ขั้นตอนวิธีการเลือกคุณลักษณะที่เหมาะสมโดยใช้ความฉลาดแบบกลุ่มในชุดข้อมูลมะเร็งเต้านม

OPTIMIZED FEATURE SELECTION ALGORITHMS USING SWARM INTELLIGENCE IN BREAST CANCER DATASET

ชลันวิชญ์ ธีระสานต์

CHALANWICH TEERASARN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2568

KMITL-2025-SC-M-002-027

OPTIMIZED FEATURE SELECTION ALGORITHMS USING SWARM INTELLIGENCE IN BREAST CANCER DATASET

CHALANWICH TEERASARN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

DEPARTMENT OF COMPUTER SCIENCE SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2025

หัวข้อวิทยานิพนธ์	ขั้นตอนวิธีการเลือกคุณลักษณะที่เหมาะสมโดยใช้ความฉลาดแบบกลุ่มในชุดข้อมูลมะเร็งเต้านม
ชื่อนักศึกษา	นายชลันวิชญ์ ธีระสานต์
รหัสประจำตัว	63605108
ปริญญา	วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2568
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร.วรางคณา กิ้มปาน

บทคัดย่อ

การเลือกคุณลักษณะที่เหมาะสมเป็นขั้นตอนสำคัญอย่างหนึ่งในงานด้านการเรียนรู้ของเครื่อง (Machine Learning) โดยมีวัตถุประสงค์หลักเพื่อปรับปรุงประสิทธิภาพของแบบจำลองการจำแนกประเภท โดยเลือกชุดของคุณลักษณะที่สำคัญที่สุดจากข้อมูลจำนวนมาก ส่งผลให้แบบจำลองสามารถทำงานได้รวดเร็วขึ้น ลดความซับซ้อน และมีความแม่นยำสูงขึ้น ซึ่งในปัจจุบันข้อมูลทางการแพทย์มักมีลักษณะหลายมิติและซับซ้อน จึงต้องอาศัยผู้เชี่ยวชาญเฉพาะทางในการวินิจฉัยข้อมูล กระบวนการนี้จึงใช้เวลาในการประมวลผลและตีความ ซึ่งส่งผลต่อการตัดสินใจทางการแพทย์ ดังนั้นงานวิจัยนี้เสนอขั้นตอนวิธีการเลือกคุณลักษณะที่เหมาะสมที่สุดในเชิงความแม่นยำและความสามารถในการลดมิติข้อมูล โดยใช้เทคนิคความฉลาดแบบกลุ่มในชุดข้อมูลมะเร็งเต้านม ผลการทดลองแสดงให้เห็นว่าการใช้อัลกอริทึมแบบผสมผสานกันระหว่าง Cuckoo Search (CS), Firefly Algorithm (FA) และ Particle Swarm Optimization (PSO) มาประยุกต์ใช้กับในการคัดเลือกคุณลักษณะ และวิธีที่นำเสนอดังกล่าวสามารถลดจำนวนคุณลักษณะลงได้อย่างมีนัยสำคัญโดยอัลกอริทึมแบบผสมผสาน PSOCS ให้ค่าความแม่นยำที่ 98.83% และจำนวนคุณลักษณะที่ถูกเลือกเพียง 14 คุณลักษณะ อัลกอริทึมแบบผสมผสาน PSOFA ให้ค่าความแม่นยำที่ 98.83% โดยเลือกคุณลักษณะที่ 15 คุณลักษณะ และอัลกอริทึมแบบผสมผสาน CSFA ให้ค่าความแม่นยำที่ 97.66% โดยเลือกคุณลักษณะ 13 คุณลักษณะ จากผลลัพธ์แสดงให้เห็นว่าการนำจุดแข็งของแต่ละอัลกอริทึมมาผสมผสานกันเพื่อใช้ในการเลือกคุณลักษณะได้อย่างประสิทธิภาพสามารถรักษาความแม่นยำของแบบจำลองและลดมิติของข้อมูลได้

คำสำคัญ : การเลือกคุณลักษณะที่เหมาะสมโดยอัลกอริทึมเมตาฮิวริสติก ข้อมูลทางการแพทย์

Thesis Title	Optimized Feature Selection Algorithms using Swarm intelligence in Breast Cancer Dataset
Student Name	Mr. Chalanwich Teerasarn
Student ID	63605108
Degree	Master of Science (Computer Science)
Department	Computer Science
Year	2025
Thesis Advisor	Asst.Prof.Dr.Warangkhana Kimpan

Abstract

Feature selection is a crucial step in machine learning, with the primary objective of enhancing the performance of classification models by selecting the most important subset of features from large datasets. This helps improve model efficiency, reduce complexity, and increase accuracy. In the medical domain, data is often high-dimensional and complex, requiring expert interpretation, which makes the diagnostic process time-consuming and impacts medical decision-making. Therefore, this research proposes an effective feature selection approach based on both classification accuracy and dimensionality reduction using swarm intelligence techniques applied to the breast cancer dataset. Experimental results demonstrate that hybrid algorithms combining Cuckoo Search (CS), Firefly Algorithm (FA), and Particle Swarm Optimization (PSO) can significantly reduce the number of selected features. The hybrid PSOCS algorithm achieved 98.83% accuracy with only 14 selected features, the PSOFA algorithm also achieved 98.83% accuracy with 15 selected features, and the CSFA algorithm attained 97.66% accuracy with 13 selected features. These results indicate that combining the strengths of different algorithms can effectively preserve model accuracy while reducing data dimensionality.

Keywords : Feature Selection, Metaheuristics, Medical Dataset

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีเพราะได้รับความกรุณาชี้แนะ และช่วยเหลือจาก ผู้ช่วยศาสตราจารย์ ดร.วรางคณา กิ้มปาน อาจารย์ที่ปรึกษาที่ได้ให้คำแนะนำ แนวคิด และบ่งชี้ข้อผิดพลาดต่างๆ มาโดยตลอด จนวิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ จึงขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณคณะกรรมการตรวจวิทยานิพนธ์ทุกท่านได้แก่ ประธานกรรมการ และกรรมการสอบวิทยานิพนธ์ที่กรุณาตรวจสอบแก้ไขและให้คำแนะนำ จนทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์ยิ่งขึ้น

สุดท้ายนี้ข้าพเจ้าขออุทิศผลงานนี้แด่บิดาและมารดาผู้ล่วงลับ ผู้เป็นแรงผลักดันทางจิตใจในชีวิตข้าพเจ้าเสมอมา คุณยายผู้อุทิศเวลาทั้งชีวิตเพื่อเลี้ยงดู สั่งสอนข้าพเจ้าจนเติบโตเป็นผู้ใหญ่ และข้าพเจ้าขอขอบพระคุณครอบครัวที่คอยช่วยเหลือสนับสนุนข้าพเจ้าเสมอมาตลอดจนประสบความสำเร็จ

นาย ชลันวิชญ์ ธีระสานต์

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการวินิจฉัยโรคทางการแพทย์มีความสำคัญอย่างยิ่งต่อการรักษาอย่างทันท่วงทีและแม่นยำ อย่างไรก็ตามกระบวนการวินิจฉัยดังกล่าวต้องอาศัยความเชี่ยวชาญของแพทย์เฉพาะทางที่มีประสบการณ์ และการตีความข้อมูลที่มีความซับซ้อน เช่น ข้อมูลภาพถ่ายทางการแพทย์ ผลตรวจทางห้องปฏิบัติการ และประวัติผู้ป่วย ซึ่งมีแนวโน้มจะเพิ่มปริมาณและความซับซ้อนมากขึ้นในยุคของข้อมูลขนาดใหญ่ (Big Data) ส่งผลให้การวินิจฉัยมีความล่าช้า และอาจเกิดข้อผิดพลาดได้

ด้วยเหตุนี้ การนำเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence: AI) และการเรียนรู้ของเครื่อง (Machine Learning) เข้ามาช่วยในการประมวลผลข้อมูลทางการแพทย์จึงได้รับความสนใจเป็นอย่างมาก โดยเฉพาะการพัฒนาแบบจำลองการจำแนกประเภท (Classification Models) ซึ่งสามารถนำไปประยุกต์ใช้เป็นระบบช่วยสนับสนุนการตัดสินใจทางคลินิก (Clinical Decision Support Systems: CDSS) ที่มีความรวดเร็วและแม่นยำ

อย่างไรก็ตามข้อมูลทางการแพทย์มักมีจำนวนคุณลักษณะ (Features) เป็นจำนวนมาก ซึ่งบางคุณลักษณะอาจไม่เกี่ยวข้องหรือมีนัยสำคัญต่ำ การใช้ข้อมูลทั้งหมดโดยไม่คัดกรองอาจนำไปสู่ปัญหา Overfitting ทำให้แบบจำลองขาดประสิทธิภาพ อีกทั้งยังเพิ่มภาระในการคำนวณ ดังนั้น กระบวนการเลือกคุณลักษณะที่เหมาะสม (Feature Selection) จึงเป็นขั้นตอนสำคัญในการลดมิติของข้อมูลคงไว้ซึ่งเฉพาะคุณลักษณะที่จำเป็น ส่งผลให้แบบจำลองมีความกระชับและแม่นยำยิ่งขึ้น

อัลกอริทึมเชิงเมตาฮิวริสติก (Metaheuristic Algorithms) เช่น Particle Swarm Optimization (PSO), Cuckoo Search (CS) และ Firefly Algorithm (FA) ได้รับการยอมรับว่าเป็นแนวทางที่มีประสิทธิภาพในการแก้ปัญหาที่ค้นหาค่าที่เหมาะสม (Optimization) ซึ่งสามารถนำมาประยุกต์ใช้กับปัญหาการเลือกคุณลักษณะได้อย่างมีประสิทธิผล โดยเฉพาะการพัฒนาอัลกอริทึมแบบผสมผสาน (Hybrid Metaheuristics) เพื่อเพิ่มศักยภาพในการค้นหาคำตอบที่ดีที่สุดจากพื้นที่ค้นหาขนาดใหญ่ ซึ่งมีแนวโน้มช่วยปรับปรุงความแม่นยำของระบบวินิจฉัย และลดจำนวนคุณลักษณะที่ใช้ในการเรียนรู้ได้อย่างมีนัยสำคัญ

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อพัฒนาแบบจำลองการเลือกคุณลักษณะที่เหมาะสมจากข้อมูลทางการแพทย์ โดยใช้เทคนิคอัลกอริทึมเมตาฮิวริสติกแบบผสมผสานระหว่าง Cuckoo Search, Firefly Algorithm และ Particle Swarm Optimization (PSO)

2) เพื่อจำแนกกลุ่มข้อมูลทางการแพทย์โดยใช้คุณลักษณะที่เหมาะสม และประเมินประสิทธิภาพของแบบจำลองการจำแนกประเภทที่ได้
3) เพื่อเปรียบเทียบประสิทธิภาพของวิธีที่เสนอ กับวิธีการเลือกคุณลักษณะแบบดั้งเดิมในแง่ของความแม่นยำในการจำแนกประเภท และจำนวนคุณลักษณะที่ใช้
4) เพื่อศึกษาความเป็นไปได้ ในการนำแบบจำลองที่ได้ไปประยุกต์ใช้เป็นเครื่องมือสนับสนุนการตัดสินใจทางการแพทย์

1.3 ขอบเขตของงานวิจัย

1) งานวิจัยนี้มุ่งเน้นการประยุกต์ใช้เทคนิคการเลือกคุณลักษณะ (Feature Selection) เพื่อคัดเลือกคุณลักษณะที่สำคัญจากชุดข้อมูลทางการแพทย์
2) อัลกอริทึมเมตาฮิวริสติกที่ใช้ในการเลือกคุณลักษณะในงานวิจัยนี้ ได้แก่ อัลกอริทึมแบบผสมผสานระหว่าง Cuckoo Search (CS), Firefly Algorithm (FA) และ Particle Swarm Optimization (PSO) โดยพัฒนาขึ้นเพื่อเพิ่มประสิทธิภาพในการค้นหาคำตอบที่เหมาะสมที่สุด
3) แบบจำลองการจำแนกประเภทที่ใช้ในการประเมินประสิทธิภาพของชุดคุณลักษณะที่เลือก จะใช้แบบจำลองพื้นฐาน Support Vector Machine (SVM) เพื่อเปรียบเทียบผลลัพธ์ความแม่นยำ
4) ชุดข้อมูลที่ใช้ในการทดลอง ได้แก่ ชุดข้อมูลทางการแพทย์ Breast Cancer Wisconsin (Diagnostic) ชุดข้อมูล SPECTF Heart และชุดข้อมูล Arrhythmia จาก UCI Machine Learning Repository ซึ่งเป็นชุดข้อมูลที่ได้รับความนิยมและมีการใช้อย่างแพร่หลายในงานวิจัยด้านการเรียนรู้ของเครื่อง
5) งานวิจัยนี้จำกัดขอบเขตการประเมินประสิทธิภาพในเชิงของ ความแม่นยำในการจำแนกประเภท และจำนวนคุณลักษณะที่เลือกได้
6) ไม่ครอบคลุมการนำแบบจำลองไปใช้งานจริงในทางคลินิก แต่เน้นการวิเคราะห์เชิงทดลองจากข้อมูลทุติยภูมิ (Secondary Data)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ได้แนวทางการเลือกคุณลักษณะที่มีประสิทธิภาพ โดยสามารถลดจำนวนฟีเจอร์ที่ไม่จำเป็นลงได้ โดยไม่ลดทอนความแม่นยำของแบบจำลอง
2) เพิ่มประสิทธิภาพของแบบจำลองการจำแนกประเภททางการแพทย์ให้สามารถวินิจฉัยโรคได้รวดเร็วและแม่นยำยิ่งขึ้น
3) ลดเวลาและทรัพยากรที่ใช้ในการประมวลผลข้อมูลขนาดใหญ่ โดยเฉพาะในขั้นตอนการฝึกแบบจำลอง

4) เป็นพื้นฐานสำหรับการวิจัยและพัฒนาระบบสนับสนุนการตัดสินใจทางการแพทย์ในอนาคต โดยเฉพาะในด้านการประยุกต์ใช้ AI กับข้อมูลผู้ป่วย

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในหัวข้อนี้กล่าวถึงความรู้เกี่ยวกับการเลือกคุณลักษณะ ตลอดจนปัญหาเกี่ยวกับการวินิจฉัยข้อมูลทางการแพทย์ และทฤษฎีที่เป็นแนวคิดสำหรับการแก้ปัญหาโดยความฉลาดแบบกลุ่มในการเลือกคุณลักษณะที่เหมาะสมที่สุด

2.1 ความรู้เบื้องต้นเกี่ยวกับการเลือกคุณลักษณะ

การเลือกคุณลักษณะ (Feature Selection) เป็นขั้นตอนสำคัญในกระบวนการเตรียมข้อมูลเพื่อการเรียนรู้ของเครื่อง (Machine Learning) และการจำแนกประเภท โดยเฉพาะอย่างยิ่งเมื่อต้องจัดการกับข้อมูลที่มีจำนวนมิติมาก (High-dimensional Data) เช่น ข้อมูลทางการแพทย์ ซึ่งมักประกอบด้วยตัวแปรจำนวนมากที่อาจไม่จำเป็นต่อการทำนายผลลัพธ์ การคัดเลือกเฉพาะคุณลักษณะที่สำคัญจึงมีบทบาทในการลดความซับซ้อนของโมเดล ลดเวลาในการประมวลผล ป้องกันการเกิดปัญหาแบบจำลองจำเพาะกับข้อมูลฝึกมากเกินไป (Overfitting) และเพิ่มความสามารถของแบบจำลองในการทำนายข้อมูลใหม่ได้อย่างมีประสิทธิภาพ [1]

ในบริบทของการแพทย์ การเลือกคุณลักษณะที่เหมาะสมสามารถช่วยให้การวิเคราะห์ข้อมูลผู้ป่วยมีความแม่นยำมากขึ้น และสนับสนุนการตัดสินใจทางคลินิกได้อย่างมีประสิทธิภาพ เช่น การเลือกตัวชี้วัดทางชีวภาพที่สำคัญต่อการวินิจฉัยโรค หรือการคัดกรองกลุ่มเสี่ยง กระบวนการเลือกคุณลักษณะที่มีประสิทธิภาพยังช่วยให้เข้าใจความสัมพันธ์ระหว่างตัวแปรกับผลลัพธ์ทางคลินิกได้ชัดเจนขึ้น ซึ่งส่งผลต่อการตีความผลลัพธ์ของโมเดลอย่างมีนัยสำคัญ โดยเฉพาะในสาขาเช่น ชีวการแพทย์ หรือวิทยาศาสตร์สุขภาพที่การตีความมีความสำคัญไม่น้อยไปกว่าความแม่นยำของโมเดลเอง [2]

2.1.1 ประเภทของวิธีการเลือกคุณลักษณะ

การเลือกคุณลักษณะสามารถแบ่งออกได้เป็น 3 แนวทางหลัก ได้แก่ วิธีการกรองคุณลักษณะ (Filter Methods) วิธีห่อหุ้มด้วยแบบจำลอง (Wrapper Methods) และวิธีฝังในกระบวนการฝึกแบบจำลอง (Embedded Methods) ซึ่งแต่ละแนวทางมีหลักการ ประสิทธิภาพ และข้อจำกัดที่แตกต่างกัน เหมาะสมกับลักษณะของข้อมูล และวัตถุประสงค์ของแบบจำลองที่หลากหลาย ดังนี้

1. **วิธีการกรองคุณลักษณะ (Filter Methods) -** เป็นแนวทางที่แยกกระบวนการเลือกคุณลักษณะออกจากกระบวนการฝึกแบบจำลอง โดยอาศัยสถิติเพื่อประเมินความสัมพันธ์ระหว่างแต่ละคุณลักษณะกับตัวแปรเป้าหมาย เช่น การใช้ค่า Pearson Correlation, Information Gain หรือ Chi-square วิธีนี้มีข้อดีคือความเร็วในการ

คำนวณ และสามารถใช้ได้กับแบบจำลองหลากหลายชนิด แต่มีข้อจำกัดที่ไม่สามารถพิจารณาปฏิสัมพันธ์ระหว่างคุณลักษณะร่วมกันได้ [3]

2. **วิธีห่อหุ้มด้วยแบบจำลอง (Wrapper Methods)** – เป็นวิธีการสร้างแบบจำลองขึ้นมาเพื่อประเมินชุดของคุณลักษณะในแต่ละรอบ เช่น การใช้ Recursive Feature Elimination (RFE) ซึ่งจะเลือกคุณลักษณะที่ทำให้โมเดลมีประสิทธิภาพดีที่สุด วิธีนี้สามารถพิจารณาความสัมพันธ์ร่วมระหว่างคุณลักษณะได้ดี แต่มีจุดอ่อนคือใช้เวลาคำนวณนานและไม่เหมาะกับชุดข้อมูลที่มีมิติมาก
3. **วิธีฝังในกระบวนการฝึกแบบจำลอง (Embedded Methods)** – ผสานการเลือกคุณลักษณะเข้ากับกระบวนการฝึกแบบจำลองโดยตรง เช่น การใช้ Lasso Regression ที่เพิ่มเงื่อนไขการปรับค่าความซับซ้อนของแบบจำลอง (Regularization) เข้าไปในฟังก์ชันค่าความสูญเสีย (Loss Function) ทำให้สามารถตัดคุณลักษณะที่ไม่จำเป็นออกได้โดยอัตโนมัติ วิธีนี้ให้ความสมดุลระหว่างความแม่นยำและความเร็วในการเลือกคุณลักษณะ และมักเป็นที่นิยมในแบบจำลองเชิงเส้นและต้นไม้ตัดสินใจ [4]

2.1.2 ความสำคัญของการเลือกคุณลักษณะในการประมวลผลข้อมูลขนาดใหญ่

ในยุคของข้อมูลขนาดใหญ่ (Big Data) การวิเคราะห์ข้อมูลที่มีจำนวนมิติมากเป็นเรื่องท้าทาย เนื่องจากข้อมูลจำนวนมากอาจมีคุณลักษณะที่ไม่จำเป็นหรือมีความซ้ำซ้อนสูง ซึ่งไม่เพียงแต่ทำให้การประมวลผลช้าลง แต่ยังเพิ่มความเสี่ยงต่อการเกิดแบบจำลองจำเพาะกับข้อมูลฝึกมากเกินไป (Overfitting) อีกด้วย การเลือกเฉพาะคุณลักษณะที่สำคัญจึงมีบทบาทในการปรับปรุงประสิทธิภาพของโมเดลทั้งในแง่ของเวลา ความซับซ้อน และความแม่นยำ โดยเฉพาะอย่างยิ่งในกรณีที่โมเดลต้องจัดการกับปัญหาข้อมูลไม่สมบูรณ์ เช่น ข้อมูลที่มีการขาดหาย (Missing Values) หรือความไม่สมดุลของคลาส (Class Imbalance) ซึ่งเป็นปัญหาพบได้ทั่วไปในงานจริง [5]

นอกจากนี้ ในงานด้านการแพทย์ซึ่งข้อมูลมีความละเอียดอ่อนและผลลัพธ์มีผลต่อชีวิต การเลือกคุณลักษณะที่เหมาะสมจึงเป็นสิ่งจำเป็นอย่างยิ่ง เพราะนอกจากจะช่วยเพิ่มความแม่นยำในการวินิจฉัยแล้ว ยังมีส่วนในการลดอคติของแบบจำลอง และทำให้ระบบช่วยตัดสินใจทางคลินิก (CDSS) มีความน่าเชื่อถือมากยิ่งขึ้น การคัดเลือกคุณลักษณะอย่างเป็นระบบจึงถือเป็นกลยุทธ์สำคัญในการพัฒนาแบบจำลองที่ปลอดภัย มีประสิทธิภาพ และสามารถใช้งานได้จริงในบริบทของระบบสาธารณสุข

2.1.3 ปัญหาและความท้าทายของการเลือกคุณลักษณะ

แม้ว่าการเลือกคุณลักษณะจะมีบทบาทสำคัญต่อประสิทธิภาพของแบบจำลอง แต่กระบวนการนี้ก็เผชิญกับปัญหาและความท้าทายหลายประการ หนึ่งในปัญหาที่พบบ่อย ได้แก่ การติดกับดักในค่าที่เหมาะสมเฉพาะจุด (Local Optima) โดยเฉพาะเมื่อใช้อัลกอริทึมเชิงฮิวริสติกที่ไม่มีหลักประกันว่าจะสามารถค้นหาคำตอบที่ดีที่สุดในระดับ (Global) ได้ นอกจากนี้ การประเมิน

คุณลักษณะแต่ละตัวแบบแยกจากกัน อาจทำให้มองข้ามความสัมพันธ์ร่วมกันระหว่างคุณลักษณะ หรือที่เรียกว่า Feature Interaction ซึ่งส่งผลให้การเลือกชุดคุณลักษณะที่ได้ไม่สอดคล้องกับลักษณะของข้อมูลจริง

อีกหนึ่งความท้าทายสำคัญคือคุณภาพของข้อมูลต้นทาง ซึ่งอาจมีลักษณะไม่สมดุล (Class Imbalance), มีสัญญาณรบกวน (Noise) หรือมีค่าที่หายไป (Missing Values) อยู่เป็นจำนวนมาก หากไม่มีการจัดการข้อมูลเหล่านี้อย่างเหมาะสม จะส่งผลต่อการเลือกคุณลักษณะที่ไม่สะท้อนคุณลักษณะที่สำคัญจริงในเชิงบริบท นำไปสู่การสร้างโมเดลที่ไม่มีความเสถียรเมื่อนำไปใช้งานจริง [6],[7]

2.1.4 การเลือกคุณลักษณะกับข้อมูลทางการแพทย์

ข้อมูลทางการแพทย์มีลักษณะเฉพาะที่แตกต่างจากข้อมูลในสาขาอื่น โดยมักประกอบด้วยตัวแปรจำนวนมาก (High-dimensional) ขณะที่จำนวนตัวอย่างกลับมีน้อย (Small-sample Size) ซึ่งเพิ่มความเสี่ยงต่อการเกิดแบบจำลองจำเพาะกับข้อมูลฝึกมากเกินไป หรือการเกิดปัญหา (Overfitting) เมื่อฝึกโมเดลกับข้อมูลลักษณะนี้ การเลือกคุณลักษณะที่เหมาะสมจึงมีความสำคัญอย่างยิ่งในการลดความซับซ้อนของข้อมูล และเพิ่มประสิทธิภาพของแบบจำลองในการวินิจฉัยหรือพยากรณ์โรค [8]

นอกจากจะช่วยลดจำนวนตัวแปรและภาระในการประมวลผลแล้ว การเลือกคุณลักษณะอย่างรอบคอบยังสนับสนุนการตีความผลลัพธ์เชิงคลินิก เช่น การวิเคราะห์ความสัมพันธ์ระหว่างคุณลักษณะกับภาวะของผู้ป่วย หรือความเสี่ยงของโรค โดยเฉพาะเมื่อใช้การเลือกคุณลักษณะร่วมกับอัลกอริทึมเมตาฮิวริสติก เช่น Particle Swarm Optimization (PSO), Firefly Algorithm (FA) และ Cuckoo Search (CS) ที่มีศักยภาพในการจัดการกับข้อมูลที่มีสัญญาณรบกวน และโครงสร้างข้อมูลที่ซับซ้อนสูงในบริบทของการแพทย์ [9]

2.1.5 บทบาทของการเลือกคุณลักษณะต่อประสิทธิภาพของแบบจำลองจำแนกประเภท

การเลือกคุณลักษณะที่เหมาะสมมีผลโดยตรงต่อประสิทธิภาพของแบบจำลองการจำแนกประเภทในหลายด้าน โดยประการแรก ได้แก่ การลดจำนวนคุณลักษณะที่ไม่จำเป็น ซึ่งส่งผลให้ลดภาระในการประมวลผลของระบบ และทำให้สามารถตอบสนองแบบเรียลไทม์ได้ดีขึ้นในระบบที่ต้องการความเร็วในการคำนวณ เช่น ระบบช่วยวินิจฉัยโรคอัตโนมัติในคลินิกหรือโรงพยาบาลที่มีข้อจำกัดด้านทรัพยากรคอมพิวเตอร์

นอกจากนี้ การเลือกคุณลักษณะที่สัมพันธ์กันอย่างมีนัยสำคัญกับผลลัพธ์ยังช่วยเพิ่มความสามารถในการทำนายข้อมูลใหม่ของแบบจำลอง ทำให้สามารถคาดการณ์ผลลัพธ์จากข้อมูลใหม่ได้อย่างแม่นยำมากขึ้น โดยเฉพาะในกรณีที่ข้อมูลมีลักษณะไม่สมดุลหรือมีความซับซ้อนสูง ซึ่งมักพบในข้อมูลทางการแพทย์ อีกทั้งยังช่วยลดความซับซ้อนของแบบจำลอง ทำให้แบบจำลองง่ายต่อการวิเคราะห์และตีความในเชิงคลินิก ซึ่งเป็นปัจจัยสำคัญต่อการยอมรับและใช้งานในทางปฏิบัติ [10]

2.2 ความฉลาดแบบกลุ่ม (Swarm Intelligence)

Swarm Intelligence (SI) คือแนวคิดที่ได้รับแรงบันดาลใจจากพฤติกรรมร่วมของสิ่งมีชีวิตในธรรมชาติ เช่น ฝูงนก ฝูงปลา หรือกลุ่มแมลง ซึ่งสามารถทำงานร่วมกันอย่างมีประสิทธิภาพโดยไม่มีศูนย์กลางควบคุม แนวคิดนี้ถูกนำมาประยุกต์ใช้ในการพัฒนาอัลกอริทึมการเพิ่มประสิทธิภาพที่ใช้ตัวแทนหลายตัว (Agents) ร่วมกันค้นหาคำตอบของปัญหาที่ซับซ้อน โดยอาศัยหลักการของการสื่อสารในกลุ่ม และการปรับพฤติกรรมตามประสบการณ์ของตนเองและผู้อื่น [11]

ในบริบทของการเลือกคุณลักษณะความฉลาดแบบกลุ่ม แสดงศักยภาพในการค้นหาคำตอบในพื้นที่ที่มีมิติสูงและโครงสร้างซับซ้อน ซึ่งอัลกอริทึมแบบดั้งเดิมอาจไม่สามารถจัดการได้อย่างมีประสิทธิภาพ อัลกอริทึมประเภทนี้สามารถหลีกเลี่ยงการติดอยู่ในค่าที่เหมาะสมเฉพาะจุด (Local Optima) และยังสามารถปรับตัวให้เข้ากับลักษณะข้อมูลที่เปลี่ยนแปลงหรือมีความไม่แน่นอนได้ดี จึงได้รับความนิยมอย่างแพร่หลายในงานด้านการเลือกคุณลักษณะ โดยเฉพาะในข้อมูลทางการแพทย์ที่มักมีลักษณะมิติสูง และมีสัญญาณรบกวนสูง [12]

2.2.1 แนวคิดพื้นฐานของอัลกอริทึมความฉลาดแบบกลุ่ม

อัลกอริทึมที่อยู่ภายใต้แนวคิดความฉลาดแบบกลุ่มมีหลักการพื้นฐานร่วมกันคือ การใช้กลุ่มของตัวแทน หรือคำตอบในประชากรที่มีปฏิสัมพันธ์กันในลักษณะที่ไม่มีศูนย์กลางควบคุม (Decentralized Control) ตัวแทนแต่ละตัวจะมีพฤติกรรมที่เรียบง่าย แต่สามารถประสานงานกันเพื่อแก้ปัญหาที่ซับซ้อนได้ผ่านการแบ่งปันข้อมูล และการอัปเดตตำแหน่งของตนเองอย่างต่อเนื่อง

หนึ่งในกลไกสำคัญของอัลกอริทึมความฉลาดแบบกลุ่ม คือ การสร้างสมดุลระหว่างการสำรวจ (Exploration) และการใช้ประโยชน์จากประสบการณ์เดิม (Exploitation) เพื่อให้การค้นหาคำตอบไม่ติดอยู่ที่ค่าที่เหมาะสมเฉพาะจุด ตัวอย่างเช่น ใน Particle Swarm Optimization (PSO) ตัวแทนแต่ละตัวจะอัปเดตตำแหน่งตามประสบการณ์ของตนเอง (Personal Best) และประสบการณ์ของกลุ่ม (Global Best) ขณะที่ใน Firefly Algorithm (FA) การเคลื่อนที่ของตัวแทนจะขึ้นอยู่กับการดึงดูดตามความเข้มของแสง ซึ่งแสดงถึงคุณภาพของคำตอบ [13]

2.2.2 อัลกอริทึม Cuckoo Search (CS)

Cuckoo Search (CS) เป็นอัลกอริทึมเชิงเมตาฮิวริสติกที่ได้รับแรงบันดาลใจจากพฤติกรรมการวางไข่ของนกคัคคู ซึ่งจะวางไข่ในรังของนกชนิดอื่น โดยมีแนวคิดหลักในการแทนที่คำตอบที่คุณภาพต่ำด้วยคำตอบใหม่ที่มีคุณภาพดีกว่าในลักษณะเดียวกับที่ไข่ของนกคัคคูที่ถูกยอมรับจะอยู่รอดต่อไปในรังของนกเจ้าบ้าน

หนึ่งในกลไกสำคัญของอัลกอริทึมนี้ ได้แก่ การใช้การเคลื่อนที่แบบ Lévy Flight ซึ่งเป็นกลยุทธ์การสุ่ม ที่มีระยะการกระโดดหลากหลายระดับ โดยอิงจากการแจกแจงแบบ Lévy Distribution

ซึ่งช่วยให้การค้นหาสามารถออกจากค่าที่เหมาะสมเฉพาะจุดได้อย่างมีประสิทธิภาพ และครอบคลุมพื้นที่การค้นหาในวงกว้างมากกว่าการสุ่มแบบปกติ

สมการที่ใช้อัปเดตตำแหน่งคำตอบใหม่ใน CS แสดงดังสมการที่ (2.1)

$$X_i^{(t+1)} = X_i^{(t)} + a \cdot Levy\ (\lambda) \tag{2.1}$$

โดยที่

- $X_i^{(t)}$ คือ ตำแหน่งคำตอบปัจจุบันของตัวแทนที่ I ในรอบที่ t
- a คือ ค่าคงที่ที่ควบคุมขนาดของการกระโดด
- $Levy\ (\lambda)$ คือ เวกเตอร์การกระโดดที่สุ่มตามการแจกแจงแบบ Lévy ซึ่ง λ มักอยู่ในช่วง $1 < \lambda \leq 3$

Algorithm 1 Cuckoo Search via Lévy Flights

```
begin
Objective function f(x), x = (x_1, ..., x_d)^T
Generate initial population of n host nests x_i (i = 1, 2, ..., n)
while t < MaxGeneration or (stop criterion) do
    Get a cuckoo randomly by Lévy flights
    Evaluate its quality/fitness F_i
    Choose a nest among n (say, j) randomly
    if F_i > F_j then
        Replace j by the new solution
    end if
    A fraction (p_a) of worse nests are abandoned and new ones are built
    Keep the best solutions (or nests with quality solutions)
    Rank the solutions and find the current best
end while
Postprocess results and visualization
end
```

รูปที่ 2.1 อัลกอริทึมของ CS

จากสมการข้างต้นเป็นส่วนหนึ่งของการอัปเดตตำแหน่งคำตอบใหม่ในขั้นตอนวิธี Cuckoo Seach แสดงดังรูปที่ 2.1 โดยสามารถอธิบายรายละเอียดในแต่ละขั้นตอนได้ดังนี้

1) ขั้นตอนที่ 1 : เริ่มต้นกำหนดพารามิเตอร์เริ่มต้นสำหรับ CS เช่น จำนวนของรังนก (Population Size) อัตราการแทนที่รัง (p_a) และจำนวนรอบการทำซ้ำ (Iterations) เป็นต้น
2) ขั้นตอนที่ 2 : สร้างประชากรเริ่มต้นของรังนกแต่ละรัง โดยการสุ่มชุดของคุณลักษณะ (Feature Subset) ในรูปแบบของไบนารี เช่น [1, 0, 0, 1, 1, ...] ซึ่งเลข 1 หมายถึงเลือกคุณลักษณะนั้น ส่วน 0 หมายถึงไม่เลือกคุณลักษณะ

3) ขั้นตอนที่ 3 : ทำซ้ำจนกว่าจะถึงจำนวนรอบที่กำหนด (Max Iterations) เพื่อทำการปรับปรุงรังด้วยการสร้างไข่ใหม่โดยใช้ Lévy Flights ซึ่งเป็นการเคลื่อนที่แบบสุ่มระยะไกล ที่ช่วยให้ค้นหาคำตอบในพื้นที่ที่กว้าง
4) ขั้นตอนที่ 4 : สุ่มเลือกรังเพื่อวางไข่ หรือการสร้าง Solution ใหม่แบบสุ่มหนึ่งชุด และเปรียบเทียบ Fitness กับรังที่เลือก หากรังใหม่ดีกว่า (Fitness ดีกว่า) ก็จะทำการแทนที่
5) ขั้นตอนที่ 5 : ใช้กลไกการแทนที่บางรัง $(1 - p_a)$ เพื่อจำลองการที่ไข่ของนกคัคคูถูกตรวจพบและทำลาย
6) ขั้นตอนที่ 6 : คัดเลือกรังที่มีไข่ดีที่สุด หรือรังที่มี Fitness ดีที่สุด (Accuracy สูงสุด หรือเลือก Feature ได้อย่างมีประสิทธิภาพที่สุด) จะถูกเก็บไว้เป็นคำตอบที่ดีที่สุด
7) ขั้นตอนที่ 7 : สิ้นสุดเมื่อถึงจำนวนรอบที่กำหนด และคืนค่าคำตอบที่ดีที่สุด

จากคุณสมบัติในการสำรวจพื้นที่ได้กว้างและการแทนที่คำตอบที่ด้อยประสิทธิภาพอย่างต่อเนื่อง Cuckoo Search จึงถูกนำมาใช้ในการเลือกคุณลักษณะ อย่างแพร่หลาย โดยเฉพาะในบริบทของข้อมูลทางการแพทย์ที่มักมีความซับซ้อนสูงและมีจำนวนคุณลักษณะมาก เช่น การเลือกตัวแปรที่สำคัญก่อนนำไปใช้ร่วมกับโมเดลจำแนกประเภท เช่น Support Vector Machine (SVM) หรือ k-Nearest Neighbors (k-NN) เพื่อเพิ่มประสิทธิภาพของการพยากรณ์และตีความผลลัพธ์ [14]

2.2.3 อัลกอริทึม Firefly Algorithm (FA)

Firefly Algorithm (FA) เป็นอัลกอริทึมเชิงเมตาฮิวริสติกที่จำลองพฤติกรรมการดึงดูดระหว่างหิ่งห้อยในธรรมชาติ โดยใช้แนวคิดว่าหิ่งห้อยแต่ละตัวจะถูกดึงดูดเข้าหาหิ่งห้อยที่มีความสว่างมากกว่า ซึ่งความสว่างในที่นี้เปรียบเทียบได้กับค่าความเหมาะสม (Fitness Value) ของคำตอบในบริบทของการเพิ่มประสิทธิภาพ แนวทางนี้ช่วยให้แต่ละตัวแทนสามารถปรับปรุงคำตอบของตนเองโดยอาศัยข้อมูลจากตัวแทนที่มีคุณภาพดีกว่าในประชากร ส่งผลให้สามารถขัดเกลาคำตอบและหลีกเลี่ยงการติดกับดักค่าที่เหมาะสมเฉพาะจุดได้อย่างมีประสิทธิภาพ

หลักการพื้นฐานของ FA อยู่ที่การลดความเข้มของแสงตามระยะทาง และการเคลื่อนที่ของหิ่งห้อยตัวที่ i ไปยังหิ่งห้อยตัวที่ j ที่สว่างกว่าสามารถคำนวณได้ตามสมการ แสดงดังสมการที่ (2.2)

$$X_i^{(t+1)} = X_i^{(t)} + \beta_0 \cdot e^{-\gamma \cdot r_{ij}^2} \cdot \left(X_j^{(t)} - X_i^{(t)}\right) + \alpha \cdot \varepsilon_i \qquad (2.2)$$

โดยที่

- $X_i^{(t)}$ คือ ตำแหน่งของหิ่งห้อยตัวที่ i ในรอบที่ t
- β_0 คือ ค่าความเข้มของแสงที่จุดเริ่มต้น (Initial Attractiveness)
- γ คือ พารามิเตอร์ควบคุมอัตราการลดลงของแสงตามระยะทาง

- r_ij คือ ระยะทางระหว่างหิ่งห้อย i และ j
- α คือ ค่าสุ่มที่ควบคุมการเคลื่อนไหวแบบสุ่ม
- ε_i คือ เวกเตอร์สุ่มที่มีการแจกแจงแบบ Gaussian

Algorithm 2 Firefly Algorithm

Objective function: $f(X), X = (x_1, x_2, \cdots, x_d)^T$
Generate initial population of fireflies X_i $(i = 1, 2, \cdots, n)$
Evaluate light intensity I_i at X_i using $f(X_i)$
Define light absorption coefficient γ
while $t <$ MaxGeneration **do**
  for $i = 1$ to n **do**
    for $j = 1$ to n **do**
      if $I_j > I_i$ **then**
        Move firefly i towards firefly j in d-dimension
        Attractiveness varies with distance r via $\exp(-\gamma r^2)$
        Evaluate new solutions and update light intensity
      end if
    end for
  end for
  Rank the fireflies and find the current best
end while
Postprocess results and visualization

รูปที่ 2.2 อัลกอริทึมของ FA

จากสมการข้างต้นเป็นส่วนหนึ่งของคำนวณการเคลื่อนที่ของหิ่งห้อยในขั้นตอนวิธี Firefly Algorithm แสดงดังรูปที่ 2.2 โดยสามารถอธิบายรายละเอียดในแต่ละขั้นตอนได้ดังนี้

1) ขั้นตอนที่ 1 : กำหนดค่าพารามิเตอร์เริ่มต้นสำหรับ FA เช่น จำนวนของหิ่งห้อย (Population Size) จำนวนรอบการทำซ้ำ (Iterations) ความสามารถในการดึงดูดเริ่มต้น (β_0) อัตราการเสื่อมของการดึงดูด (γ) และอัตราการสุ่ม (α) เป็นต้น
2) ขั้นตอนที่ 2 : สร้างประชากรเริ่มต้นของหิ่งห้อย โดยแต่ละตัวแทนชุดของคุณลักษณะที่เลือก กำหนดตำแหน่งเริ่มต้นของหิ่งห้อย และคำนวณค่าความสว่างตาม Fitness
3) ขั้นตอนที่ 3 : ทำซ้ำจนกว่าจะครบจำนวนรอบที่กำหนด โดยในแต่ละรอบหิ่งห้อยทุกตัวจะค้นหาหิ่งห้อยที่สว่างกว่าตนเอง จากนั้นเคลื่อนที่เข้าไปใกล้หิ่งห้อยนั้น โดยตำแหน่งจะเปลี่ยนแปลงตามระดับความสว่างและระยะห่าง จากนั้นเพิ่มความสุ่ม (Randomness) เพื่อเพิ่มความหลากหลายของการสำรวจ
4) ขั้นตอนที่ 4 : อัปเดตตำแหน่ง (Solution) การเคลื่อนที่ของหิ่งห้อยจะใช้สมการอัปเดตตำแหน่งโดยรวมแรงดึงดูดและสุ่ม (Random Walk) ในบริบทของ Feature Selection การอัปเดตจะมีการแปลงค่าต่อท้ายเป็นไบนารี

5) ขั้นตอนที่ 5 : คำนวณค่า Fitness ใหม่ และอัปเดตความสว่าง โดยใช้โมเดล SVM ประเมินคุณภาพของ Feature Subset ใหม่ แล้วอัปเดตค่าความสว่างตามผลลัพธ์
6) ขั้นตอนที่ 6 : จดจำหิ่งห้อยที่ดีที่สุด (Best Solution) ค่าที่ดีที่สุดตลอดการทำซ้ำจะถูกเก็บไว้ เพื่อใช้เป็นผลลัพธ์สุดท้ายของอัลกอริทึม

อัลกอริทึม FA ได้รับความนิยมในการประยุกต์ใช้กับปัญหาเลือกคุณลักษณะ (Feature Selection) เนื่องจากมีความสามารถในการค้นหาคำตอบที่เหมาะสมในปัญหาที่มีลักษณะไม่เป็นเชิงเส้น (Nonlinear) หรือมีมิติมาก (High-dimensional) โดยเฉพาะในบริบทของข้อมูลทางการแพทย์ ซึ่งมักมีข้อมูลซับซ้อน และมีสัญญาณรบกวนสูง การใช้ FA ช่วยปรับปรุงประสิทธิภาพของโมเดลจำแนกประเภทได้อย่างมีนัยสำคัญ [15]

2.2.4 อัลกอริทึมผสม Cuckoo Search และ Firefly (Hybrid Cuckoo–Firefly Algorithm: CSFA)

แม้อัลกอริทึม Cuckoo Search (CS) และ Firefly Algorithm (FA) ต่างก็มีจุดแข็งเฉพาะด้านในการแก้ปัญหาเพิ่มประสิทธิภาพ (Optimization) แต่ก็มีข้อจำกัดบางประการที่ส่งผลต่อประสิทธิภาพในบางกรณี เช่น CS มีแนวโน้มสำรวจพื้นที่ค้นหาได้กว้างด้วย Lévy Flight แต่การปรับแต่งคำตอบ (Exploitation) อาจยังไม่ละเอียดพอ ขณะที่ FA มีพฤติกรรมในการปรับแต่งคำตอบที่เข้มแข็งแต่การสำรวจในพื้นที่ใหม่อาจจำกัด ด้วยเหตุนี้จึงเกิดแนวคิดในการผสมผสานอัลกอริทึมทั้งสอง เพื่อเสริมจุดแข็งซึ่งกันและกัน

Hybrid CSFA มีหลักการโดยทั่วไปคือการใช้โครงสร้างหลักของ FA ในการเคลื่อนที่ของคำตอบ (หิ่งห้อย) ร่วมกับกลไกการกระโดดแบบ Lévy Flight จาก CS เพื่อเสริมความสามารถในการสำรวจพื้นที่ค้นหาใหม่ที่หลากหลายมากขึ้น โดยอาจสลับหรือเลือกใช้วิธีการอัปเดตตำแหน่งจากทั้งสองอัลกอริทึมในสัดส่วนที่เหมาะสม การปรับสมดุลระหว่างการใช้ FA เพื่อขัดเกลาคำตอบที่อยู่ใกล้ๆ กับการใช้ CS เพื่อสุ่มกระโดดสำรวจพื้นที่ที่ไกลออกไป เป็นกุญแจสำคัญที่ทำให้ CSFA มีประสิทธิภาพสูงกว่าอัลกอริทึมเดี่ยวในหลายกรณี

ในงานวิจัยด้านการเลือกคุณลักษณะ (Feature Selection) โดยเฉพาะในข้อมูลการแพทย์ที่มีลักษณะมิติสูง และมีข้อมูลสัญญาณรบกวน การใช้อัลกอริทึมไฮบริดอย่าง CSFA มีแนวโน้มช่วยลดจำนวนคุณลักษณะที่ไม่จำเป็นลงได้มาก พร้อมกับรักษาหรือเพิ่มความแม่นยำของแบบจำลองจำแนกประเภท เช่น SVM ได้ดียิ่งขึ้น โดยเฉพาะเมื่อต้องการลด Overfitting และเพิ่มความสามารถในการทำนายผลลัพธ์กับข้อมูลใหม่

2.2.5 อัลกอริทึม Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) เป็นอัลกอริทึมเชิงเมตาฮิวริสติกที่ได้รับแรงบันดาลใจจากพฤติกรรมการเคลื่อนที่ของฝูงนกหรือฝูงปลา โดยมีแนวคิดพื้นฐานว่าตัวแทนแต่ละตัว (Particle) จะเคลื่อนที่ไปในพื้นที่ค้นหาโดยอาศัยทั้งประสบการณ์ของตนเองและประสบการณ์ของเพื่อนร่วมฝูง เป้าหมายคือการหาตำแหน่งที่ให้ค่าความเหมาะสมสูงที่สุดในพื้นที่ค้นหาของปัญหาที่กำลังพิจารณา [16]

ใน PSO แต่ละอนุภาคจะมีตำแหน่งและความเร็ว โดยตำแหน่งแทนคำตอบของปัญหา และความเร็วเป็นค่าที่กำหนดทิศทางในการเคลื่อนที่ สมการที่ใช้อัปเดตค่าความเร็ว แสดงดังสมการที่ (2.3) และสมการอัปเดตตำแหน่งของแต่ละอนุภาค ดังแสดงในสมการที่ (2.4)

$$V_i^{(t+1)} = W \cdot V_i^{(t)} + c_1 \cdot r_1 \cdot \left(pbest_i - X_i^{(t)}\right) + c_2 \cdot r_2 \cdot \left(gbest - X_i^{(t)}\right) \quad (2.3)$$

$$X_i^{(t+1)} = X_i^{(t)} + V_i^{(t+1)} \quad (2.4)$$

โดยที่

- $X_i^{(t)}$ คือ ตำแหน่งของอนุภาคตัวที่ i ในรอบที่ t
- $V_i^{(t)}$ คือ ความเร็วอนุภาคตัวที่ i ในรอบที่ t
- $pbest_i$ คือ ตำแหน่งที่ดีที่สุดที่อนุภาค i เคยพบ
- $gbest$ คือ ตำแหน่งที่ดีที่สุดที่อนุภาคทุกตัวในฝูงเคยพบ
- W คือ พารามิเตอร์น้ำหนักโมเมนตัม (Inertia Weight)
- c_1 และ c_2 คือ ค่าสัมประสิทธิ์การเรียนรู้ (Learning Factors)
- r_1 และ r_2 คือ ค่าสุ่มระหว่าง 0 ถึง 1

PSO มีจุดเด่นคือโครงสร้างที่เรียบง่าย ปรับใช้ได้ง่าย และสามารถหาคำตอบที่ดีได้ภายในเวลาอันสั้น ทำให้ได้รับความนิยมในงานที่มีพื้นที่ค้นหาขนาดใหญ่ โดยเฉพาะในงานเลือกคุณลักษณะ PSO ถูกนำมาใช้ในการเลือกชุดคุณลักษณะที่เหมาะสมกับปัญหาการจำแนกประเภท โดยเฉพาะข้อมูลทางการแพทย์ที่มีจำนวนตัวแปรจำนวนมากและมีความสัมพันธ์ที่ซับซ้อน การเลือกคุณลักษณะด้วย PSO ช่วยลดจำนวนคุณลักษณะที่ไม่จำเป็น ขณะเดียวกันยังคงรักษาความแม่นยำของโมเดลไว้ได้อย่างมีประสิทธิภาพ [17]

2.2.6 อัลกอริทึมแบบผสมผสานระหว่าง Particle Swarm Optimization (PSO) กับ Cuckoo Search (CS)

แนวคิดของการผสมผสานอัลกอริทึม PSO และ CS (Hybrid PSOCS) มุ่งเน้นไปที่การนำข้อดีของแต่ละอัลกอริทึมมาทำงานร่วมกัน โดยอัลกอริทึม PSO มีข้อได้เปรียบในการค้นหาคำตอบได้อย่าง

รวดเร็วในช่วงเริ่มต้นของการเรียนรู้ ในขณะที่อัลกอริทึม CS มีจุดแข็งในการหลีกเลี่ยงการติดอยู่ในค่าที่ดีที่สุดเฉพาะที่ผ่านการกระโดดค้นหาด้วย Lévy Flight การออกแบบอัลกอริทึม Hybrid PSOCS อาจดำเนินการได้หลายรูปแบบ เช่น การใช้กลไกของ PSO ในการอัปเดตตำแหน่งของประชากรหลัก การใช้ Lévy Flight จาก CS ในบางช่วงของการอัปเดต เพื่อเสริมการสำรวจพื้นที่ค้นหา หรือการปรับแต่งอนุภาคใกล้ตำแหน่งที่ดีที่สุดของ PSO ด้วยการเคลื่อนที่แบบสุ่มจาก CS

จากงานวิจัยที่ผ่านมา พบว่าการผสมผสานระหว่าง PSO และ CS ให้ผลลัพธ์ที่ดีขึ้นเมื่อเทียบกับการใช้อัลกอริทึมเดี่ยว โดยเฉพาะอย่างยิ่งในการแก้ปัญหาที่มีความซับซ้อนสูง [18], [19] ในการเลือกคุณลักษณะได้รับความสนใจอย่างมากในช่วงไม่กี่ปีที่ผ่านมา เนื่องจากสามารถลดจำนวนฟีเจอร์ที่ไม่จำเป็นออกไปได้อย่างมีประสิทธิภาพ และยังคงรักษาค่าความแม่นยำในการจำแนกประเภทไว้ได้ในระดับสูง โดยเฉพาะเมื่อนำไปใช้ร่วมกับตัวจำแนกประเภท เช่น Support Vector Machine (SVM) [20],[21]

2.2.7 อัลกอริทึมแบบผสมผสานระหว่าง Particle Swarm Optimization (PSO) กับ Firefly Algorithm (FA)

แนวคิดของการผสมผสานระหว่าง PSO และ FA (PSOFA) คือการรวมข้อดีของทั้งสองอัลกอริทึมเข้าด้วยกัน เพื่อให้ได้ทั้งความสามารถในการแสวงหาคำตอบที่ดีอย่างรวดเร็วจาก PSO และความสามารถในการกระจายการค้นหาและหลีกเลี่ยง Local Optima จาก FA ลักษณะการผสมผสานอาจดำเนินการได้หลายรูปแบบ เช่น ใช้ PSO ในการอัปเดตประชากรหลัก และใช้ FA เป็นกลไกช่วยสำรวจรอบบริเวณ gbest ใช้ FA เพื่อสุ่มกระจายอนุภาคบางส่วนเมื่อการค้นหาของ PSO หยุดนิ่ง หรือสลับขั้นตอนระหว่างการอัปเดตของ PSO และการดึงดูดของ FA

การวิจัยที่ผ่านมาพบว่า PSOFA ช่วยเพิ่มความสามารถในการสำรวจพื้นที่ค้นหาและทำให้ผลลัพธ์มีความเสถียรมากขึ้น โดยเฉพาะในการประยุกต์กับการเลือกคุณลักษณะ ซึ่งต้องเผชิญกับพื้นที่ค้นหาขนาดใหญ่และซับซ้อน [22]

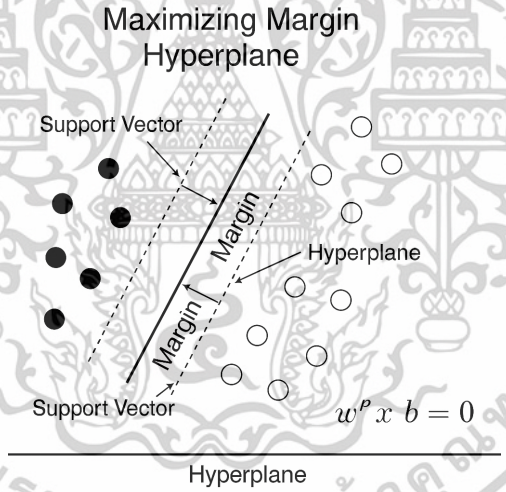
2.3 การจำแนกประเภทด้วย Support Vector Machine

Support Vector Machine (SVM) เป็นเทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่ได้รับความนิยมสูงในการจำแนกประเภท (Classification) โดยเฉพาะในกรณีที่ข้อมูลมีจำนวนตัวอย่างน้อย แต่มีจำนวนคุณลักษณะสูง ซึ่งเป็นลักษณะสำคัญของข้อมูลทางชีวการแพทย์ SVM มีหลักการพื้นฐานในการสร้าง ไฮเปอร์เพลน (Hyperplane) ที่สามารถแบ่งข้อมูลออกเป็นสองกลุ่มให้ห่างจากจุดข้อมูลที่ใกล้ที่สุดในแต่ละกลุ่ม (เรียกว่า Support Vectors) ด้วย ระยะขอบ (Margin) ที่กว้างที่สุด แสดงดังรูปที่ 2.3 เพื่อให้ได้โมเดลที่มีความสามารถในการจำแนกข้อมูลใหม่ได้อย่างมีประสิทธิภาพ [23]

ในกรณีที่ข้อมูลไม่สามารถแยกได้ด้วยเส้นตรงในมิติเดิม SVM สามารถขยายขอบเขตการทำงานได้ด้วย Kernel Trick ซึ่งเป็นการแปลงข้อมูลจากมิติดั้งเดิมไปยัง มิติที่สูงขึ้น (Feature Space) ที่เหมาะสมกว่า โดยไม่ต้องคำนวณพิกัดจริงในมิตินั้น เทคนิคนี้ช่วยให้สามารถสร้างไฮเปอร์เพลนในลักษณะที่ไม่เป็นเส้นตรงได้อย่างมีประสิทธิภาพ Kernel ที่ได้รับความนิยม ได้แก่:

- Radial Basis Function (RBF): จับความสัมพันธ์ที่ไม่เชิงเส้นได้ดี เหมาะกับข้อมูลทางการแพทย์ที่มีลักษณะซับซ้อน
- Polynomial Kernel: ใช้สำหรับกรณีที่ความสัมพันธ์มีลักษณะเป็นพหุนาม
- Sigmoid Kernel: คล้ายกับการทำงานของโครงข่ายประสาทเทียม

ในการประยุกต์ใช้ SVM ร่วมกับกระบวนการเลือกคุณลักษณะ (Feature Selection) SVM จะทำหน้าที่เป็น ตัวจำแนกหลัก (Base Classifier) ที่ใช้ประเมินคุณภาพของชุดคุณลักษณะที่ถูกคัดเลือกโดยอัลกอริทึมเชิงวิวัฒนาการ หรืออัลกอริทึมความฉลาดแบบกลุ่ม เช่น PSO, GA, CS หรือ FA โดยค่าความแม่นยำ (Accuracy) ของโมเดล SVM จะถูกใช้เป็น ฟังก์ชันวัดความเหมาะสม (Fitness Function) ในการคัดเลือกชุดคุณลักษณะที่ให้ผลลัพธ์ดีที่สุด [24]



รูปที่ 2.3 SVM Hyperplane and Margins in 2D Classification

2.4 การประเมินประสิทธิภาพของแบบจำลองการจำแนกประเภท

ในการประเมินประสิทธิภาพของกระบวนการเลือกคุณลักษณะร่วมกับแบบจำแนกประเภท เช่น SVM จำเป็นต้องใช้ตัวชี้วัดที่สามารถสะท้อนคุณภาพของโมเดลได้อย่างครอบคลุม โดยเฉพาะในบริบทของข้อมูลทางการแพทย์ที่ผลลัพธ์ของการจำแนกประเภทอาจส่งผลกระทบโดยตรงต่อการวินิจฉัยและการรักษาผู้ป่วย ตัวชี้วัดหลักที่นิยมใช้ ได้แก่ Accuracy, Precision, Recall และ F1-score ในบางกรณีอาจรวมถึง ROC-AUC เพื่อประเมินความสามารถของแบบจำลองในการแยกกลุ่ม

ข้อมูลออกจากกันอย่างแม่นยำ [25] สำหรับการประเมินประสิทธิภาพของกระบวนการทั้งหมด การใช้ตัวชี้วัดที่เหมาะสมถือเป็นสิ่งสำคัญ โดยตัวชี้วัดที่นิยมใช้ ได้แก่

- Accuracy คือ สัดส่วนของจำนวนตัวอย่างที่ถูกจำแนกได้อย่างถูกต้อง ทั้งในกลุ่มบวกและกลุ่มลบ เมื่อเทียบกับจำนวนตัวอย่างทั้งหมด เป็นตัวชี้วัดที่ง่ายต่อการตีความและนิยมใช้เป็นเบื้องต้นในการประเมินประสิทธิภาพของโมเดลจำแนกประเภท แสดงดังสมการที่ (2.5)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

- Precision คือ อัตราส่วนของตัวอย่างที่ถูกจำแนกเป็นกลุ่มบวกอย่างถูกต้อง (True Positive) ต่อจำนวนตัวอย่างที่แบบจำลองจำแนกว่าเป็นบวกทั้งหมด (รวมทั้ง FP) เป็นตัวชี้วัดที่สะท้อนความน่าเชื่อถือของแบบจำลองเมื่อแสดงผลว่าเป็นบวก แสดงดังสมการที่ (2.6)

$$Precision = \frac{TP}{TP + FN} \tag{2.6}$$

- Recall หรือ Sensitivity คือ อัตราส่วนของตัวอย่างบวกทั้งหมดที่แบบจำลองสามารถตรวจจับได้อย่างถูกต้อง เป็นตัวชี้วัดความสามารถของแบบจำลองในการระบุกลุ่มเป้าหมายที่แท้จริง แสดงดังสมการที่ (2.7)

$$Recall = \frac{TP}{TP + FN} \tag{2.7}$$

- F1-score เป็นค่าเฉลี่ยเชิงฮาร์มอนิก (Harmonic Mean) ระหว่าง Precision และ Recall โดยเน้นความสมดุลระหว่างทั้งสองค่ามากกว่าค่าเฉลี่ยเลขคณิต เหมาะสำหรับกรณีที่ต้องการให้ความสำคัญกับทั้งการลดผลลบเทียม (FN) และผลบวกเทียม (FP) พร้อมกัน โดยเฉพาะในข้อมูลที่ไม่สมดุล ซึ่ง F1-score สามารถเป็นตัวชี้วัดหลักในการตัดสินคุณภาพของแบบจำลองได้อย่างมีประสิทธิภาพ แสดงดังสมการที่ (2.8)

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \tag{2.8}$$

- AUC (Area Under the ROC Curve) คือ พื้นที่ใต้เส้นโค้งซึ่งแสดงค่าความสัมพันธ์ระหว่าง True Positive Rate (Recall) และ False Positive Rate ที่หลากหลายค่า Threshold เป็นตัวชี้วัดที่ประเมินความสามารถของแบบจำลองในการแยกแยะระหว่างกลุ่มบวก และลบอย่างต่อเนื่อง โดยค่าที่เข้าใกล้ 1 หมายถึง แบบจำลองสามารถจัดลำดับความน่าจะเป็นได้แม่นยำสูง ในขณะที่ค่าใกล้ 0.5 บ่งชี้ว่าแบบจำลองไม่มีความสามารถในการจำแนกเลย

การเลือกใช้ตัวชี้วัดควรขึ้นอยู่กับวัตถุประสงค์ของระบบ เช่น หากต้องการระบุผู้ป่วยที่แท้จริง Recall จะมีความสำคัญมากกว่า แต่ถ้าต้องการลดความผิดพลาดที่เกิดจากการจำแนกผิดเป็นโรค Precision จะมีบทบาทมากกว่า ดังนั้นการประเมินแบบหลากหลายตัวชี้วัดจึงเป็นแนวทางที่เหมาะสม โดยเฉพาะในการประยุกต์ใช้กับงานด้านการแพทย์ซึ่งผลลัพธ์มีผลกระทบโดยตรงต่อการรักษาผู้ป่วย [26]

2.5 งานวิจัยที่เกี่ยวข้อง

การเลือกคุณลักษณะ หรือ Feature Selection เป็นกระบวนการสำคัญในการลดมิติของข้อมูล เพิ่มประสิทธิภาพในการเรียนรู้ของแบบจำแนกประเภท และลดเวลาในการประมวลผล โดยเฉพาะในงานด้านการแพทย์ที่มีลักษณะข้อมูลซับซ้อนและมีจำนวนคุณลักษณะจำนวนมาก อัลกอริทึมความฉลาดแบบกลุ่ม (SI) ได้รับความสนใจอย่างมากเนื่องจากสามารถค้นหาคำตอบที่เหมาะสมได้ในพื้นที่คำตอบขนาดใหญ่ภายใต้ข้อจำกัดเชิงเวลา

ในช่วงที่ผ่านมา มีการประยุกต์ใช้อัลกอริทึมความฉลาดแบบกลุ่ม อย่างแพร่หลายในการเลือกคุณลักษณะ เนื่องจากอัลกอริทึมประเภทนี้มีความสามารถในการสำรวจพื้นที่คำตอบขนาดใหญ่ได้อย่างมีประสิทธิภาพ ไม่ติดอยู่กับความจำเพาะเฉพาะที่ และสามารถประยุกต์ใช้กับปัญหาที่ไม่มีโครงสร้างตายตัวได้ดี อัลกอริทึมยอดนิยม ได้แก่ Particle Swarm Optimization (PSO), Cuckoo Search (CS), Firefly Algorithm (FA) ตลอดจนรูปแบบแบบลูกผสม (Hybrid Algorithms) ซึ่งถูกพัฒนาเพื่อลดข้อจำกัดของอัลกอริทึมเดี่ยวและเพิ่มประสิทธิภาพในการค้นหาคำตอบที่ดีที่สุด [27], [28]

ในปี ค.ศ. 2018 Gherboudj et al. [29] ได้เสนอการประยุกต์ใช้ Cuckoo Search ร่วมกับ SVM สำหรับการเลือกคุณลักษณะจากข้อมูลโรคหัวใจ และสามารถเพิ่มความแม่นยำได้มากกว่า 90% โดยใช้คุณลักษณะเพียงบางส่วนของข้อมูลต้นฉบับ ส่วนในปี ค.ศ. 2020 Nandy et al. [30] ใช้ Firefly Algorithm เพื่อคัดเลือกฟีเจอร์จากข้อมูลสุขภาพ และเปรียบเทียบประสิทธิภาพกับอัลกอริทึมอื่น ๆ พบว่า FA ให้ผลลัพธ์ที่แม่นยำและเสถียรกว่าในกรณีข้อมูลไม่สมดุล ในปี ค.ศ. 2015 Alshamlan et al. [31] ใช้ PSO เพื่อเลือกยีนที่สำคัญจากข้อมูลไมโครอาร์เรย์ และได้ผลการจำแนกมะเร็งที่แม่นยำอย่างมีนัยสำคัญ ในปี ค.ศ. 2015 Wang et al. [32] ได้นำเสนอแนวทางการผสมผสานระหว่าง Particle Swarm Optimization (PSO) และ Cuckoo Search (CS) เพื่อเพิ่มประสิทธิภาพในการเลือกคุณลักษณะ โดยใช้กลไกการอัปเดตความเร็วและตำแหน่งของอนุภาคตามหลักการของ PSO ร่วมกับกลไกการกระโดดค้นหาแบบ Lévy Flight จาก CS เพื่อเสริมสร้างความหลากหลายให้กับประชากร การผสมผสานนี้ช่วยลดโอกาสการติดอยู่ในจุดที่ดีที่สุดเฉพาะที่ และเพิ่มความสามารถในการสำรวจพื้นที่ค้นหาได้กว้างขึ้น ผลการทดลองในหลายชุดข้อมูลแสดงให้เห็นว่า Hybrid PSOCS สามารถให้ค่าความแม่นยำสูง และลดจำนวนฟีเจอร์ที่เลือกได้อย่างมีประสิทธิภาพเมื่อเทียบกับอัลกอริทึมแบบเดี่ยว ในปี ค.ศ. 2020 Al-Betar et al. [33] ได้ทำการวิเคราะห์และสรุป

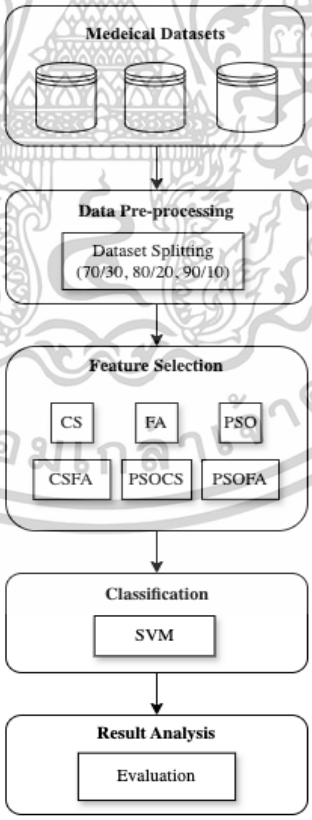
กลยุทธ์การผสมผสานของอัลกอริทึม Swarm Intelligence ที่ได้รับความนิยม เช่น PSO, CS และ FA โดยระบุว่า PSOCS เป็นหนึ่งในแนวทาง Hybrid ที่มีศักยภาพสูงในการแก้ปัญหาการเลือกคุณลักษณะ โดยเฉพาะอย่างยิ่งในปัญหาที่มีลักษณะซับซ้อนและมีมิติจำนวนมาก ในปี ค.ศ. 2017 Azzalini และ Mousavi [34] ได้นำเสนออัลกอริทึม Hybrid PSOFA ซึ่งผสานความสามารถของ PSO ในการค้นหาคำตอบเบื้องต้นได้อย่างรวดเร็ว กับความสามารถของ FA ในการหลีกเลี่ยงการติดอยู่ใน Local Optima ผ่านกลไกการดึงดูดแบบพฤติกรรมหิ่งห้อย อัลกอริทึมที่ได้ถูกนำมาประยุกต์ใช้ในงานเลือกคุณลักษณะบนชุดข้อมูลมาตรฐานหลายชุด และพบว่าสามารถเพิ่มค่าความแม่นยำในการจำแนกข้อมูล พร้อมทั้งลดจำนวนฟีเจอร์ลงได้อย่างชัดเจน นอกจากนี้ในปี ค.ศ. 2020 มีงานของ Naik และ Rautray [35] นำเสนออัลกอริทึม Hybrid PSOFA สำหรับการเลือกคุณลักษณะในงานวินิจฉัยทางการแพทย์ โดยเปรียบเทียบกับ PSO, FA และ GA แบบดั้งเดิม ผลการทดลองแสดงให้เห็นว่า Hybrid PSOFA มีความสามารถในการค้นหาคำตอบที่มีความเสถียร และสามารถเพิ่มประสิทธิภาพของโมเดล SVM ได้ดีกว่าอัลกอริทึมแบบเดี่ยว

จากงานวิจัยที่กล่าวมาข้างต้น แสดงให้เห็นว่าอัลกอริทึมความฉลาดแบบกลุ่มสามารถประยุกต์ใช้ในการเลือกคุณลักษณะในข้อมูลทางการแพทย์ได้อย่างมีประสิทธิภาพ โดยช่วยลดจำนวนคุณลักษณะที่ไม่จำเป็น ลดความซับซ้อนของแบบจำลอง และเพิ่มความแม่นยำของผลการจำแนก นอกจากนี้ยังช่วยลดเวลาในการประมวลผล ทำให้เหมาะสมกับงานที่ต้องการประสิทธิภาพเชิงเวลา โดยเฉพาะอย่างยิ่งเมื่อทำงานร่วมกับแบบจำแนกประเภทอย่าง Support Vector Machine (SVM) ซึ่งต้องอาศัยข้อมูลที่มีคุณภาพสูง การใช้วิธีการเลือกคุณลักษณะร่วมกับอัลกอริทึมเหล่านี้จึงสามารถปรับปรุงคุณภาพของแบบจำลองได้อย่างมีนัยสำคัญ และเป็นแนวทางที่มีแนวโน้มจะได้รับความนิยมในงานด้านการแพทย์และชีวสารสนเทศในอนาคต

บทที่ 3

วิธีการดำเนินงานวิจัย

การวิจัยนี้มุ่งเน้นการพัฒนากระบวนการเลือกคุณลักษณะ หรือ Feature Selection ที่เหมาะสมจากข้อมูลทางการแพทย์ แสดงดังรูปที่ 3.1 โดยใช้ขั้นตอนวิธีที่อิงกับความฉลาดแบบกลุ่ม (Swarm Intelligence Algorithms) ได้แก่ Cuckoo Search (CS), Firefly Algorithm (FA), Hybrid Cuckoo Search–Firefly Algorithm (CSFA), Particle Swarm Optimization (PSO), Hybrid Particle Swarm Optimization-Cuckoo Search (PSOCS) และ Hybrid Particle Swarm Optimization– Firefly Algorithm (PSOFA) เพื่อลดจำนวนคุณลักษณะ และเพิ่มประสิทธิภาพของแบบจำลองการจำแนกประเภท (Classification Model) โดยใช้ Support Vector Machine (SVM) เป็นตัวจำแนก และประเมินผลด้วยตัวชี้วัดมาตรฐาน ได้แก่ Accuracy, Precision, Recall และ F1-Score บนชุดข้อมูล Wisconsin Breast Cancer Diagnostic (WBCD), SPECTF Heart และ Arrhythmia โดยทั้งหมดเป็นชุดข้อมูลทางการแพทย์ เพื่อทดสอบความสามารถในการประยุกต์ใช้ของขั้นตอนวิธีต่างๆ



รูปที่ 3.1 ขั้นตอนภาพรวมของกระบวนการวิจัยสำหรับการเลือกคุณลักษณะและการจำแนกประเภทในข้อมูลทางการแพทย์

3.1 ชุดข้อมูลที่ใช้ในงานวิจัย

3.1.1 Wisconsin Diagnostic Breast Cancer (WDBC)

จัดทำโดย Dr. William H. Wolberg จาก University of Wisconsin Hospitals โดยข้อมูลถูกจัดเก็บจากการตรวจวินิจฉัยตัวอย่างเซลล์เต้านมผ่านกระบวนการ Fine Needle Aspiration (FNA) และแปลงเป็นคุณลักษณะเชิงตัวเลขที่สะท้อนลักษณะเฉพาะของเซลล์มะเร็ง ข้อมูลประกอบด้วยทั้งหมด 569 ตัวอย่างข้อมูล โดยจำแนกเป็น 2 กลุ่ม คือ

1. มะเร็งชนิดร้ายแรง (Malignant) จำนวน 212 ตัวอย่าง
2. มะเร็งชนิดไม่ร้ายแรง (Benign) จำนวน 357 ตัวอย่าง

โดยแต่ละตัวอย่างประกอบด้วย 30 คุณลักษณะเชิงตัวเลข (Numerical Features) ซึ่งได้จากการคำนวณค่าทางสถิติต่างๆ ของภาพเซลล์ ได้แก่ ค่าเฉลี่ย (Mean), ส่วนเบี่ยงเบนมาตรฐาน (Standard Error) และค่ามากสุด (Worst Value) ของคุณสมบัติตามรายการคุณลักษณะโดยแสดงดังตารางที่ 3.1

ตารางที่ 3.1 คุณลักษณะในชุดข้อมูล WDBC

ลำดับ	ชื่อคุณลักษณะพื้นฐาน	คำอธิบาย
1	Radius (mean)	ระยะเฉลี่ยจากศูนย์กลางถึงขอบเขตเซลล์ (cell boundary) - ค่าเฉลี่ย
2	Radius (se)	ระยะเฉลี่ยจากศูนย์กลางถึงขอบเขตเซลล์ (cell boundary) - ส่วนเบี่ยงเบนมาตรฐาน
3	Radius (worst)	ระยะเฉลี่ยจากศูนย์กลางถึงขอบเขตเซลล์ (cell boundary) - ค่ามากที่สุด
4	Texture (mean)	ความแปรปรวนของค่าความเข้มของพิกเซล (standard deviation) - ค่าเฉลี่ย
5	Texture (se)	ความแปรปรวนของค่าความเข้มของพิกเซล (standard deviation) - ส่วนเบี่ยงเบนมาตรฐาน
6	Texture (worst)	ความแปรปรวนของค่าความเข้มของพิกเซล (standard deviation) - ค่ามากที่สุด
7	Perimeter (mean)	ความยาวเส้นรอบวงของเซลล์ - ค่าเฉลี่ย
8	Perimeter (se)	ความยาวเส้นรอบวงของเซลล์ - ส่วนเบี่ยงเบนมาตรฐาน
9	Perimeter (worst)	ความยาวเส้นรอบวงของเซลล์ - ค่ามากที่สุด

ตารางที่ 3.1 (ต่อ) คุณลักษณะในชุดข้อมูล WDBC

ลำดับ	ชื่อคุณลักษณะพื้นฐาน	คำอธิบาย
10	Area (mean)	พื้นที่ของเซลล์ - ค่าเฉลี่ย
11	Area (se)	พื้นที่ของเซลล์ - ส่วนเบี่ยงเบนมาตรฐาน
12	Area (worst)	พื้นที่ของเซลล์ - ค่ามากที่สุด
13	Smoothness (mean)	ความเรียบของขอบเซลล์ (local variation) - ค่าเฉลี่ย
14	Smoothness (se)	ความเรียบของขอบเซลล์ (local variation) - ส่วนเบี่ยงเบนมาตรฐาน
15	Smoothness (worst)	ความเรียบของขอบเซลล์ (local variation) - ค่ามากที่สุด
16	Compactness (mean)	ค่าที่คำนวณจาก (perimeter² / area - 1.0) - ค่าเฉลี่ย
17	Compactness (se)	ค่าที่คำนวณจาก (perimeter² / area - 1.0) - ส่วนเบี่ยงเบนมาตรฐาน
18	Compactness (worst)	ค่าที่คำนวณจาก (perimeter² / area - 1.0) - ค่ามากที่สุด
19	Concavity (mean)	ความเว้าของขอบเซลล์ (severity of concave portions) - ค่าเฉลี่ย
20	Concavity (se)	ความเว้าของขอบเซลล์ (severity of concave portions) - ส่วนเบี่ยงเบนมาตรฐาน
21	Concavity (worst)	ความเว้าของขอบเซลล์ (severity of concave portions) - ค่ามากที่สุด
22	Concave Points (mean)	จำนวนจุดเว้าบนขอบเซลล์ - ค่าเฉลี่ย
23	Concave Points (se)	จำนวนจุดเว้าบนขอบเซลล์ - ส่วนเบี่ยงเบนมาตรฐาน
24	Concave Points (worst)	จำนวนจุดเว้าบนขอบเซลล์ - ค่ามากที่สุด
25	Symmetry (mean)	ความสมมาตรของเซลล์ - ค่าเฉลี่ย
26	Symmetry (se)	ความสมมาตรของเซลล์ - ส่วนเบี่ยงเบนมาตรฐาน
27	Symmetry (worst)	ความสมมาตรของเซลล์ - ค่ามากที่สุด
28	Fractal Dimension (mean)	ความซับซ้อนของขอบเซลล์ - ค่าเฉลี่ย
29	Fractal Dimension (se)	ความซับซ้อนของขอบเซลล์ - ส่วนเบี่ยงเบนมาตรฐาน
30	Fractal Dimension (worst)	ความซับซ้อนของขอบเซลล์ - ค่ามากที่สุด

ชุดข้อมูล WDBC ถือเป็นข้อมูลมาตรฐานที่ถูกนำมาใช้ในงานวิจัยด้านการเรียนรู้ของเครื่องและการแพทย์อย่างแพร่หลาย เนื่องจากมีความสมบูรณ์ และสามารถนำไปประยุกต์ใช้กับอัลกอริทึมต่างๆ ได้อย่างมีประสิทธิภาพ

3.1.2 SPECTF Heart

ชุดข้อมูล SPECTF Heart เป็นหนึ่งในชุดข้อมูลทางการแพทย์ที่ได้รับความนิยมในการศึกษาเกี่ยวกับการวินิจฉัยโรคหัวใจ โดยเผยแพร่ผ่านฐานข้อมูล UCI Machine Learning Repository เช่นเดียวกับชุดข้อมูล WDBC จุดมุ่งหมายของชุดข้อมูลนี้คือการใช้ข้อมูลที่ได้จากภาพสแกนหัวใจด้วยเทคนิค Single Photon Emission Computed Tomography (SPECT) เพื่อทำนายภาวะหัวใจขาดเลือด โดยชุดข้อมูลนี้ประกอบไปด้วยข้อมูลทั้งหมด 267 ตัวอย่าง โดยจำแนกเป็น 2 กลุ่ม คือ

1. กลุ่มผู้ป่วยที่มีภาวะหัวใจขาดเลือด (Abnormal) จำนวน 131 ตัวอย่าง
2. กลุ่มปกติ (Normal) จำนวน 136 ตัวอย่าง

ข้อมูลในแต่ละตัวอย่างประกอบด้วย 44 คุณลักษณะแบบเลขจำนวนเต็ม (Integer) ซึ่งได้จากค่าความเข้มของสัญญาณภาพ SPECT ในแต่ละตำแหน่ง จากการประเมินมุมมองทั้งแนวหน้าและแนวหลังของหัวใจ แสดงดังตารางที่ 3.2 คุณลักษณะเหล่านี้สะท้อนถึงระดับของการดูดกลืนสารรังสีในกล้ามเนื้อหัวใจ ซึ่งบ่งชี้ถึงการไหลเวียนของเลือด และสามารถใช้ในการวิเคราะห์ภาวะผิดปกติของหัวใจได้อย่างแม่นยำ

ตารางที่ 3.2 คุณลักษณะในชุดข้อมูล SPECTF Heart

ลำดับ	ชื่อคุณลักษณะพื้นฐาน	ตัวอย่างคุณลักษณะ	คำอธิบาย
1-22	SPECT ภาพแนวหน้า (Anterior)	SPECT Anterior 1, SPECT Anterior 2, SPECT Anterior 3	ความเข้มของสัญญาณจากตำแหน่งต่างๆ ในภาพสแกนมุมมองด้านหน้า
23-44	SPECT ภาพแนวหลัง (Posterior)	SPECT Posterior 1, SPECT Posterior 2, SPECT Posterior 3	ความเข้มของสัญญาณจากตำแหน่งต่างๆ ในภาพสแกนมุมมองด้านหลัง

ชุดข้อมูล SPECTF Heart จึงเหมาะสมอย่างยิ่งสำหรับการนำมาใช้ในการทดลองเปรียบเทียบประสิทธิภาพของอัลกอริทึมการเลือกคุณลักษณะ เนื่องจากมีระดับความซับซ้อนของข้อมูลอยู่ในระดับปานกลาง และมีจำนวนคุณลักษณะมากกว่าชุดข้อมูล WDBC ซึ่งช่วยให้สามารถประเมินศักยภาพของอัลกอริทึมได้อย่างครอบคลุมมากยิ่งขึ้น

3.1.2 Arrhythmia

ชุดข้อมูล Arrhythmia เป็นข้อมูลทางการแพทย์ที่เกี่ยวข้องกับการวินิจฉัยภาวะหัวใจเต้นผิดจังหวะ (Cardiac Arrhythmia) ซึ่งเผยแพร่โดย UCI Machine Learning Repository จุดมุ่งหมายของชุดข้อมูลนี้คือการจัดกลุ่มประเภทของคลื่นไฟฟ้าหัวใจ (ECG) เพื่อช่วยในการวิเคราะห์สภาวะความผิดปกติของระบบการเต้นของหัวใจ โดยชุดข้อมูลนี้ประกอบไปด้วย 452 ตัวอย่างข้อมูล โดยจำแนกเป็น 16 กลุ่ม แสดงดังตารางที่ 3.3

ตารางที่ 3.3 รายการจำแนก 16 กลุ่มในชุดข้อมูล Arrhythmia

กลุ่มที่	ชื่อกลุ่ม	จำนวน	คำอธิบาย
1	Normal	245	หัวใจเต้นปกติ
2	Ischemic changes (Coronary Artery Disease)	44	ภาวะขาดเลือดที่เกิดจากโรคหลอดเลือดหัวใจ
3	Old Anterior Myocardial Infarction	15	เคยเกิดกล้ามเนื้อหัวใจตายบริเวณด้านหน้า
4	Old Inferior Myocardial Infarction	15	เคยเกิดกล้ามเนื้อหัวใจตายบริเวณด้านล่าง
5	Sinus Tachycardia	13	ภาวะหัวใจเต้นเร็วเกินไปแบบไซนัส (มากกว่า 100 ครั้ง/นาที)
6	Sinus Bradycardia	25	ภาวะหัวใจเต้นช้าเกินไปแบบไซนัส (น้อยกว่า 60 ครั้ง/นาที)
7	Ventricular Premature Contraction (VPC)	3	หัวใจห้องล่างเต้นก่อนเวลา
8	Supraventricular Premature Contraction	2	หัวใจห้องบนเต้นก่อนเวลา
9	Left Bundle Branch Block (LBBB)	9	การนำไฟฟ้าผิดปกติในแขนงซ้ายของหัวใจ
10	Right Bundle Branch Block (RBBB)	50	การนำไฟฟ้าผิดปกติในแขนงขวาของหัวใจ
11	1st Degree Atrioventricular Block	0	การนำไฟฟ้าช้าระหว่างห้องบนและล่างของหัวใจ ระดับที่ 1

ตารางที่ 3.3 (ต่อ) รายการจำแนก 16 กลุ่มในชุดข้อมูล Arrhythmia

กลุ่มที่	ชื่อกลุ่ม	จำนวน	คำอธิบาย
12	2nd Degree AV Block	0	การนำไฟฟ้าขาดหายเป็นบางครั้งระหว่างห้องบนและล่างของหัวใจ
13	3rd Degree AV Block	0	การนำไฟฟ้าขาดหายทั้งหมดระหว่างห้องบนและล่างของหัวใจ
14	Left Ventricular Hypertrophy	4	หัวใจห้องล่างซ้ายหนา
15	Atrial Fibrillation (AF)	5	ภาวะหัวใจห้องบนสั่นพริ้ว
16	Others	22	กลุ่มอื่นๆ ที่ไม่อยู่ในประเภทข้างต้น

โดยมีจำนวนคุณลักษณะทั้ง 279 คุณลักษณะเชิงตัวเลข ได้มาจากค่าพารามิเตอร์ที่ได้จากการวิเคราะห์สัญญาณคลื่นไฟฟ้าหัวใจ (Electrocardiogram: ECG) ซึ่งสามารถจำแนกออกเป็นหมวดหมู่หลักได้หลากหลายประเภท แสดงดังตารางที่ 3.4

ตารางที่ 3.4 กลุ่มพารามิเตอร์หลักจากสัญญาณคลื่นไฟฟ้าหัวใจในชุดข้อมูล Arrhythmia

กลุ่มพารามิเตอร์	ตัวอย่างคุณลักษณะ	คำอธิบาย
ช่วงเวลา (Time Intervals)	PR Interval, QRS Duration, QT Interval	ช่วงเวลาระหว่างจุดสำคัญต่างๆ บนคลื่น ECG ซึ่งสะท้อนถึงการนำกระแสไฟฟ้าในหัวใจ
ความสูงของคลื่น (Amplitude)	P Wave Amplitude, T Wave Amplitude, R Peak	ความสูงของแต่ละคลื่นสะท้อนถึงความแรงของสัญญาณไฟฟ้าในแต่ละช่วงการทำงานของหัวใจ
ความถี่หรืออัตราการเต้น (Rate)	Heart Rate, RR Interval	ความถี่ในการเต้นของหัวใจ และช่วงเวลาระหว่างการเต้นแต่ละครั้ง
ความสัมพันธ์ของคลื่น (Morphology Ratios)	ST-T Ratio, P-R Segment Ratio	อัตราส่วนระหว่างคลื่นหรือช่วงต่างๆ ซึ่งช่วยวิเคราะห์ความผิดปกติทางไฟฟ้าของหัวใจ
ลักษณะเฉพาะทางไฟฟ้าอื่นๆ	Axis Deviation, Electrical Axis	ทิศทางการกระจายของสัญญาณไฟฟ้าภายในหัวใจ ซึ่งมีผลต่อการวินิจฉัยโรคต่างๆ

3.2 การแบ่งชุดข้อมูลเพื่อการทดลอง (Dataset Splitting for Experimentation)

เพื่อให้การประเมินประสิทธิภาพของกระบวนการเลือกคุณลักษณะและแบบจำลองจำแนกประเภทมีความเที่ยงตรงและน่าเชื่อถือ งานวิจัยนี้ได้ดำเนินการแบ่งชุดข้อมูลที่ใช้ในการทดลองออกเป็นสองชุด ได้แก่ ชุดข้อมูลฝึกสอน (Training Set) และ ชุดข้อมูลทดสอบ (Testing Set) โดยกำหนดสัดส่วนการแบ่งเป็น 3 รูปแบบ ได้แก่ 70:30, 80:20 และ 90:10 ตามลำดับ ทั้งนี้เพื่อศึกษาผลกระทบของปริมาณข้อมูลที่ใช้ในการฝึกสอนต่อประสิทธิภาพของอัลกอริทึมการเลือกคุณลักษณะและแบบจำลองจำแนกประเภท โดยมุ่งเน้นการตรวจสอบความเสถียรของผลลัพธ์ภายใต้ขนาดข้อมูลที่หลากหลาย

3.3 การเลือกคุณลักษณะด้วยขั้นตอนวิธีความฉลาดแบบกลุ่ม (Swarm Intelligence-Based Feature Selection)

ในการวิจัยนี้ ผู้วิจัยได้เลือกใช้ขั้นตอนวิธีเชิงวิวัฒนาการและความฉลาดแบบกลุ่ม (Swarm Intelligence) สำหรับการเลือกคุณลักษณะจากชุดข้อมูลทางการแพทย์ โดยมีเป้าหมายเพื่อลดจำนวนคุณลักษณะที่ไม่จำเป็น เพิ่มความแม่นยำในการจำแนกประเภท และลดความซับซ้อนของแบบจำลอง

3.3.1 Cuckoo Search (CS)

ขั้นตอนวิธี Cuckoo Search (CS) ได้รับแรงบันดาลใจจากพฤติกรรมการวางไข่ของนกคัคคูที่อาศัยรังของนกชนิดอื่นในการเลี้ยงลูกอ่อน โดยจำลองแนวคิดนี้ผ่านกระบวนการค้นหาคำตอบที่เหมาะสมที่สุด ซึ่งในบริบทของการเลือกคุณลักษณะ ไข่แต่ละฟองจะแทนชุดของคุณลักษณะที่ถูกเลือก (Feature Subset) และ ค่าความเหมาะสม (Fitness) ของไข่แต่ละฟองจะถูกประเมินด้วยโมเดล SVM โดยมีการกำหนดพารามิเตอร์เริ่มต้น แสดงดังตารางที่ 3.5

ตารางที่ 3.5 การกำหนดพารามิเตอร์เริ่มต้นสำหรับ CS

Parameters	Value
n	ตามจำนวนคุณลักษณะ
Max Iteration	200
p_a	0.25
λ	1.5

3.3.2 Firefly Algorithm (FA)

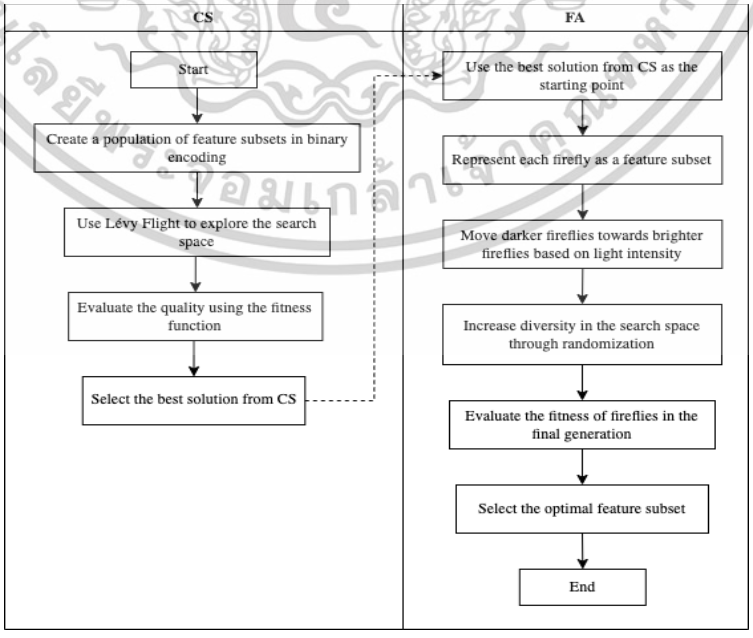
Firefly Algorithm (FA) เป็นขั้นตอนวิธีความฉลาดแบบกลุ่ม (Swarm Intelligence) ที่ได้รับแรงบันดาลใจจากพฤติกรรมของหิ่งห้อยซึ่งจะเคลื่อนที่เข้าหาแหล่งแสงที่สว่างกว่า หิ่งห้อยแต่ละตัวในบริบทของการเลือกคุณลักษณะจะแทนชุดของคุณลักษณะที่ถูกเลือก (Feature Subset) และระดับความสว่างจะสะท้อนถึงคุณภาพของคำตอบนั้นผ่านฟังก์ชันวัดความเหมาะสม (Fitness) การกำหนดค่าพารามิเตอร์เริ่มต้นสำหรับ FA แสดงดังตารางที่ 3.6

ตารางที่ 3.6 การกำหนดพารามิเตอร์เริ่มต้นสำหรับ FA

Parameters	Value
n	ตามจำนวนคุณลักษณะ
Max Iteration	200
β_0	1.0
γ	1.0
α	0.2

3.3.3 Hybrid Cuckoo Search - Firefly Algorithm (CSFA)

ขั้นตอนวิธีการผสมระหว่าง Cuckoo Search (CS) และ Firefly Algorithm (FA) ที่เรียกว่า CSFA ได้รับการออกแบบมาเพื่อรวมข้อดีของทั้งสองอัลกอริทึม โดยมุ่งเน้นการเพิ่มประสิทธิภาพในการสำรวจ (Exploration) และการแสวงหาคำตอบเชิงลึก (Exploitation) ซึ่งเป็นหัวใจสำคัญของกระบวนการเลือกคุณลักษณะที่เหมาะสม



รูปที่ 3.2 อัลกอริทึมของ CSFA

การทำงานถูกแบ่งออกเป็น 2 ระยะโดยเริ่มจาก CS ในระยะแรกเพื่อค้นหาชุดของคุณลักษณะที่ดีที่สุด แล้วส่งต่อไปยัง FA ในระยะที่สองเพื่อปรับแต่ง แสดงดังรูปที่ 3.2 โดยสามารถอธิบายรายละเอียดในแต่ละขั้นตอนได้ดังนี้

1) ขั้นตอนที่ 1 : เริ่มต้นสร้างประชากรของชุดคูณลักษณะในรูปแบบไบนารี
2) ขั้นตอนที่ 2 : ใช้กลไก Lévy Flight ในการสำรวจพื้นที่คำตอบ (Search Space) อย่างกว้างขวาง
3) ขั้นตอนที่ 3 : ประเมินคุณภาพของแต่ละรัง (คำตอบ) ด้วยฟังก์ชันวัดความเหมาะสม เช่น ความแม่นยำจาก SVM
4) ขั้นตอนที่ 4 : เลือกชุดคำตอบที่ดีที่สุดจาก CS เพื่อใช้เป็นจุดเริ่มต้นให้กับ FA
5) ขั้นตอนที่ 5 : ใช้ชุดคุณลักษณะที่ดีที่สุดจาก CS เป็นจุดเริ่มต้นของการค้นหาใน FA
6) ขั้นตอนที่ 6 : แทนแต่ละหิ่งห้อยเป็นชุดคุณลักษณะ และดำเนินการเคลื่อนที่ของหิ่งห้อยที่สว่างน้อยไปยังหิ่งห้อยที่สว่างกว่า
7) ขั้นตอนที่ 7 : เพิ่มความหลากหลายของคำตอบผ่านกลไกการสุ่ม (Randomization)
8) ขั้นตอนที่ 8 : ประเมิน fitness ของชุดคุณลักษณะในรุ่นสุดท้าย และเลือกชุดที่ดีที่สุดเป็นผลลัพธ์สุดท้ายของ CSFA

3.4 การจำแนกประเภท (Classification)

หลังจากดำเนินกระบวนการเลือกคุณลักษณะโดยใช้ขั้นตอนวิธีความฉลาดแบบกลุ่ม (CS, FA, CSFA, PSO, GA และ PSOGA) แล้ว ชุดคุณลักษณะที่ได้จะถูกนำไปใช้ในการจำแนกประเภทด้วยแบบจำลอง Support Vector Machine (SVM) เพื่อประเมินความสามารถในการแยกแยะกลุ่มข้อมูลของแต่ละอัลกอริทึม ในการทดลองนี้ SVM ถูกใช้เป็น ตัววัดความสามารถของการเลือกคุณลักษณะ (Wrapper Evaluation Method) โดยอาศัยผลลัพธ์ของความแม่นยำ (Accuracy) จาก SVM เป็นฟังก์ชันวัดความเหมาะสม (Fitness Function) ให้กับอัลกอริทึมที่ใช้เลือกคุณลักษณะ

การฝึกและทดสอบโมเดล SVM ดำเนินการโดยใช้ชุดข้อมูลที่แบ่งไว้แล้วในขั้นตอนก่อนหน้า และประเมินผลด้วยวิธี cross-validation และค่าตัวชี้วัดต่างๆ เช่น Accuracy, Precision, Recall และ F1-score เพื่อเปรียบเทียบประสิทธิภาพระหว่างแต่ละอัลกอริทึมเลือกคุณลักษณะ

บทที่ 4
ผลการวิจัยและการอภิปรายผล

ในบทนี้จะกล่าวถึงผลการวิจัยและการอภิปรายผลสำหรับการคัดเลือกคุณลักษณะโดยใช้ความฉลาดแบบกลุ่มกับชุดข้อมูลมะเร็งเต้านม และมีการขยายการประยุกต์ใช้ขั้นตอนวิธีดังกล่าวกับชุดข้อมูลทางการแพทย์อื่น เช่น ชุดข้อมูลการทำนายภาวะหัวใจขาดเลือด และชุดข้อมูลการวินิจฉัยภาวะการเต้นผิดปกติของหัวใจ ตามที่ได้เสนอไปในหัวข้อก่อนหน้า

โดยการทดลองได้มีการใช้ 3 ชุดข้อมูล และมีการแบ่งข้อมูลออกเป็น 3 ขนาดเพื่อใช้กับการคัดเลือกคุณลักษณะหลากหลายขั้นตอนวิธี แสดงดังตารางที่ 4.1

ตารางที่ 4.1 กรณีของการทดลองทั้งหมด

ลำดับ	ชุดข้อมูล	ขนาดของข้อมูล	ขั้นตอนวิธี	การจำแนกประเภท	การประเมินผลลัพธ์
1	มะเร็งเต้านม	90/10	CS FA PSO CSFA PSOCS PSOFA	SVM	Accuracy Precision F1-Score Recall AUC-ROC Selected Feature Computation Time
2		80/20			
3		70/30			
4	ภาวะหัวใจขาดเลือด	90/10			
5		80/20			
6		70/30			
7	การเต้นผิดปกติของหัวใจ	90/10			
8		80/20			
9		70/30			

จากขั้นตอนในการแบ่งขนาดของข้อมูล และใช้ขั้นตอนวิธีความฉลาดแบบกลุ่มกับการเลือกคุณลักษณะ เพื่อนำไปใช้ในการฝึกฝนแบบจำลองการจำแนกประเภทและวัดประเมินประสิทธิภาพผลลัพธ์ที่ได้จะสามารถบ่งบอกได้ถึงความเหมาะสมของการประยุกต์ใช้ขั้นตอนวิธีและการเลือกชุดข้อมูลที่เหมาะสมตามลักษณะของข้อมูล

4.1 ผลลัพธ์ของการทดลองด้วยชุดข้อมูลมะเร็งเต้านม

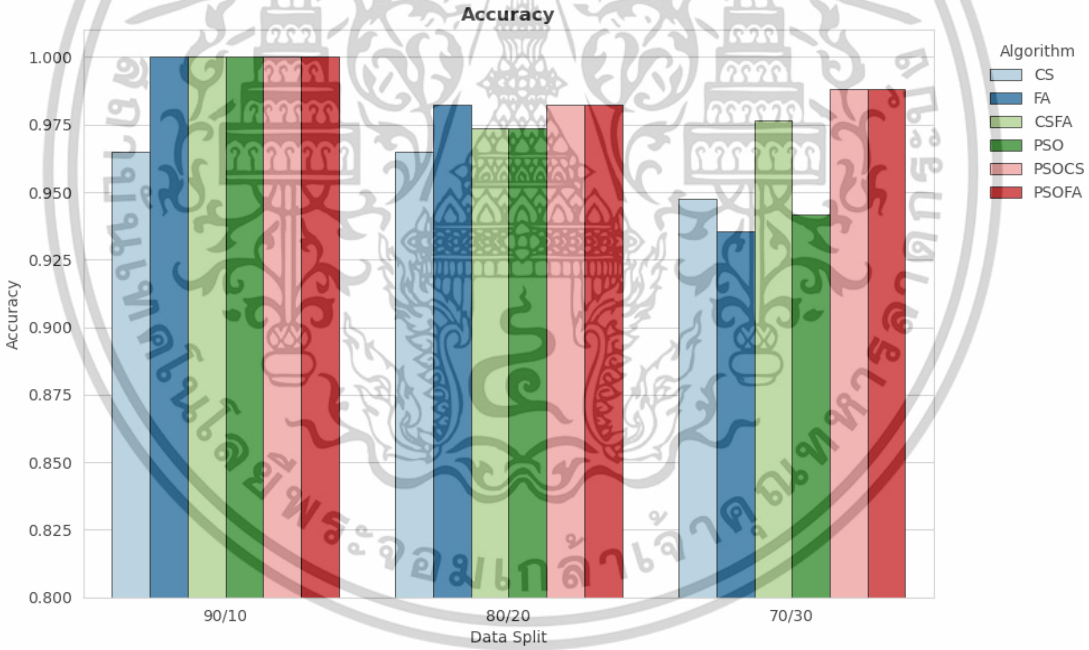
จากผลการทดลอง แสดงดังตารางที่ 4.2 พบว่าในขนาดของข้อมูล 90/10 ค่าความแม่นยำจะสูงมากในขนาดข้อมูลนี้ และลดน้อยลงในขนาดข้อมูล 80/20 และ 70/30 ซึ่งสะท้อนถึงปัญหาการจดจำข้อมูลฝึกมากเกินไป (Overfitting) ในทางตรงกันข้าม เมื่อแบ่งข้อมูลเป็นสัดส่วน 80/20 และ 70/30 ทำให้ขนาดของชุดข้อมูลการทดสอบมีมากขึ้น พบว่าค่าความแม่นยำมีแนวโน้มลดลงเล็กน้อย และมีการกระจายค่าของตัวชี้วัดอื่น เช่น Precision และ Recall มากขึ้น

ตารางที่ 4.2 ผลลัพธ์การทดลองด้วยชุดข้อมูลมะเร็งเต้านม

ขนาดข้อมูล	อัลกอริทึม	Accuracy	Precision	Recall	F1-Score	AUC-ROC	จำนวนคุณลักษณะที่ถูกเลือก
90/10	CS	96.49	95.65	95.65	95.65	96.35	13
	FA	100	100	100	100	100	15
	PSO	100	100	100	100	100	19
	CSFA	100	100	100	100	100	14
	PSOCS	100	100	100	100	100	16
	PSOFA	100	100	100	100	100	22
80/20	CS	96.49	97.50	92.85	95.12	95.73	17
	FA	98.24	100	95.23	97.56	97.61	15
	PSO	97.36	100	92.85	96.29	96.42	18
	CSFA	97.36	97.56	95.23	96.38	96.92	18
	PSOCS	98.24	100	95.23	97.56	97.61	15
	PSOFA	98.24	100	95.23	97.56	97.61	16
70/30	CS	94.73	92.18	93.65	92.91	94.51	15
	FA	93.56	90.62	92.06	91.33	93.25	12
	PSO	94.15	92.06	92.06	92.06	93.71	19
	CSFA	97.66	100	93.65	96.72	96.82	13
	PSOCS	98.83	100	96.82	98.38	98.41	14
	PSOFA	98.83	100	96.82	98.38	98.41	15

จากการเปรียบเทียบทั้ง 3 รูปแบบการแบ่งข้อมูล พบว่าอัลกอริทึมในกลุ่ม Hybrid ให้ค่าความแม่นยำสูง และสม่ำเสมอในทุกเงื่อนไขของข้อมูล ขณะที่อัลกอริทึมแบบเดี่ยวมีแนวโน้มลดลงอย่างชัดในชุดข้อมูลฝึกฝนน้อย โดยเฉพาะในขนาด 70/30 ซึ่งถือเป็นเงื่อนไขที่สะท้อนความสามารถของแบบจำลองได้ใกล้เคียงกับความเป็นจริงมากที่สุด อีกทั้งการผสมผสานยังสามารถเลือกคุณลักษณะได้อย่างเหมาะสม ไม่มากหรือน้อยเกินไป ซึ่งช่วยลดความเสี่ยงจากการจดจำข้อมูลฝึกมากเกินไป และยังรักษาความแม่นยำในระดับสูง

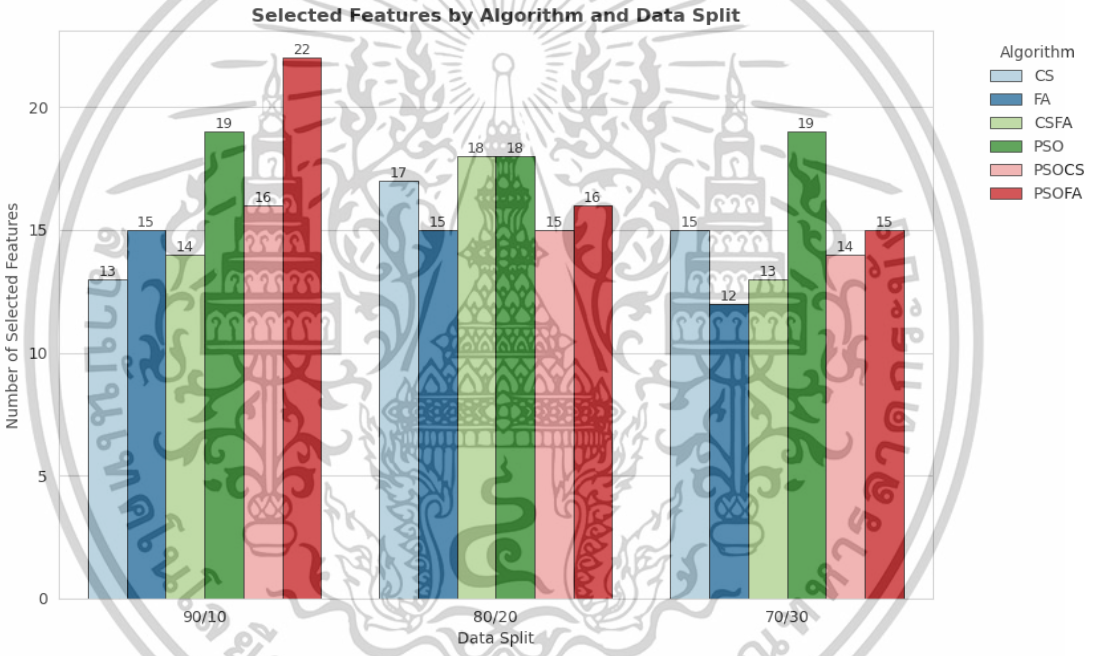
จากผลลัพธ์การประเมินประสิทธิภาพโดยใช้ตัวชี้วัดหลักทั้งในเชิงประสิทธิภาพการจำแนก ความสามารถในการทำนายข้อมูลใหม่ได้ (Generalize) อัลกอริทึม PSOCS (Hybrid PSO + Cuckoo Search) แสดงผลลัพธ์ที่โดดเด่นที่สุด โดยสามารถรักษาความแม่นยำระดับสูงในทุกขนาดการแบ่งข้อมูล โดยเฉพาะในขนาด 70/30 และยังสามารถเลือกคุณลักษณะได้อย่างมีประสิทธิภาพและใช้เวลาในการประมวลผลที่ต่ำกว่า PSOFA หรือว่า CSFA อย่างมีนัยสำคัญ จึงสามารถสรุปได้ว่า PSOCS เป็นอัลกอริทึมที่เหมาะสมที่สุดในการเลือกคุณลักษณะเพื่อการจำแนกประเภทในชุดข้อมูลมะเร็งเต้านม



รูปที่ 4.1 ผลการทดลองความแม่นยำการจำแนกประเภทในชุดข้อมูลมะเร็งเต้านม

แสดงให้เห็นผลลัพธ์ด้านความแม่นยำของแต่ละอัลกอริทึมและขนาดการแบ่งข้อมูล สังเกตได้ว่าอัลกอริทึมผสมผสานได้แก่ PSOCS และ PSOFA ให้ผลลัพธ์ที่ดีที่สุดตามด้วย CSFA โดยเห็นผลได้อย่างชัดเจนเมื่อเปรียบเทียบกับอัลกอริทึมแบบเดี่ยว แสดงดังรูปที่ 4.1 การผสมผสานให้ผลลัพธ์ที่ดีเนื่องจากอาศัยจุดแข็งของแต่ละตัวมาทำการผสมผสานกัน เช่น PSOCS โดย PSO เด่นในด้านการลู่เข้าหาคำตอบ (Convergence) อย่างรวดเร็วด้วยแนวคิดการเคลื่อนไหวตามประสบการณ์ที่ดีที่สุด ส่วน CS เสริมในเรื่องการสำรวจ (Exploration) โดยใช้ Lévy Flight ทำให้ Particle เคลื่อนที่ได้แบบ

กระโดดเพื่อหลีกเลี่ยงการติดอยู่ในคำตอบดีที่สุดในบริเวณหนึ่ง (Local Optima) ดังนั้นประโยชน์ที่เกิดขึ้นเมื่อนำมาผสมผสานกัน เมื่อ PSO เริ่มติดอยู่ในคำตอบดีที่สุดในบริเวณหนึ่ง CS จะทำการปรับปรุงการเคลื่อนที่ ทำให้แบบจำลองยังสามารถค้นหาคำตอบที่ดีกว่าได้ เช่นกันกับในส่วนของ PSOFA ซึ่งใช้แรงดึงดูดตามระดับความสว่างของ FA ซึ่งเป็นการปรับแต่งที่ดีในบริเวณที่ใกล้คำตอบ เมื่อนำมาผสมผสานกันจึงได้รับประโยชน์จากการปรับแต่งคุณลักษณะอย่างละเอียด หลังจาก PSO พาเข้าสู่พื้นที่ของคำตอบที่ดีแล้ว และการผสมผสานกันของ CSFA เป็นการเน้นการสำรวจ ในแต่ละด้าน CS ใช้การสุ่มแบบกว้างโดย Lévy Flight และ FA เคลื่อนที่เข้าหาคำตอบที่ดีกว่าในเชิง Local จึงทำให้เกิดความสมดุลในการสำรวจพื้นที่ใหม่ และการปรับคำตอบอย่างต่อเนื่อง ทำให้ประสิทธิภาพคงที่แม้ข้อมูลการฝึกฝนลดลง ด้วยหลักการผสมผสานข้างต้นจึงส่งผลต่อผลลัพธ์ความแม่นยำที่ดีขึ้นของแบบจำลองการจำแนกประเภท



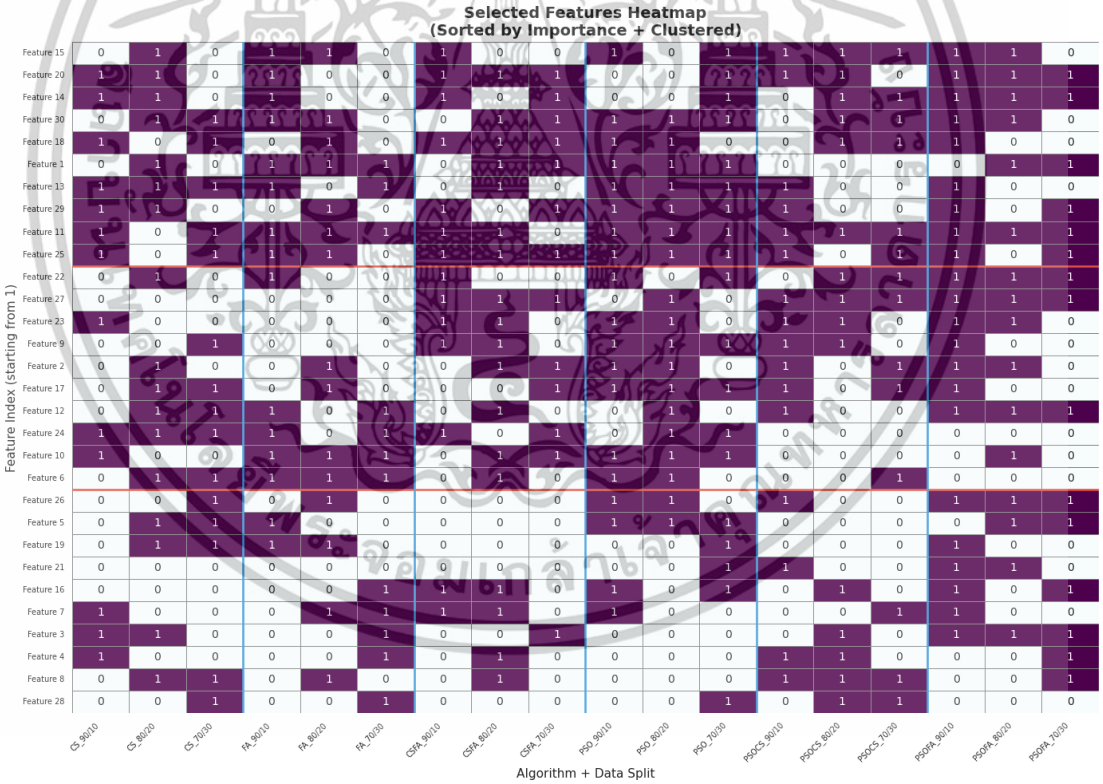
รูปที่ 4.2 จำนวนคุณลักษณะที่ถูกเลือกในชุดข้อมูลมะเร็งเต้านม

ผลลัพธ์จำนวนคุณลักษณะที่ถูกเลือกโดยแต่ละอัลกอริทึมภายใต้การแบ่งชุดข้อมูลที่แตกต่างกัน พบว่าแต่ละอัลกอริทึมมีแนวโน้มการเลือกจำนวนคุณลักษณะที่แตกต่างกันออกไป แสดงดังรูปที่ 4.2 ซึ่งสามารถอธิบายได้จากพฤติกรรมเชิงโครงสร้างของอัลกอริทึมแบบเดี่ยวและแบบผสมผสาน

อัลกอริทึม Cuckoo Search (CS) มีแนวโน้มเลือกคุณลักษณะในปริมาณที่น้อยที่สุดในทุกสัดส่วนของการแบ่งข้อมูล ซึ่งสอดคล้องกับธรรมชาติของอัลกอริทึมที่เน้นการสำรวจเชิงลึกแบบสุ่มโดย Lévy Flight และคัดกรองจุดที่สำคัญ ส่งผลให้อัลกอริทึมมีความสามารถในการลดความซับซ้อนของแบบจำลองได้ดี อัลกอริทึม Particle Swarm Optimization (PSO) เป็นอัลกอริทึมที่เลือกคุณลักษณะในปริมาณสูงอย่างต่อเนื่อง แสดงถึงลักษณะการทำงานที่เน้นการหาจุดที่ดีที่สุดอย่างรวดเร็วผ่านการปรับปรุงตำแหน่งตามฝูง การเลือกจำนวนคุณลักษณะที่มากเกินไปอาจสะท้อนถึง

ความเสี่ยงในการเกิด Overfitting โดยเฉพาะในชุดข้อมูลขนาดเล็ก ในส่วนของอัลกอริทึม Firefly Algorithm (FA) มีจำนวนคุณลักษณะที่ถูกเลือกอยู่ในระดับปานกลาง โดยเฉพาะในสัดส่วน 70/30 ที่ลดลงอย่างชัดเจน ซึ่งอาจจะเป็นผลมาจากกลไกการเคลื่อนที่ตามความสว่างที่เน้นจุดที่มีคุณภาพสูงเท่านั้น ส่งผลให้เกิดการกรองคุณลักษณะที่ไม่จำเป็นได้ดี และความแม่นยำมีความเสถียรภาพ

ในส่วนของอัลกอริทึมแบบผสมผสาน CSFA พบว่ามีพฤติกรรมการเลือกคุณลักษณะที่สมดุลที่สุด โดยเฉพาะในชุดข้อมูล 80/20 ที่เลือกมากที่สุดในกลุ่มของการผสมผสานอัลกอริทึม และลดลงในชุดข้อมูล 70/30 ซึ่งแสดงถึงการผสมผสานระหว่างการสำรวจของ CS และการแสวงหาของ FA ได้อย่างมีประสิทธิภาพ อัลกอริทึมแบบผสมผสาน PSOCS มีแนวโน้มการเลือกคุณลักษณะระดับปานกลาง โดยมีความสม่ำเสมอสูง แสดงถึงการที่ CS เข้ามาช่วยลดพฤติกรรม Over-selection ของ PSO ได้ดี ในส่วนของอัลกอริทึมการผสมผสาน PSOFA มีลักษณะเด่นคือเลือกคุณลักษณะมากที่สุดในชุดข้อมูล 90/10 ซึ่งแสดงถึงความพยายามในการชดเชยความไม่แน่นอนของข้อมูลฝึกที่มีปริมาณน้อยผ่านการขยายพื้นที่ของคุณลักษณะและในชุดข้อมูลขนาดอื่นๆ แสดงถึงการปรับตัวที่ดีขึ้นเมื่อข้อมูลมีความหลากหลาย



รูปที่ 4.3 ภาพ Heatmap คุณลักษณะที่ถูกเลือกในชุดข้อมูลมะเร็งเต้านม

ผลลัพธ์จากการเลือกคุณลักษณะ แสดงดังรูปที่ 4.3 บ่งบอกว่าคุณลักษณะบางรายการได้รับการเลือกอย่างสม่ำเสมอในหลายอัลกอริทึม ซึ่งแสดงถึงความสำคัญเชิงโครงสร้างของคุณลักษณะเหล่านี้ในการจำแนกชุดข้อมูลมะเร็งเต้านม โดยการถูกเลือกอย่างต่อเนื่องสะท้อนให้

เห็นถึงความมั่นคงของตัวแปรที่สามารถใช้เป็นแกนกลางของแบบจำลองในการจำแนกประเภท และในบางคุณลักษณะมีแนวโน้มถูกเลือกเฉพาะบางอัลกอริทึมและบางขนาดของชุดข้อมูลเท่านั้น ซึ่งอาจแสดงถึงลักษณะเฉพาะของอัลกอริทึมที่เน้นการสำรวจในมิติข้อมูลต่างๆ หรืออาจเป็นคุณลักษณะที่มีความสัมพันธ์ต่ำกับผลลัพธ์โดยรวม จึงมีโอกาสที่จะไม่ได้รับเลือกอย่างสม่ำเสมอ

เมื่อพิจารณาพฤติกรรมของแต่ละอัลกอริทึมพบว่าอัลกอริทึม PSO และ PSOFA มีแนวโน้มในการเลือกคุณลักษณะจำนวนมากและซ้ำกัน ซึ่งสอดคล้องกับผลจากการวิเคราะห์จำนวนคุณลักษณะที่ถูกเลือกก่อนหน้า มีลักษณะการกระจายที่หนาแน่นแสดงถึงแนวโน้มการเลือกคุณลักษณะที่มากเกินไปเป็นผลจากการลู่เข้าคำตอบอย่างรวดเร็ว ในส่วนของอัลกอริทึม CS และ FA มีแนวโน้มเลือกคุณลักษณะน้อยลง และแตกต่างกันตามสัดส่วนของข้อมูล ซึ่งแสดงถึงพฤติกรรมการเลือกตัวแปรอย่างระมัดระวัง ความสามารถของทั้งสองอัลกอริทึมนี้คือเลือกคุณลักษณะที่มีนัยสำคัญจริงๆ ช่วยลดความซับซ้อนของแบบจำลองและอาจลดความเสี่ยงต่อการจำเพาะเจาะจงกับข้อมูลฝึกฝนมากเกินไป

สำหรับอัลกอริทึมแบบผสมผสานพบว่ามีการเลือกคุณลักษณะที่ผสมผสานจุดแข็งของอัลกอริทึมต้นแบบ CSFA แสดงความสามารถในการเลือกคุณลักษณะสำคัญอย่างสม่ำเสมอและยังสามารถปรับเปลี่ยนคุณลักษณะตามลักษณะข้อมูลได้ แสดงถึงความยืดหยุ่นและความสมดุล PSOCS มีแนวโน้มเลือกคุณลักษณะน้อยลงเมื่อเปรียบเทียบกับ PSO แสดงถึง ผลการผสมผสานกลไกการสำรวจจาก CS ที่ช่วยลดการเลือกที่มากเกินความจำเป็น ส่วน PSOFA แสดงถึงการปรับตัวตามอัลกอริทึมได้ตามบริบทของข้อมูล สังเกตได้จากจำนวนการเลือกคุณลักษณะที่ลดลงตามสัดส่วนของข้อมูล

4.2 ผลลัพธ์ของการทดลองด้วยชุดข้อมูลภาวะหัวใจขาดเลือด

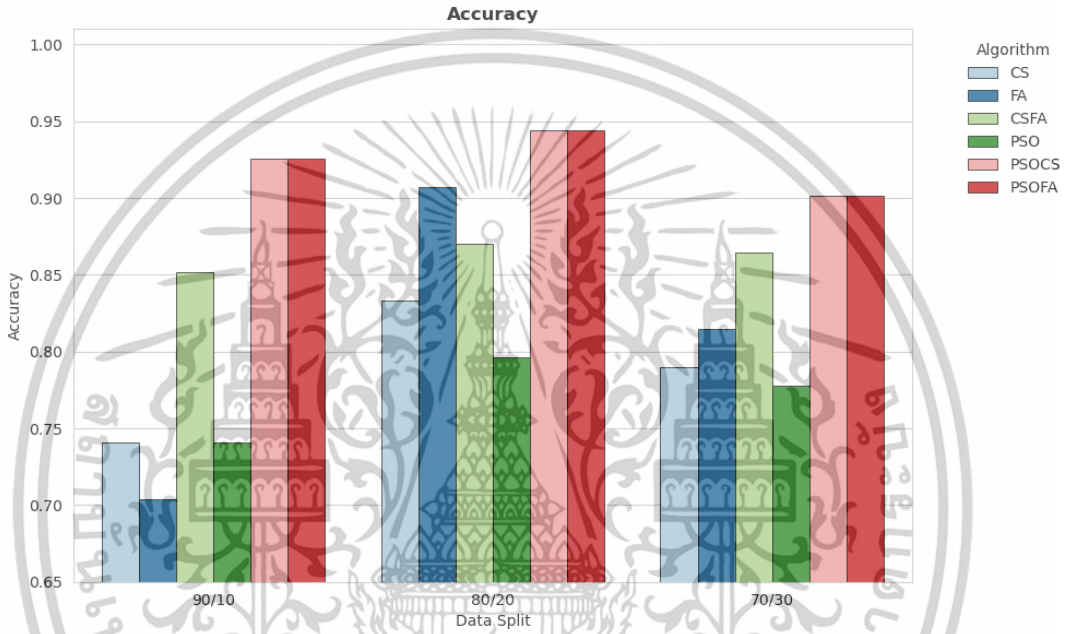
จากผลการทดลองการเลือกคุณลักษณะด้วยอัลกอริทึมทั้ง 6 รูปแบบ บนชุดข้อมูล SPECTF Heart พบว่าอัลกอริทึมแต่ละตัวให้ผลลัพธ์แตกต่างกันอย่างมีนัยสำคัญทั้งในด้านความแม่นยำ ความสมดุลระหว่าง Precision และ Recall ค่าพื้นที่ใต้กราฟ รวมถึงจำนวนคุณลักษณะที่ถูกเลือก ภายใต้การแบ่งข้อมูลเป็น 3 สัดส่วนได้แก่ 90/10, 80/20 และ 70/30 ผลลัพธ์แสดงดังตารางที่ 4.3

ตารางที่ 4.3 ผลลัพธ์การทดลองในชุดข้อมูลภาวะหัวใจขาดเลือด

ขนาดข้อมูล	อัลกอริทึม	Accuracy	Precision	Recall	F1-Score	AUC-ROC	จำนวนคุณลักษณะที่ถูกเลือก
90/10	CS	74.07	80.95	85.00	82.92	63.92	20
	FA	70.37	77.27	85.00	80.95	56.78	10
	PSO	74.07	88.23	75.00	81.08	73.21	26
	CSFA	85.18	90.00	90.00	90.00	80.71	23
	PSOCS	92.59	95.00	95.00	95.00	90.35	22
	PSOFA	92.59	90.90	100	95.23	85.71	22
80/20	CS	83.33	94.87	84.09	89.15	82.04	18
	FA	90.74	93.33	95.45	94.38	82.72	19
	PSO	79.62	90.24	84.09	87.05	72.04	22
	CSFA	87.03	89.36	95.45	92.30	72.72	17
	PSOCS	94.44	93.61	100	96.70	85.00	20
	PSOFA	94.44	97.67	95.45	96.55	92.72	24
70/30	CS	79.01	87.69	86.36	87.02	66.51	20
	FA	81.48	81.48	100	89.70	50.00	6
	PSO	77.77	91.37	80.30	85.48	73.48	26
	CSFA	86.41	93.65	89.39	91.47	81.36	21
	PSOCS	90.12	96.77	90.90	93.74	88.78	26
	PSOFA	90.12	90.27	98.48	94.20	75.90	21

ในสัดส่วนของข้อมูลขนาด 90/10 อัลกอริทึม PSOCS และ PSOFA ให้ค่าความแม่นยำสูงสุดเท่ากันที่ 92.59% โดย PSOFA สามารถให้ค่า Recall 100% และ F1-Score ที่ 95.23 ซึ่งแสดงให้เห็นความสามารถของอัลกอริทึมในการครอบคลุมผู้ป่วยที่เป็นโรคหัวใจได้อย่างครบถ้วน แต่ในส่วนของจำนวนการเลือกคุณลักษณะสูงถึง 22 คุณลักษณะ แต่ FA เลือกคุณลักษณะเพียง 10 คุณลักษณะ และมีค่าความแม่นยำต่ำกว่าสิ่งนี้แสดงให้เห็นว่าการเลือกคุณลักษณะอย่างระวังส่งผลต่อความสามารถในการจำแนกข้อมูล อัลกอริทึม CSFA แสดงผลลัพธ์อยู่ในระดับดี โดยค่าความแม่นยำอยู่ที่ 85.18% และ F1-Score ที่ 90% จำนวนคุณลักษณะที่ถูกเลือก 23 คุณลักษณะ ซึ่งแสดงถึงความสามารถของการผสมผสานกันในการสำรวจและแสวงหาได้อย่างสมดุล

ในข้อมูลสัดส่วน 80/20 อัลกอริทึมที่ให้ประสิทธิภาพโดยรวมดีที่สุดคือ PSOCS และ PSOFA ซึ่งให้ค่าความแม่นยำที่ 94.44% โดย PSOFA ให้ Precision สูงสุดถึง 97.67% และ F1-Score ที่ 96.55 ซึ่งแสดงถึงการจำแนกผู้ป่วยได้แม่นยำและครอบคลุมสูงมาก ขณะที่ FA ให้ F1-Score รองลงมาที่ 94.38 แต่ใช้คุณลักษณะเพียง 19 รายการ แสดงถึงความสามารถในการสร้างแบบจำลองที่มีประสิทธิภาพด้วยความซับซ้อนต่ำ ในทางกลับกัน PSO แม้จะมีค่าความแม่นยำเพียง 79.62% แต่เลือกคุณลักษณะมากถึง 22 รายการ สะท้อนถึงปัญหาการเลือกคุณลักษณะมากเกินไปและความเสี่ยงต่อการจำเพาะเจาะจงกับข้อมูลฝึกมากเกินไป



รูปที่ 4.4 ผลการทดลองความแม่นยำการจำแนกประเภทในชุดข้อมูลภาวะหัวใจขาดเลือด

เมื่อเพิ่มข้อมูลฝึกให้มากขึ้นอัลกอริทึม PSOFA ให้ค่าความแม่นยำสูงสุดที่ 90.12% แสดงดังรูปที่ 4.4 และ Recall สูงถึง 98.48% แสดงถึงศักยภาพในการทำนายข้อมูลที่ดีขึ้นเมื่อมีข้อมูลมากขึ้น ขณะที่ PSOCS มี F1-Score ที่ดีที่สุด (93.74) พร้อมจำนวนคุณลักษณะ 26 คุณลักษณะ ซึ่งอาจแลกมากับความซับซ้อนของแบบจำลอง อัลกอริทึม FA แม้ให้ค่า Recall 100% แต่มีค่าความแม่นยำต่ำเพียง 81.48% และ AUC-ROC ต่ำมากที่ 50.00 บ่งชี้ว่าอาจเกิด Bias จากการเลือกคุณลักษณะเพียง 6 คุณลักษณะ ซึ่งอาจไม่เพียงพอในการสร้างแบบจำลองที่ดีภายใต้ข้อมูลที่หลากหลาย

4.3 ผลลัพธ์ของการทดลองด้วยชุดข้อมูลภาวะหัวใจเต้นผิดจังหวะ

ชุดข้อมูล Arrhythmia เป็นชุดข้อมูลที่มีจำนวนคุณลักษณะมากเมื่อเทียบกับจำนวนข้อมูล (High-dimensional) และมีความไม่สมดุลของคลาส ซึ่งส่งผลโดยตรงต่อความสามารถในการจำแนกของแต่ละอัลกอริทึม โดยจากผลการทดลองพบว่าค่าความแม่นยำโดยรวมลดลงอย่างชัดเจนเมื่อเพิ่ม

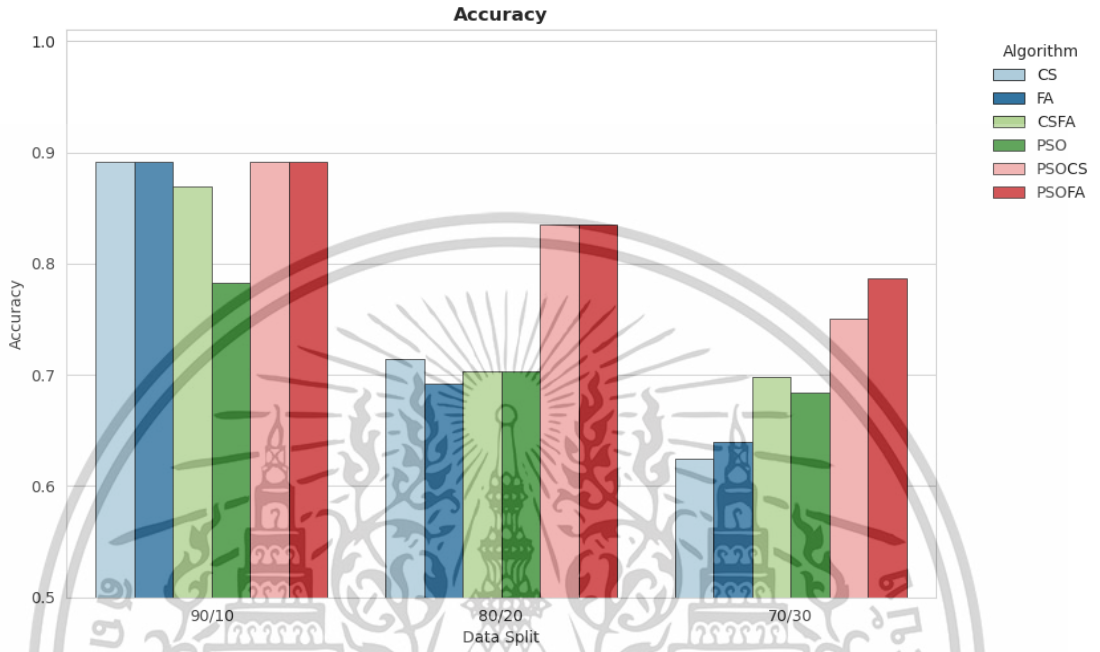
อัตราส่วนของข้อมูลทดสอบ อัลกอริทึม PSOCS และ PSOFA ให้ผลลัพธ์ดีกว่าอัลกอริทึมอื่นในหลายกรณี ทั้งในด้านความแม่นยำและ F1-Score ผลลัพธ์แสดงดังตารางที่ 4.4

ตารางที่ 4.4 ผลลัพธ์การทดลองในชุดข้อมูลภาวะหัวใจเต้นผิดจังหวะ

ขนาดข้อมูล	อัลกอริทึม	Accuracy	Precision	Recall	F1-Score	จำนวนคุณลักษณะที่ถูกเลือก
90/10	CS	89.13	73.75	77.50	74.57	136
	FA	89.13	73.75	70.83	71.90	145
	PSO	78.26	55.14	54.78	54.77	138
	CSFA	86.95	65.97	70.47	66.13	140
	PSOCS	89.13	76.81	80.47	77.64	134
	PSOFA	89.13	76.70	72.97	73.04	143
80/20	CS	71.42	63.32	50.36	52.87	130
	FA	69.23	39.58	36.90	37.08	60
	PSO	70.32	47.96	48.29	47.60	138
	CSFA	70.32	52.42	46.78	47.99	126
	PSOCS	83.51	69.75	58.76	62.00	147
	PSOFA	83.51	62.92	57.87	59.58	143
70/30	CS	62.50	44.57	40.75	41.79	121
	FA	63.97	52.03	46.48	44.88	75
	PSO	68.38	50.25	45.27	47.18	127
	CSFA	69.85	57.30	50.49	52.79	129
	PSOCS	75.00	60.40	53.49	52.39	150
	PSOFA	78.67	61.32	55.03	54.19	127

ในชุดข้อมูลขนาด 90/10 อัลกอริทึม CS, FA, PSOCS, PSOFA ให้ค่าความแม่นยำเท่ากันที่ 89.13% โดย PSOCS ให้ Recall และ F1-Score สูงสุด ส่วนอัลกอริทึม PSO มีประสิทธิภาพต่ำที่สุดในทุกตัวชี้วัดแม้เลือกคุณลักษณะมาก ดังนั้นจึงแสดงให้เห็นว่ามีปัญหาในความจำเพาะมากเกินไปของข้อมูลฝึกฝน หรือว่าปัญหาการเลือกคุณลักษณะมากเกินไป ในขณะเดียวกัน CSFA ให้ค่าความแม่นยำสูง 86.95% แต่ F1-Score ต่ำที่ 66.13% ซึ่งอาจเกิดการ Bias ต่อบางคลาส

ในชุดข้อมูลขนาด 80/20 ค่าความแม่นยำโดยรวมลดลงทุกอัลกอริทึม แสดงดังรูปที่ 4.5 โดยเฉพาะ FA เหลือเพียง 69.23% และ F1-Score อยู่ที่ 37.08% ในส่วนของอัลกอริทึมแบบผสมผสาน PSOCS และ PSOFA ให้ค่าความแม่นยำที่ 83.51% พร้อม Recall ที่ดี แต่อัลกอริทึม CSFA เริ่มสูญเสียประสิทธิภาพ โดยค่า F1-Score อยู่ที่ 47.99%



รูปที่ 4.5 ผลการทดลองความแม่นยำการจำแนกประเภทในชุดข้อมูลภาวะหัวใจเต้นผิดจังหวะ

โดยในขนาดชุดข้อมูล 70/30 ผลลัพธ์โดยรวมแย่ลงกว่าทุกขนาดชุดข้อมูล โดยเฉพาะ CS และ PSO ต่ำกว่า 63% อัลกอริทึม FA แม้ใช้คุณลักษณะน้อย 75 คุณลักษณะแต่ค่า Recall 46.48% อยู่ในระดับต่ำ และ F1-Score ต่ำกว่า 45% เสี่ยงต่อสภาวะแบบจำลองไม่สามารถเรียนรู้จากข้อมูลฝึกได้เพียงพอ ในส่วนของ PSOFA ให้ค่าความแม่นยำที่ 78.67% และ Recall ที่ 55.03% ถือว่าอยู่ในระดับที่ดีที่สุดของชุดข้อมูลนี้

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของอัลกอริทึมการเลือกคุณลักษณะด้วยแนวทาง Swarm Intelligence ทั้งในรูปแบบเดี่ยว ได้แก่ CS, FA และ PSO และแบบผสมผสาน ได้แก่ CSFA, PSOCS และ PSOFA โดยใช้ Support Vector Machine (SVM) เป็นตัวจำแนก เพื่อประเมินความแม่นยำในการจำแนกข้อมูลทางการแพทย์ภายใต้ชุดข้อมูลที่มีลักษณะแตกต่างกัน ได้แก่ ชุดข้อมูลมะเร็งเต้านม (WBCD), SPECTF Heart และ Arrhythmia โดยเน้นการวิเคราะห์เชิงลึกจากชุดข้อมูลมะเร็งเต้านมเป็นหลัก และใช้ชุดข้อมูลอีกสองชุดเพื่อขยายผลในเชิงเปรียบเทียบ

จากผลการทดลองพบว่าอัลกอริทึมแบบ Hybrid ได้แก่ PSOFA และ PSOCS ให้ผลลัพธ์ที่โดดเด่นทั้งในด้านความแม่นยำ F1-Score และ AUC-ROC โดยเฉพาะในสัดส่วนข้อมูล 70/30 ซึ่ง PSOFA ให้ค่าความแม่นยำสูงถึง 90.12% และ Recall สูงถึง 98.48% สะท้อนถึงความสามารถในการครอบคลุมผู้ป่วยที่เป็นมะเร็งได้อย่างแม่นยำ ขณะที่ PSOCS มีค่า F1-Score สูงที่สุดที่ 93.74% แสดงถึงความสมดุลระหว่าง Precision และ Recall ได้อย่างมีประสิทธิภาพ

ในด้านจำนวนคุณลักษณะที่ถูกเลือก พบว่า PSO และ PSOCS มีแนวโน้มเลือกคุณลักษณะมากกว่ากลุ่มอื่น ๆ (สูงถึง 26 ตัวแปร) ขณะที่ CS และ FA มักเลือกจำนวนน้อยกว่า ซึ่งชี้ให้เห็นถึงความเสี่ยงในการเลือกคุณลักษณะมากเกินไปของบางอัลกอริทึม ส่งผลต่อความซับซ้อนของแบบจำลองและโอกาสในการเกิดความจำเพาะกับข้อมูลฝึกฝน จากการทดลองแสดงให้เห็นว่ามี ชุดของคุณลักษณะหลักที่ถูกเลือกอย่างสม่ำเสมอในทุกอัลกอริทึม เช่น คุณลักษณะที่ 14, 15, 20 และ 30 ซึ่งสะท้อนว่าคุณลักษณะเหล่านี้มีความสำคัญร่วมกันในการจำแนกกลุ่มโรค และสามารถนำไปใช้เป็น Feature Set พื้นฐานสำหรับการใช้งานได้

ในขณะเดียวกันเมื่อนำอัลกอริทึมเหล่านี้ทดลองกับชุดทางการแพทย์อื่น เช่น ชุดข้อมูลทำนายภาวะหัวใจขาดเลือด ผลลัพธ์แสดงให้เห็นว่าอัลกอริทึมผสมผสานอย่าง PSOCS และ PSOFA ยังคงแสดงประสิทธิภาพได้ดีโดยมีค่าความแม่นยำสูงสุดถึง 94.44% และ F1-Score สูงถึง 96.55% ในสัดส่วนข้อมูล 80/20 ขณะที่ FA มีแนวโน้มให้ผลลัพธ์ใกล้เคียงกับอัลกอริทึมผสมผสานแม้จะเลือกคุณลักษณะน้อยกว่า ซึ่งแสดงให้เห็นถึงความสามารถของอัลกอริทึมแบบเดี่ยวในกรณีที่ข้อมูลมีลักษณะเฉพาะทางและจำนวนคุณลักษณะไม่มาก และในส่วนของชุดข้อมูลการวินิจฉัยภาวะหัวใจเต้นผิดจังหวะที่มีลักษณะมิติสูง และมีความไม่สมดุลจของคลาสอย่างชัดเจน ส่งผลให้ผลการจำแนกของหลายอัลกอริทึมลดลง โดยเฉพาะ PSO ซึ่งมักให้ F1-Score ต่ำกว่า 50% ในหลายสัดส่วน ขณะที่อัลกอริทึมแบบผสมผสานโดยเฉพาะ PSOFA และ PSOCS ยังคงสามารถรักษาค่าความแม่นยำและ

F1-Score ให้อยู่ในระดับสูงกว่ากลุ่มอื่นๆ ได้ แม้จะต้องแลกมากับจำนวนคุณลักษณะที่เลือกสูงถึง 140–150 รายการ

อัลกอริทึมแบบผสมผสานโดยเฉพาะ PSOFA และ PSOCS แสดงประสิทธิภาพที่โดดเด่นและสม่ำเสมอในทุกชุดข้อมูล โดยเฉพาะในปัญหาที่มีความซับซ้อน เช่นความไม่สมดุลของคลาส หรือข้อมูลที่มีมิติสูง ขณะที่อัลกอริทึมแบบเดี่ยว เช่น FA และ CS แม้จะมีจุดแข็งในเรื่องการลดความซับซ้อนของแบบจำลอง แต่มีความเสี่ยงต่อการไม่เรียนรู้เมื่อข้อมูลมีความหลากหลายหรือไม่สมดุล ผลจากงานวิจัยนี้ชี้ให้เห็นถึงความสำคัญของการเลือกอัลกอริทึมให้เหมาะสมกับลักษณะของข้อมูล และสนับสนุนแนวทางการพัฒนา อัลกอริทึมแบบผสมผสานสำหรับการเลือกคุณลักษณะ เพื่อเพิ่มประสิทธิภาพของระบบวินิจฉัยในงานข้อมูลทางการแพทย์

5.2 ข้อเสนอแนะ

จากผลลัพธ์ที่ได้ในงานวิจัยนี้ยังคงมีข้อจำกัดในการประเมินประสิทธิภาพของอัลกอริทึม ซึ่งในครั้งนี้อ้างอิงผลลัพธ์จากเพียงชุดข้อมูลมะเร็งเต้านม ชุดข้อมูลภาวะหัวใจขาดเลือด และชุดข้อมูลภาวะหัวใจเต้นผิดจังหวะกับกลุ่มของอัลกอริทึมเพียงบางตัวเท่านั้น ทำให้ยังไม่สามารถสรุปได้อย่างครอบคลุมว่าอัลกอริทึมเหล่านี้สามารถประยุกต์ใช้ได้กับข้อมูลชนิดอื่นหรือในสภาพแวดล้อมที่แตกต่างกันได้หรือไม่ จึงจำเป็นต้องมีการทดลองเพิ่มเติมกับชุดข้อมูลที่หลากหลายขึ้น ทั้งในแง่ของความซับซ้อน ปริมาณข้อมูล และลักษณะของปัญหา รวมถึงการศึกษาเชิงลึกว่าอัลกอริทึมสามารถเรียนรู้และปรับตัวต่อข้อมูลใหม่ได้ดีเพียงใด ซึ่งจะช่วยยืนยันถึงความสามารถในการใช้งานในสถานการณ์จริงในอนาคต

ขณะเดียวกันข้อมูลที่ได้จากการวิจัยชี้ให้เห็นว่า ความสามารถของแบบจำลองไม่ควรพิจารณาจากค่าความแม่นยำเพียงอย่างเดียว แต่ควรประเมินร่วมกับตัวชี้วัดอื่น เช่น F1-Score, AUC-ROC และ Recall โดยเฉพาะในกรณีของข้อมูลที่มีปัญหาความไม่สมดุลของคลาส ร่วมถึงการเลือกใช้อัลกอริทึมควรพิจารณาตามลักษณะของข้อมูลและข้อจำกัดด้านการใช้งานจริง เช่น ความเร็วในการประมวลผล หรือความสามารถในการตีความของแบบจำลอง โดยเฉพาะในบริบทของงานทางการแพทย์

สำหรับงานวิจัยในอนาคต ควรมีการศึกษาประยุกต์ใช้อัลกอริทึมแบบผสมผสานร่วมกับข้อมูลที่มีโครงสร้างซับซ้อนมากขึ้น หรือข้อมูลที่ไม่เป็นเชิงโครงสร้าง เช่น ข้อมูลภาพทางการแพทย์ หรือข้อมูลเวชระเบียนอิเล็กทรอนิกส์ รวมถึงการออกแบบกลไกการเรียนรู้แบบต่อเนื่องเพื่อให้อัลกอริทึมสามารถปรับตัวได้ในสถาพแวดล้อมของข้อมูลที่เปลี่ยนแปลง เพื่อสนับสนุนการนำแบบจำลองไปประยุกต์ใช้ในระบบวินิจฉัยหรือระบบช่วยตัดสินใจทางคลินิกได้ในอนาคต

เอกสารอ้างอิง

[1] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157–1182.

[2] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Systems, 86, 33–45.

[3] Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5(Oct), 1205–1224.

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

[5] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In Data Classification: Algorithms and Applications (pp. 37–64). CRC Press.

[6] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Systems, 86, 33–45.

[7] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507–2517.

[8] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28.

[9] Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. Artificial Intelligence, 151(1–2), 155–176.

[10] Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation, 20(4), 606–626.

[11] Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. In Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1200–1205). IEEE.

[12] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28.

เอกสารอ้างอิง (ต่อ)

[13] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In Proceedings of ICNN'95 - International Conference on Neural Networks (Vol. 4, pp. 1942–1948). IEEE.

[14] Yang, X.-S., & Deb, S. (2009). Cuckoo search via Lévy flights. In 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC) (pp. 210–214). IEEE. [DOI: 10.1109/NABIC.2009.5393690]

[15] Yang, X. S. (2010). Firefly algorithms for multimodal optimization. In International symposium on stochastic algorithms (pp. 169–178). Springer.

[16] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In Proceedings of ICNN'95 - International Conference on Neural Networks (Vol. 4, pp. 1942–1948). IEEE.

[17] Yang, X. S. (2010). Nature-inspired metaheuristic algorithms (2nd ed.). Luniver Press.

[18] Wang, G. G., Deb, S., & Cui, Z. (2015). A hybrid metaheuristic approach for feature selection. Neurocomputing, 149, 188–199.

[19] Al-Betar, M. A., et al. (2020). Natural metaheuristics and hybridization strategies for feature selection: A review. Information Fusion, 52, 90–111.

[20] Elaziz, M. A., et al. (2020). Improved binary whale optimization algorithm for feature selection using SVM classifier. Computers & Electrical Engineering, 87, 106776.

[21] Zawbaa, H. M., Emary, E., & Grosan, C. (2016). Feature selection via chaotic antlion optimization. PLoS ONE, 11(3), e0150652

[22] Azzalini, M., & Mousavi, S. M. (2017). A hybrid firefly algorithm for feature selection. Expert Systems with Applications, 80, 67–74.

[23] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[24] Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2010). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

เอกสารอ้างอิง (ต่อ)

[25] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.

[26] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

[27] Chuang, L. Y., Yang, C. H., & Li, J. C. (2008). Chaotic maps based genetic algorithm for feature selection. Information Sciences, 179(20), 3179–3191.

[28] Xue, B., Zhang, M., & Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. IEEE Transactions on Cybernetics, 43(6), 1656–1671.

[29] Gherboudj, A. S., Benchabane, F., & Settouti, N. (2018). Feature selection using cuckoo search: Application to medical data. Computational Biology and Chemistry, 75, 24–32.

[30] Nandy, S., Sinha, D., & Das, A. (2020). Firefly algorithm based feature selection in health data analysis. Procedia Computer Science, 167, 2290–2299.

[31] Alshamlan, H. M., Badr, G. H., & Alohali, Y. A. (2015). A PSO-based approach for gene selection in microarray data. BioMed Research International, 2015, Article ID 423149.

[32] Wang, G. G., Deb, S., & Cui, Z. (2015). A hybrid metaheuristic approach for feature selection. Neurocomputing, 149, 188–199.

[33] Al-Betar, M. A., et al. (2020). Natural metaheuristics and hybridization strategies for feature selection: A review. Information Fusion, 52, 90–111.

[34] Azzalini, M., & Mousavi, S. M. (2017). A hybrid firefly algorithm for feature selection. Expert Systems with Applications, 80, 67–74.

[35] Naik, B., & Rautray, R. (2020). An enhanced hybrid PSO-FA for feature selection in medical diagnosis. Biomedical Signal Processing and Control, 59, 101921.

ภาคผนวก

ภาคผนวก ก

ผลงานวิจัยที่ตีพิมพ์

Feature Selection Method Based on Hybrid Cuckoo Search and Firefly Algorithm for Breast Cancer Prediction

Chalanwich Teerasarn
Dept. of Computer Science, School of Science
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
63605108@kmitl.ac.th

Warangkhana Kimpan
Dept. of Computer Science, School of Science
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
warangkhana.ki@kmitl.ac.th

2025 4th Asia Conference on Algorithms, Computing and Machine Learning (CACML) | 979-8-3315-4314-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/CACML64929.2025.11010928

***Abstract*—This research focuses on developing an efficient feature selection process using a hybrid technique combining Cuckoo Search algorithm and Firefly Algorithm. The Wisconsin Diagnostic Breast Cancer dataset is utilized to evaluate the capability of selecting significant features and eliminating irrelevant ones. The experimental results demonstrate that using the hybrid technique significantly improves the accuracy of machine learning compared to using the Cuckoo Search and Firefly Algorithm individually. Additionally, an analysis was conducted on the impact of splitting the dataset for training and testing, with splits of 70/30, 80/20, and 90/10. The experiments revealed a relationship between the number of selected features and the model accuracy. The findings from this study can serve as a guideline for developing appropriate feature selection process for complex data analysis problems.**

Keywords—Feature Selection, Swarm Intelligence, Cuckoo Search, Firefly Algorithm, Breast Cancer

I. Introduction

Machine learning is an essential tool in many domains, including engineering, finance, and medicine, especially in the interpretation of medical data. Feature selection is considered an essential step that helps reduce data complexity, increase processing speed, and improve the performance of learning models. Selecting appropriate features enables the model to focus on learning only the data directly relevant to the problem being analyzed.

Metaheuristic algorithms are a group of algorithms designed to solve complex optimization problems including Simulated Annealing (SA), Tabu Search (TS), Genetic Algorithm (GA), Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Differential Evolution (DE), and others. Cuckoo Search (CS) and Firefly Algorithm (FA) are also interesting bio-inspired algorithms which have distinct strengths. CS excels in exploring the search space broadly but lacks deep refinement capabilities, while FA specializes in fine-tuning solutions but has limited exploration capabilities. This led to the idea of hybridizing the two algorithms to balance exploration and exploitation effectively.

Therefore, this research aims to develop and evaluate the performance of a hybrid technique combining the Cuckoo Search and Firefly Algorithm for feature selection. The Wisconsin Diagnostic Breast Cancer dataset is used to analyze the performance and the impact of data splitting with proportions of 70/30, 80/20, and 90/10. The experimental results will confirm the capability of the hybrid technique in improving the accuracy of learning models.

II. Related Work

The research in feature selection has developed various techniques and algorithms to reduce unnecessary features and enhance the performance of learning models. Many studies have explored metaheuristic algorithms as well as hybrid approaches to improve outcomes. This section will cover key algorithms and relevant approaches.

Feature selection is a crucial step in the data analysis process, aiming to reduce unnecessary features and retain only those that are significant. Research by H. Xie, L. Zhang, C.P. Lim, Y. Yu, and H. Liu [1] utilized Particle Swarm Optimization (PSO) for feature selection and demonstrated that the selected features could improve accuracy. The study by C. Ozgur [2] presented a Genetic Algorithm (GA) for feature selection in large datasets, highlighting the potential of metaheuristic algorithms in handling complex data.

A. Cuckoo Search

Cuckoo Search (CS) is an algorithm inspired by the egg-laying behavior of cuckoo birds, using Lévy Flights to find optimal solutions within the search space. The CS algorithm was developed and demonstrated its capability in solving optimization problems in complex scenarios [3].

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Le\acute{v}y(\lambda) \tag{1}$$

Equation (1) describes the process of generating new solution in the Cuckoo Search by employing the concept of Lévy Flights, which is a random walk where the step length is determined by a Lévy distribution. In this equation, $x_i^{(t)}$ represents the current position of the solution for cuckoo i at iteration t, while $x_i^{(t+1)}$ indicates the updated position after the adjustment. This update depends on the parameter α, which

controls the step size to balance between exploring new areas (exploration) and refining solution in nearby regions (exploitation). Additionally, Lévy(λ) the step length drawn from a Lévy distribution, which ensures an effective balance between local and global search in the optimization process.

$$Lévy \sim u = t^{-\lambda}, 1 < \lambda \leq 3 \quad (2)$$

Equation (2) defines the Lévy distribution, a heavy-tailed distribution where most steps are short, but occasional long steps occur. The step length (u) derived from this distribution is used in Equation (1) to determine the distance for updating solutions. The parameter λ, within the range $1 < \lambda \leq 3$, shapes the distribution: when λ is close to 1, it emphasizes long jumps for exploring new areas (exploration), whereas when λ is close to 3, it favors short steps for refining solutions in the local region (exploitation). The power law from $t^{-\lambda}$ ensures the algorithm can effectively balance local and global search, improving optimization efficiency.

The study by A. Joshi and R. Aziz [4] proposed a hybrid method combining Cuckoo Search and Spider Monkey Optimization for feature selection to identify gene sets that assist in predicting early-state cancer. This method demonstrated its effectiveness in improving classification accuracy and cancer prediction from large datasets. B. Aljorani and A. Hasan [5] introduced a hybrid method between Cuckoo Search and Crossover Operators to explore unvisited areas and avoid being trapped in local optima. The experiments showed that the developed algorithm achieved higher efficiency in terms of classification accuracy and the number of selected features compared to the traditional CS method. B. Elizabeth et al. [6] used the Chi-Square technique to rank features in a heart disease dataset and employed the Cuckoo Search Optimization algorithm to select the most relevant features. The selected features were then used to train models with k-nearest neighbor (KNN) and Support Vector Machine (SVM). The experimental results demonstrated that the proposed method could effectively reduce features and improve the heart disease diagnostic system. Y. Kaya [7] proposed the Cuckoo Search algorithm for feature selection outperformed the Genetic Algorithm (GA) by selecting fewer features (25 vs. 27) while achieving a higher classification accuracy (99.04% vs. 96.15%) on the Sonar dataset. H. Abdulwahab, S. Ajitha, M. Saif, B. Murshed, and F. Ghanem [8] proposed the Multi-Objective Binary Cuckoo Search Algorithm (MOBCSA) effectively selected optimal gene subsets for classification tasks in bioinformatics, achieving high classification accuracy (92.79% to 98.42%) while reducing the number of selected genes (15.67 to 27.88), outperforming other multi-objective feature selection methods. The Cuckoo Search algorithm is described in Fig 1.

Algorithm 1 Cuckoo Search via Lévy Flights

```
begin
Objective function f(x), x = (x_1, ..., x_d)^T
Generate initial population of n host nests x_i (i = 1, 2, ..., n)
while t < MaxGeneration or (stop criterion) do
    Get a cuckoo randomly by Lévy flights
    Evaluate its quality/fitness F_i
    Choose a nest among n (say, j) randomly
    if F_i > F_j then
        Replace j by the new solution
    end if
    A fraction (p_a) of worse nests are abandoned and new ones are built
    Keep the best solutions (or nests with quality solutions)
    Rank the solutions and find the current best
end while
Postprocess results and visualization
end
```

Fig. 1. Pseudocode of the Cuckoo Search (CS)

B. *Firefly Algorithm*

Firefly Algorithm (FA) is inspired by the flashing behavior of fireflies, which is used for attraction and communication. Research by Yang (2010) demonstrated the efficiency of FA in finding optimal solutions for nonlinear problems [9].

$$I(r) = I_0 e^{-\gamma r^2} \quad (3)$$

Equation (3) describes the light intensity, which represents the quality of a solution in the search space. The light intensity (I) decreases with the distance (r) between two fireflies, where I_0 is the initial intensity at the source, and γ is the absorption coefficient that governs the rate of intensity reduction as the distance increases.

$$\beta = \beta_0 e^{-\gamma r^2} \quad (4)$$

Equation (4) represents the attractiveness of a firefly (β), which decreases with the distance (r) between two fireflies. The initial attractiveness at the same position ($r = 0$) is denoted by β_0 and γ is the absorption coefficient that determines the rate at which the attractiveness diminishes as the distance increases.

$$r_{ij} = \sqrt{\sum_{k=1}^{d} \left(x_{i,k} - x_{j,k}\right)^2} \quad (5)$$

Equation (5) represents the distance between fireflies i and j, calculated using the Euclidean distance (r_{ij}), where $x_{i,k}$ denotes the position of firefly i in dimension k and d is the number of dimensions the search space. The movement of firefly i, which is attracted toward firefly j (with higher brightness) represents in equation (6).

$$x_i = x_i + \beta e^{-\gamma r_{ij}^2}\left(x_j - x_i\right) + \alpha(rand - 0.5) \quad (6)$$

Where β is the attraction coefficient the decreases with distance, α is the randomization factor to enhance diversity in the search space process, and $rand$ is a random value from the range [0,1] to add stochasticity to the movement direction. The Firefly Algorithm is described in Fig 2.

Algorithm 2 Firefly Algorithm

Objective function: $f(X), X = (x_1, x_2, \cdots, x_d)^T$
Generate initial population of fireflies X_i $(i = 1, 2, \cdots, n)$
Evaluate light intensity I_i at X_i using $f(X_i)$
Define light absorption coefficient γ
while $t <$ MaxGeneration **do**
  for $i = 1$ to n **do**
    for $j = 1$ to n **do**
      if $I_j > I_i$ **then**
        Move firefly i towards firefly j in d-dimension
        Attractiveness varies with distance r via $\exp(-\gamma r^2)$
        Evaluate new solutions and update light intensity
      end if
    end for
  end for
  Rank the fireflies and find the current best
end while
Postprocess results and visualization

Fig. 2. Pseudocode of the Firefly Algorithm (FA)

In applying FA to feature selection, the study by T. Badriyah, I. Syarif, and D. Prakoso [10] proposed using the Firefly Algorithm for feature selection in high-dimensional datasets. S. Chemmalar, G. Selvi, Y. Gokul, G. Srivastava, and T. Gagekallu [11] used the Firefly Algorithm as a technique for selecting the most important features for breast cancer diagnosis. The results showed that this method is highly effective. The study by P. Rana, I. Batra, and A. Malik [12] proposed using the FA, particularly in its hybrid form, exhibited a remarkable ability to enhance feature selection by significantly improving classification accuracy and efficiently reducing dimensionality. This optimization technique outperformed conventional methods such as PSO and other widely used algorithms in various domains. Furthermore, T. Sindhu, N. Kumaratharan, and P. Anandan [13] using the FA proved to be an effective optimization tool for selecting relevant features in high-dimensional datasets, leading to substantial improvements in classification accuracy while outperforming other traditional methods, including PSO and GA, due to its superior exploration-exploitation balance, adaptability, and robustness in handling complex data structures across various real-world applications. Additionally, S. Maza and D. Zouache [14] proposed the Binary Firefly Algorithm (BFA) demonstrated its superior capability in feature selection by effectively eliminating redundant features while ensuring high classification accuracy. Experimental results confirmed that BFA consistently outperformed PSO across multiple datasets, reinforcing its robustness and adaptability for feature selection tasks.

The integration of multiple algorithms has become a popular approach in recent research. For example, Q. Chen, Y. Chen, and W. Jiang [15] combined PSO and GA to enhance the efficiency of feature selection. B. Asgarali, G. Habib, and T. Omid [16] integrated Cuckoo Search and Differential Evolution to develop a highly effective hybrid technique.

In order to enhance and balance the exploration and exploitation of swarms effectively, we proposed the idea of hybridizing the two algorithms: Cuckoo Search and Firefly Algorithm, called CSFA that can improve accuracy and reduce unnecessary features.

TABLE I. CLASSES

Classes	Instances	Percents
Malignant	212	37.3
Benign	357	62.7

III. THE HYBRID CUCKOO SEARCH AND FIREFLY ALGORITHM IN FEATURE SELECTION

This section provides details about the dataset, feature selection techniques and models used for breast cancer classification, which are key components of this research. The primary objective is to select only the most relevant features to reduce the number of features, minimize the risk of overfitting, and enhance the performance of the classification model in terms of both accuracy and efficient data processing capabilities. The architecture of the proposed hybrid method, called CSFA as shown in Fig 3.

A. Dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is widely recognized for its use in machine learning and big data management. It is available for download from the UCI Machine Learning Repository. The primary purpose of this dataset is to diagnose breast cancer based on clinical features derived from measurements of cell nucleus characteristics.

This dataset consists of 560 instances, 30 features, and 2 target classes. The data comprises of 30 numerical values obtained from digital images of cell samples collected through the Fine Needle Aspirate (FNA) process. The main features include: the mean distance from the center to point on the perimeter (Perimeter), area (Area), the variance of radial lengths in a specific area (Smoothness), compactness (perimeter^2 / area - 1.0), the severity of concave portions of the contour (Concavity), the number of concave portions of the contour (Concave Points), symmetry (Symmetry), and the fractal dimension (coastline approximation - 1). These features are calculated using three metrics: mean, standard error, and mean of the three largest values (Worst). This dataset is divided into 212 malignant cases (37.3%) and 357 benign cases (62.7%) as shown in Table 1.

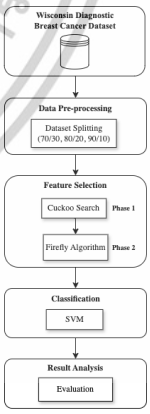


Fig. 3. Architecture of proposed hybrid method

B. Feature Selection Techniques

Feature selection is a crucial step in the machine learning process, aiming to select only the significant and relevant feature from the dataset while eliminating unnecessary or complex ones. This method enhances the model's performance, reduces the risk of overfitting and simplifies processing complexity to achieve more accurate and efficient results.

Fig 4 shows the integration of the Cuckoo Search (CS) and Firefly Algorithm (FA), called CSFA for feature selection leverages the strengths of both algorithms. The CS performs a comprehensive global search to identify high-quality initial feature sets, while FA refines these feature sets in depth to achieve the optimal results. This hybrid method clearly divides the process into two stages to maximize the efficiency of search and refinement.

In Phase 1, a global search is conducted using the CS algorithm, which is designed to explore the search space comprehensively. Lévy Flight is employed as a key mechanism. The process begins by generating a population of feature sets in binary encoding. It is then used to search and explore areas that may contain optimal solutions. These feature sets are evaluated using a fitness function, which considers classification accuracy and feature reduction to enhance the compactness of the model. Finally, the best solution from CS is passed to Phase 2 for further refinement.

In Phase 2, the FA uses the best solution from CS as the starting point for refinement. Each firefly represents a feature set and the movement of a dimmer firefly toward a brighter one is based on the intensity of the light, represents the quality of the fitness function and the distance, which reduces the importance of farther options. This attraction and movement principle helps improve the search for solutions. Additionally, randomness in the movement process enhances diversity in the search space (exploration). Finally, the firefly with the highest fitness in the final generation represents the optimal feature set.

C. Classification

Once the optimal feature set is obtained, Support Vector Machine or SVM is used as the classifier. SVM is chosen due to its ability to handle high-dimensional data and its effectiveness in binary classification problems. Additionally, the kernel and parameters of the SVM are fine-tuned to achieve the best results.

TABLE II. PARAMETER SETTINGS OF ALL THE OPTIMIZATION ALGORITHMS

Algorithm	Parameter
CS	Population Size (n) = 50, Iterations (T) = 200, Discovery Rate of Alien Eggs (p_a) = 0.1, Step Size (α) = 0.1, Levy Flight Distribution Parameter (λ) = 1.3
FA	Population Size (n) = 50, Iterations (T) = 200, Light Absorption Coefficient (γ) = 1.618, Attractiveness Coefficient ($\beta_{\min}$) = 1, Randomness Factor (α) = 1
CSFA	Population Size (n) = 50, Iterations (T) = 200, Discovery Rate of Alien Eggs (p_a) = 0.1, Step Size (α) = 0.1, Levy Flight Distribution Parameter (λ) = 1.3, Light Absorption Coefficient (γ) = 1.618, Attractiveness Coefficient ($\beta_{\min}$) = 1, Randomness Factor (α) = 1

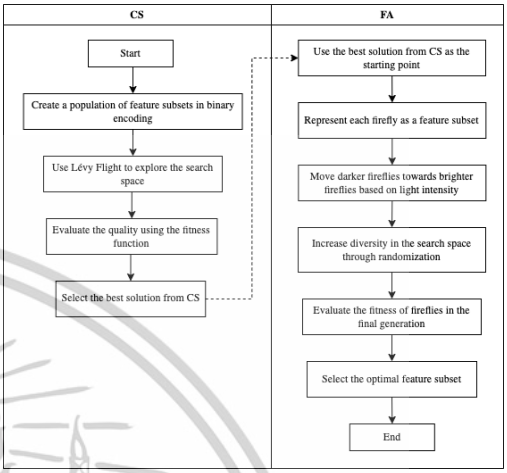


Fig. 4. Workflow diagram of feature selection using CSFA method

IV. Experimental Results

The experimental results demonstrate the effectiveness of the hybrid feature selection method combining Cuckoo Search and Firefly Algorithm (CSFA) using the WDBC dataset. The evaluation focuses on classification accuracy, the distribution of selected features and performance comparisons between dataset splitting methods and feature selection techniques.

Table 2 shows the parameter settings for each algorithm in the experiment, ensuring a fair performance evaluation by maintaining the same iteration count and population size. Table 3 shows the number of selected features for each algorithm across different data splitting ratios. Notably, CSFA selected the highest number of features in the 70/30 split, indicating its effectiveness in conducting a comprehensive search.

Table 4 shows that the hybrid algorithm of CSFA demonstrates the highest efficiency in feature selection and data classification. Compared to using CS and FA individually, the hybrid method converges to the solution more quickly and exhibits significantly higher stability as shown in Fig 5.

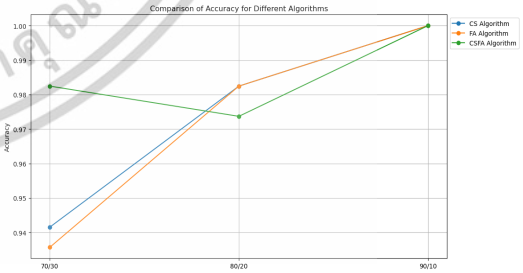


Fig. 5. A comparison chart of accuracy values across different algorithms

TABLE III. THE NUMBER OF SELECTED FEATURES OF ALL THE OPTIMIZATION ALGORITHMS

Algorithm	Data splitting	Number of selected features
CS	70/30	15
	80/20	15
	90/10	12
FA	70/30	9
	80/20	11
	90/10	17
CSFA	70/30	18
	80/20	13
	90/10	14

In term of model accuracy, the experimental results show that using the hybrid technique (CSFA) achieves higher accuracy compared to using CS and FA individually, especially with datasets split into 70/30 and 80/20. This demonstrates its ability to handle complex data and reduce the impact of selecting irrelevant features.

The high F1-Score of CSFA reflects the balance between Recall and Precision, which are key factors in improving disease detection performance. Especially in cases of imbalanced data, this hybrid approach significantly reduces error rates and enhances the accuracy of the results.

TABLE IV. EVALUATE RESULTS

Algorithm	Data Splitting	Accuracy	Precision	Recall	F1-Score	Specificity	AUC-ROC
CS	70/30	94.15	92.06	92.06	92.06	95.37	93.71
	80/20	98.24	100	95.23	97.57	100	97.61
	90/10	100	100	100	100	100	100
FA	70/30	93.56	88.23	95.23	91.60	92.59	93.91
	80/20	98.24	100	95.23	97.56	100	97.61
	90/10	100	100	100	100	100	100
CSFA	70/30	98.24	100	95.23	97.56	100	97.61
	80/20	97.36	100	92.85	96.29	100	96.42
	90/10	100	100	100	100	100	100

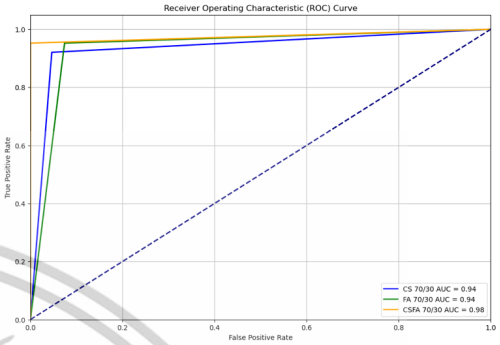


Fig. 6. ROC graph of the algorithms with a 70/30 dataset split

Fig 6 shows the Receiver Operating Characteristic (ROC) curve graph comparing the classification performance of the CS, FA, and CSFA algorithms. From the ROC graph, it is observed that CSFA achieves the highest AUC (Area Under the Curve) value of 0.98 compared to CS and FA, which both have an AUC of 0.94, in the case of a 70/30 dataset splitting. The CSFA graph demonstrates better data separation capability, with a higher True Positive rate and lower False Positive rate. This indicates that CSFA can create a more efficient and accurate model for this dataset. The AUC comparison between the algorithms reflects the potential of the hybrid method in improving models for disease detection.

Fig 7 is a heatmap showing the features selected by each algorithm across different dataset splits. The hybrid CSFA method effectively selects important features, demonstrating the algorithm's capability to eliminate irrelevant features. The heatmap data compared feature selection between CS, FA, and CSFA, highlighting the hybrid method's ability to balance feature selection better than individual algorithms.

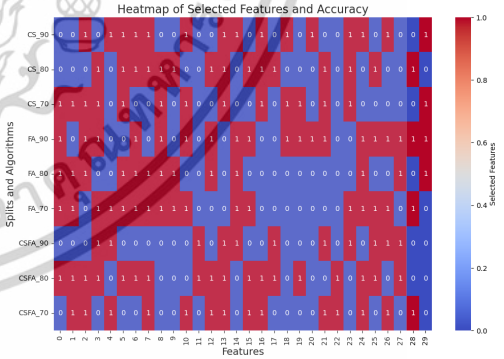


Fig. 7. Heatmap of feature selection

The use of CSFA accelerates the convergence process to the optimal solution, demonstrating that this technique has the potential to efficiently explore and refine the search space compared to using CS and FA individually. For the 90/10 dataset splitting, although CSFA achieved a perfect metric score of 100%, this result may indicate an overfitting issue. Additional testing using cross-validation or increasing the diversity of the dataset could help mitigate this problem. The results from CSFA demonstrate higher potential in terms of accuracy and stability, particularly in the context of breast cancer detection.

V. CONCLUSION

This research presents a hybrid feature selection method combining Cuckoo Search (CS) and Firefly Algorithm (FA), called CSFA for breast cancer detection using the Wisconsin Diagnostic Breast Cancer dataset. The experimental results demonstrate that the hybrid method significantly improves accuracy and efficiency in data classification, especially when compared to using CS and FA individually.

The analysis of the results indicates that CSFA can effectively handle complex problems, such as selecting significant features in high-dimensional data, and reduce the number of unnecessary features. Additionally, the use of CSFA helps mitigate the risk of overfitting to an appropriate degree.

However, there are certain limitations to consider. The Wisconsin Diagnostic Breast Cancer dataset from the UCI Machine Learning Repository is relatively small, containing only 569 samples, which is generally insufficient for training a model effectively. This limitation often leads to overfitting. Additionally, the dataset is divided into two classes, where 357 instances belong to the benign (B) class and 212 instances belong to the malignant (M) class. This distribution highlights a class imbalance, which can impact model performance by favoring the majority class.

For future work, we will focus on applying the approach to other medical datasets and performing additional hyperparameter tuning, as this study uses only a single set of parameter values, which may not fully optimize the model's performance. Subsequently, the improved method will be compared with other hybrid algorithms to ensure a comprehensive evaluation of algorithm performance, highlighting the potential of the hybrid method combining CS and FA in feature selection, which can be further applied in research related to data analysis and machine learning in various contexts.

REFERENCES

[1] H. Xie, L. Zhang, C.P. Lim, Y. Yu and H. Liu, "Feature selection using Enhanced Particle Swarm Optimisation for classification models," Sensors. 2021; 21(5):1816.

[2] C. F. Ozgur, "Genetic algorithm based feature selection in high dimensional text dataset classification," WSEAS Transactions on Information Sciences and Application. 2015. 12.

[3] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," World Congress on Nature & Biologically Inspired Computing (NaBIC), IEEE, Coimbatore, India, 2009.

[4] A. A. Joshi and R. M. Aziz, "A two-phase cuckoo search based approach for gene selection and deep learning classification of cancer disease using gene expression data with a novel fitness function," Multimedia Tools and Applications, vol. 83, pp. 71721–71752, 2024.

[5] B. K. Aljorani and A. H. Hasan, "An enhanced binary cuckoo search algorithm using crossover operators for features selection," International Conference on Advanced Computer Applications (ACA2021), IEEE, Imam ALkadhum College, Maysan, Iraq, 2021.

[6] B. A. Elizabeth, A. M. Usman, A. M. Hussein, W. O. Ebinum, T. U. Maigari, A. U. Abdullahi, J. A. Ibrahim, M. J. Usman, J. Philips, and M. Dawaki, "Chi-square and cuckoo search based feature selection for heart disease prediction," International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG), IEEE, Omu-Aran, Nigeria, 2024.

[7] Y. Kaya, "Comparison of using the genetic algorithm and cuckoo search for feature selection," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-5, doi: 10.1109/IDAP.2018.8620823.

[8] H. M. Abdulwahab, S. Ajitha, M. A. N. Saif, B. A. H. Murshed, and F. A. Ghanem, "MOBCSA: Multi-Objective Binary Cuckoo Search Algorithm for features selection in bioinformatics," in IEEE Access, vol. 12, pp. 21840-21867, 2024, doi: 10.1109/ACCESS.2024.3362228.

[9] X.-S. Yang, "Firefly algorithm, stochastic test functions and design optimization," International Journal of Bio-Inspired Computation, pp. 78–84, 2010.

[10] T. Badriyah, I. Syarif, and D. D. L. Prakoso, "Feature selection implementation on high-dimensional data using firefly algorithm," Eighth International Conference on Informatics and Computing (ICIC), IEEE, Manado, Indonesia, 2023.

[11] S. Y. Chemmalar, G. Selvi, Y. Gokul, G. Srivastava, and T. R. Gagekallu, "Firefly optimized federated SVM model for breast cancer prediction," Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, Regina, SK, Canada, 2023.

[12] P. Rana, I. Batra, and A. Malik, "An innovative approach: hybrid firefly algorithm for optimal feature selection," 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), Greater Noida, India, 2024, pp. 1-4.

[13] T. S. Sindhu, N. Kumaratharan, and P. Anandan, "A review on optimization algorithms for feature selection," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 851-855, doi: 10.1109/ICSCSS57650.2023.10169444.

[14] S. Maza and D. Zouache, "Binary Firefly Algorithm for feature selection in classification," 2019 International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS), Skikda, Algeria, 2019, pp. 1-6.

[15] Q. Chen, Y. Chen, and W. Jiang, "Genetic Particle Swarm Optimization–based feature selection for very-high-resolution remotely sensed imagery object change detection," Sensors. 2016; 16(8):1204. https://doi.org/10.3390/s16081204.

[16] B. Asgarali, G. Habib, and T. Omid, "An efficient hybrid algorithm using cuckoo search and differential evolution for data clustering," Indian Journal of Science and Technology, 2015.

ประวัติผู้เขียน

ชื่อ	นาย ชลันวิชญ์ ธีระสานต์
วัน เดือน ปีเกิด	28 มิถุนายน พ.ศ. 2541
ที่อยู่ปัจจุบัน	99/18 หมู่ 5 ตำบลโคกสะอาด อำเภอเมือง จังหวัดอุดรธานี 41000
ประวัติการศึกษา	(2563) วิศวกรรมศาสตรบัณฑิต เกียรตินิยมอันดับ 1 สาขาวิศวกรรมคอมพิวเตอร์และการสื่อสาร เกรดเฉลี่ย 3.72 มหาวิทยาลัยราชภัฏอุดรธานี
ผลงานทางวิชาการ	Teerasarn, C., & Kimpan, W. 2025. “Feature Selection Method Based on Hybrid Cuckoo Search and Firefly Algorithm for Breast Cancer Prediction” 2025 4th Asia Conference on Algorithms, Computing and Machine Learning (CACML)