป

10. การเปิดเผยการใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์ (Declaration of Generative AI)

การใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์ งานวิจัยนี้ได้ประยุกต์ใช้ Gemini ในฐานะเครื่องมือช่วยร่างโครงสร้างโปรแกรมพื้นฐาน (Code Scaffolding) เพื่อเพิ่มประสิทธิภาพในการทำงาน อย่างไรก็ตาม งานวิจัยนี้ได้ดำเนินการตรวจสอบและทวนสอบ (Review and Verification) ตรรกะของคำสั่งทั้งหมดด้วยตนเอง โดยครอบคลุมตั้งแต่การนำเข้าข้อมูลเข้า การตรวจสอบกระบวนการสลับสุ่มข้อมูล (Manual Inspection of Data Shuffling) การควบคุมค่า Seed เพื่อให้สามารถทำซ้ำการทดลองได้ (Reproducibility) รวมถึงการตรวจสอบความถูกต้องของผลลัพธ์จากการวิเคราะห์ เพื่อให้มั่นใจว่าคำสั่งที่ใช้ในการทดลองมีความเที่ยงตรงตามหลักการวิจัย

11. การแสดงบทบาทผู้เขียน (CRediT Author Statement)

Yotsapat Ruangpaisarn: Conceptualization, Methodology, Software, Validation, Writing-Original Draft.

Sakulchai Saramat: Data Curation, Visualization.

Auttanit Wongjak: Software, Validation, Formal Analysis.

Sukanya Chansamorn: Investigation, Writing-Review & Editing, Supervision.

All authors read and approved the final manuscript.

12. เอกสารอ้างอิง (References)

- Ahmed, A. (2025). *Healthcare Risk Factors Dataset* [Dataset]. Kaggle. <https://www.kaggle.com/datasets/abdallaahmed77/healthcare-risk-factors-dataset>
- Bashir, S., Khattak, I. U., Khan, A., Khan, F. H., Gani, A., & Shiraz, M. (2022). A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches. *Complexity*, 2022(1), 8190814. <https://doi.org/10.1155/2022/8190814>
- Clarke, R., Ransom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., & Wang, Y. (2008). The Properties of High-Dimensional Data Spaces: Implications for Exploring Gene and Protein Expression Data. *Nature Reviews Cancer*, 8(1), 37-49. <https://doi.org/10.1038/nrc2294>
- Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157-1182.
- Jothi, N., Syed-Mohamed, S. M., & Rajagopal, H. (2022, July 20-22). Hybrid Feature Selection using Shapley Value and ReliefF for Medical Datasets. *2022 International Conference on Inventive Computation Technologies*, 351-355. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/icict54344.2022.9850833>

- Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, 104-116. <https://doi.org/10.1016/j.csbj.2016.12.005>
- Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260), 583-621. <https://doi.org/10.1080/01621459.1952.10483441>
- Majed, R. J., Al-Heddi, R. M., & Zeki, A. M. (2023, October 25-26). Parkinson's Disease Detection Using Data Mining Models: A Comparative Study. *2023 4th International Conference on Data Analytics for Business and Industry*, 290-294. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/icdabi60145.2023.10629302>
- Narwane, S. V., & Sawarkar, S. D. (2021, August 27-29). Dimensionality Reduction of Unbalanced Datasets: Principal Component Analysis. *2021 Asian Conference on Innovation in Technology*, 1-6. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/asiancon51346.2021.9544971>
- Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-Fold Cross-Validation on ML Classifiers for Predicting Cervical Cancer. *Frontiers in Nanotechnology*, 4, 972421. <https://doi.org/10.3389/fnano.2022.972421>
- Remeseiro, B., & Bolon-Canedo, V. (2019). A Review of Feature Selection Methods in Medical Applications. *Computers in Biology and Medicine*, 112, 103375. <https://doi.org/10.1016/j.compbimed.2019.103375>
- Saha, P., Patikar, S., & Neogy, S. (2020, October 2-4). A Correlation-Sequential Forward Selection Based Feature Selection Method for Healthcare Data Analysis. *2020 IEEE International Conference on Computing, Power and Communication Technologies*, 69-72. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/gucon48875.2020.9231205>
- Salehin, S., Islam, A. J., Iqbal, A. M., Barua, L., Islam, K., & Uddin, A. (2024, December 7-8). A Comparative Analysis of Feature Selection Methods on the Accuracy of Heart Disease Prediction Models. *2024 International Conference on Recent Progresses in Science, Engineering and Technology*, 1-6. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/icrpset64863.2024.10955882>
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379-423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Siburian, R. H., Km Nasution, M., & Tarigan, J. T. (2025, January 21). Optimization Of KNN, SVM, And SVM Kernel in Water Potability Prediction with Hyperparameter Approach. *2025 International Conference on Computer Sciences, Engineering, and Technology Innovation*, 576-581. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/icocseti63724.2025.11019055>
- Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. *BMC Medical Informatics and Decision Making*, 19(1), 281. <https://doi.org/10.1186/s12911-019-1004-8>