JIST Journal of Information Science and Technology
Volume 11, NO 2 | JUL – DEC 2021 | 49-61
ISSN: 2651-1053 (Online)
ISSN: 1906-9553 (Print)

การเปรียบเทียบความแม่นยำการจำแนกประเภทข้อมูลอนุกรมเวลาในปริภูมิเวกเตอร์ ระหว่างวิธีแซ็คและวิธีบอส: กรณีศึกษา ข้อมูลคลื่นไฟฟ้าหัวใจ

The Accuracy Comparison of Time Series Classification in Vector Space between SAX and BOSS Methods: A Case Study of Electrocardiogram

นภัสสร แก้วกล้า และ อัครินทร์ ไพบูลย์พานิช*

Napatsorn Kaewkla and Akarin Phaibulpanich*

ภาควิชาสถิติ จุฬาลงกรณ์มหาวิทยาลัย

Department of Statistics, Chulalongkorn University

Received: November 25, 2021; Revised: December 17, 2021; Accepted: December 23, 2021; Published: December 28, 2021

ABSTRACT – The electrocardiogram (ECG) is an important procedure used to diagnose heart disorders. However, the ECG may contain different types of noise due to various of factors, potentially resulting in diagnostic errors. This research compares Symbolic Aggregate Approximation in Vector Space (SAXVSM) and Bag of Symbolic Fourier Approximation Symbols in Vector Space (BOSSVS) methods for classifying ECG data with noise. To choose a suitable classification algorithm for ECG5000 dataset, which is available in the Physionet database. Four types of ECG noises were simulated and then added to the data as follow: 1) Electromyography (EMG) 2) Powerline Interference 3) Baseline Wander and 4) Composite at 25%, 50% and 100% levels for the performance comparison of the ECG classification between normal and abnormal heart rhythms with SAXVSM and BOSSVS. The results show that both algorithms have similar high performance for all 13 datasets: accuracy and F1 score are 97-99%, precision is 95-99%, and recall is 97-100%, but BOSSVS has a longer running time than SAXVSM.

KEYWORDS: Time Series Classification, SAX, BOSS, Vector Space, Electrocardiogram

บทคัดย่อ -- การตรวจคลื่นไฟฟ้าหัวใจ เป็นหัตถการสำคัญที่ใช้วินิจฉัยความผิดปกติของหัวใจ แต่การตรวจวัดคลื่นไฟฟ้าหัวใจก็อาจมีสัญญาณรบกวนแบบต่าง ๆ ที่เกิดขึ้นจากหลายสาเหตุ ซึ่งอาจทำให้ผลการวินิจฉัยทางการแพทย์ผิดพลาด งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบอัลกอริทึมสำหรับการจำแนกประเภทข้อมูลคลื่นไฟฟ้าหัวใจที่มีสัญญาณรบกวนด้วย Symbolic Aggregate Approximation in Vector Space (SAXVSM) และ Bag of Symbolic Fourier

*Corresponding Author: 6280156126@student.chula.ac.th

Approximation Symbols in Vector Space (BOSSVS) เพื่อเลือกอัลกอริทึมการจำแนกประเภทข้อมูลคลื่นไฟฟ้าหัวใจอย่างเหมาะสม โดยใช้ข้อมูลคลื่นไฟฟ้าหัวใจ ECG5000 จากฐานข้อมูล Physionet และผู้วิจัยได้จำลองสัญญาณรบกวนในคลื่นไฟฟ้าหัวใจ 4 แบบ ได้แก่ 1) Electromyography (EMG) 2) Powerline Interference 3) Baseline Wander และ 4) Composite ที่ระดับ 25% 50% และ 100% เพื่อเปรียบเทียบประสิทธิภาพของการจำแนกประเภทการเต้นของหัวใจปกติและผิดปกติด้วย SAXVSM และ BOSSVS จากการวิจัยสามารถสรุปได้ว่า สำหรับข้อมูล 13 ชุด ทั้ง SAXVSM และ BOSSVSM มีประสิทธิภาพดีใกล้เคียงกัน โดยมีค่าความถูกต้องและคะแนน F1 อยู่ที่ 97-99% ค่าความแม่นยำอยู่ที่ 95-99% และค่าความระลึกอยู่ที่ 97-100% แต่ BOSSVS ใช้เวลาในการประมวลผลนานกว่า SAXVSM

คำสำคัญ: การจำแนกประเภทข้อมูลอนุกรมเวลา, แซ็ค, บอส, ปริภูมิเวกเตอร์, คลื่นไฟฟ้าหัวใจ

1. บทนำ

การตรวจคลื่นไฟฟ้าหัวใจ (Electrocardiogram: ECG) เป็นการวัดความต่างศักย์ของกระแสไฟฟ้าที่ไหลผ่านจุดกำเนิดไฟฟ้าไปยังเซลล์กล้ามเนื้อหัวใจ ซึ่งจะสัมพันธ์กับการบีบและคลายตัวของกล้ามเนื้อหัวใจ ดังนั้นการตรวจคลื่นไฟฟ้าหัวใจ จึงเป็นหัตถการสำคัญที่ใช้ในการวินิจฉัยภาวะผิดปกติของหัวใจ โดยจะใช้อิเล็กโทรด (Electrode) หรือขั้วไฟฟ้า ซึ่งต่ออยู่กับเครื่องบันทึกคลื่นไฟฟ้าหัวใจ มาวางไว้เหนือผิวหนังบริเวณต่าง ๆ ทำให้เกิดความต่างศักย์ไฟฟ้าระหว่างส่วนต่าง ๆ หรือที่เรียกว่าลีด (Lead) และบันทึกกระแสไฟฟ้าที่เกิดขึ้นเป็นรูปคลื่นต่าง ๆบนกระดาษกราฟ อย่างไรก็ดีการตรวจวัดคลื่นไฟฟ้าหัวใจนก็อาจมีปัจจัยหรือสัญญาณรบกวน (Noise) ซึ่งเกิดขึ้นจากหลายสาเหตุ เช่น ตำแหน่งของการติดเครื่องมือผิดพลาด การเคลื่อนไหวร่างกาย การรบกวนของสนามแม่เหล็กไฟฟ้า เป็นต้น ซึ่งสัญญาณรบกวนเหล่านี้อาจทำให้เกิดความผิดพลาดในการวินิจฉัยของแพทย์ได้ [1] ที่ผ่านมามีงานวิจัยที่นำเสนอตัวแบบการจำแนกประเภทข้อมูลอนุกรมเวลาที่มีประสิทธิภาพดี ได้แก่ Symbolic Aggregate Approximation - Vector Space Model (SAXVSM) [2, 3] และ Bag-Of-SFA-Symbols in Vector Space (BOSSVS) [4-6]

ขั้นตอนวิธี Symbolic Aggregate Approximation (SAX) และ Bag of Symbolic Fourier Approximation Symbols (BOSS) มีแนวคิดสำหรับการวิเคราะห์ข้อมูลอนุกรมเวลาด้วยการแบ่งอนุกรมเวลาย่อยและพยายามแปลงข้อมูลอนุกรมเวลาย่อยนี้ให้เป็นลำดับของตัวอักษร เพื่อลดมิติของข้อมูล โดยใช้ Piecewise Aggregate Approximation (PAA) และ Discrete Fourier Transform (DFT) ตามลำดับ ในขณะเดียวกันก็ยังคงเก็บคุณลักษณะสำคัญของอนุกรมเวลาเดิมไว้ นอกจากนี้ยังไม่ค่อยมีผลกระทบต่อสัญญาณรบกวน แต่ยังไม่มีงานวิจัยใดใช้สองวิธีนี้เพื่อศึกษาประสิทธิภาพการจำแนกประเภทคลื่นไฟฟ้าหัวใจภาวะหัวใจห้องล่างเต้นผิดจังหวะแบบ PVC ที่มีสัญญาณรบกวนแบบต่าง ๆ ผู้วิจัยจึงประยุกต์ใช้ SAX และ BOSS ร่วมกับปริภูมิเวกเตอร์ (Vector Space) ในการจำแนกประเภทข้อมูลและจำลองสัญญาณรบกวนในคลื่นไฟฟ้าหัวใจไว้ 4 แบบ คือ 1) Electromyography (EMG) 2) Powerline Interference 3) Baseline Wander และ 4) Composite Noise เพื่อเปรียบเทียบประสิทธิภาพการจำแนกประเภทของทั้งสองวิธีนี้

2. งานวิจัยและทฤษฎีที่เกี่ยวข้อง

Kaya & Pehlivan (2015) [7] เปรียบเทียบประสิทธิภาพ การจำแนกประเภทข้อมูลคลื่นไฟฟ้าหัวใจในภาวะหัวใจเต้นผิดจังหวะ ที่เกิดจากกระแสไฟฟ้าออกจากหัวใจห้องล่างแบบ Premature Ventricular Contraction (PVC) โดยใช้ข้อมูลคลื่นไฟฟ้าหัวใจจาก The MIT-BIH Arrhythmia Database แล้วสุ่มจังหวะการเต้นของหัวใจทั้งหมด 7,000 จังหวะ ประกอบด้วยปกติ (3,500 จังหวะ) และ PVC (3,500 จังหวะ) แล้วผ่านขั้นตอนการกรองสัญญาณรบกวนด้วยค่ามัธยฐาน (Median Filtering) จากนั้นได้เปรียบเทียบวิธีการลดมิติข้อมูล (Feature Reduction) ด้วย 3 วิธี คือ 1) การวิเคราะห์องค์ประกอบหลัก (Principal Components Analysis: PCA) 2) การวิเคราะห์องค์ประกอบอิสระ (Independent Component Analysis: ICA) 3) การทำแผนที่โยงก่อร่างตัวเอง (Self-Organizing Maps: SOM) แล้วเปรียบเทียบประสิทธิภาพการจำแนกประเภทด้วย 4 ตัวแบบ คือ 1) โครงข่ายประสาทเทียม (Neural Networks: NN)

2) การหาเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbour: K-NN) 3) ซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machines: SVM) และ 4) ต้นไม้ตัดสินใจ (Decision Tree: DT) ในภาพรวมตัวแบบ K-NN ใช้เวลาในการจำแนกประเภทน้อยและมีประสิทธิภาพการจำแนกประเภทดีที่สุด

Senin & Malinchik (2013) [3] นำเสนอตัวแบบ Symbolic Aggregate Approximation - Vector Space Model (SAX-VSM) แนวคิดหลักของตัวแบบนี้ คือ การแบ่งอนุกรมเวลาออกเป็นอนุกรมเวลาย่อย (Sliding Window) และใช้ Piecewise Aggregate Approximation (PAA) เพื่อลดมิติของข้อมูลอนุกรมเวลา แล้วจึงใช้แซ็ค (SAX) เพื่อแปลงข้อมูลอนุกรมเวลาให้เป็นรูปแบบกลุ่มตัวอักษรและใช้ตัวแบบปริภูมิเวกเตอร์ (Vector Space Model) เพื่อจำแนกประเภทข้อมูลอนุกรมเวลาจากรูปแบบกลุ่มตัวอักษรเหล่านี้ ในงานวิจัยนี้ได้เปรียบเทียบอัตราความคลาดเคลื่อนการจำแนกประเภทด้วย 5 ตัวแบบ คือ 1) 1NN-Euclidean 2) 1NN-Dynamic Time Wraping (DTW) 3) Fast Shapelets และ 4) Bag of Patterns และ 5) SAX-VSM โดยใช้ข้อมูลจาก UCR Time Series ทั้งหมด 19 ชุด แสดงให้เห็นว่า มีข้อมูล 12 ชุดจากทั้งหมด 19 ชุด ที่ SAX-VSM ให้อัตราความคลาดเคลื่อน (Error Rate) ต่ำที่สุด

Schäfer (2014, 2015) [4-6] นำเสนอตัวแบบ Bag-Of-SFA-Symbols in Vector Space (BOSS VS) มีแนวคิดมาจากการแปลงสัมประสิทธิ์ฟูเรียร์แบบไม่ต่อเนื่อง (Discrete Fourier Transform: DFT) จากงานวิจัยได้ทำการเปรียบเทียบประสิทธิภาพและเวลาสะสมที่ใช้ในการเรียนรู้ (Training) และทดสอบ (Testing) โดยทดลองกับข้อมูลอนุกรมเวลาจาก UCR Time Series และข้อมูลอื่น ๆ รวมทั้งหมด 91 ชุด พบว่า BOSS ใช้เวลาการประมวลผลสะสมน้อยและประสิทธิภาพการจำแนกประเภทโดยเฉลี่ยดีกว่า เมื่อเทียบกับ 1NN-Dynamic Time Wraping (DTW) ในงานวิจัยระบุว่า DFT มีหลักการ คือ แปลงข้อมูลอนุกรมเวลาที่เป็นสัญญาณจากโดเมนเวลา (Time Domain) ไปเป็นโดเมนความถี่ (Frequency Domain) ทำให้สามารถลดสัญญาณรบกวนได้

คลื่นไฟฟ้าหัวใจ (Electrocardiogram: ECG) เป็นข้อมูลอนุกรมเวลา (Time Series) ซึ่งประกอบด้วย $T=(t_1,t_2,\ldots,t_n)$ คือ ลำดับของค่าข้อมูลที่ถูกบันทึกตามช่วงเวลา โดยที่ $n \in N$

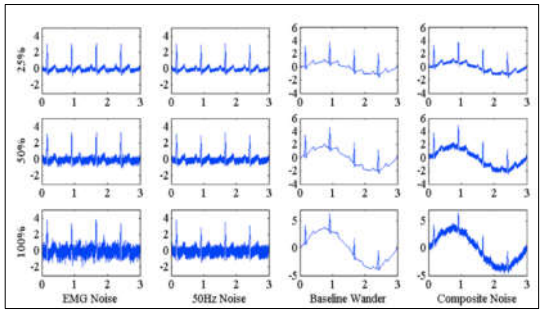
สัญญาณรบกวนในคลื่นไฟฟ้าหัวใจ คือ สัญญาณผิดปกติที่เกิดขึ้นในขณะบันทึกคลื่นไฟฟ้าหัวใจ ซึ่งอาจส่งผลต่อประสิทธิภาพของการตรวจจับสัญญาณการเต้นของหัวใจ โดยสัญญาณรบกวนที่มักจะพบในคลื่นไฟฟ้าหัวใจและนำมาศึกษาในงานวิจัยนี้มี 4 แบบ [8] ได้แก่

1) Electromyographic Noise (EMG) เป็นสัญญาณรบกวนที่เกิดจากการหดเกร็งของกล้ามเนื้อหรือการเคลื่อนไหวของร่างกายอย่างกระทันหัน ความถี่ใน EMG ทับซ้อนกันมาก ทำให้ตรวจจับยาก

2) Powerline Interference (50 Hz) เป็นสัญญาณรบกวนความถี่สูง โดยทั่วไปจะมีความถี่ประมาณ 50 เฮิรตซ์ มีลักษณะเป็นสัญญาณรบกวนแบบคลื่นไซน์ อาจจะมาพร้อมฮาร์มอนิก (Harmonic) จำนวนหนึ่ง กล่าวคือ กระแสหรือแรงดันในรูปสัญญาณคลื่นไซน์ของสัญญาณหรือปริมาณในคาบใด ๆ ที่มีความถี่เป็นจำนวนเท่าของความถี่หลักมูล (Fundamental Frequency) สาเหตุหลักเกิดจากการรบกวนคลื่นแม่เหล็กไฟฟ้าจากสายไฟ (Electromagnetic Field : EMF) หรือจากเครื่องจักรในระยะใกล้ๆ การต่อสายดินที่ไม่ดี เป็นต้น

3) Baseline Wander เป็นสัญญาณรบกวนความถี่ต่ำมีลักษณะรบกวนจากแนวแกน x สูงขึ้นหรือลดต่ำลงจากแนวเส้นฐาน (Baseline) สาเหตุหลักเกิดจากการหายใจ การติดขั้วไฟฟ้าไม่ดี เป็นต้น

4) Composite Noise เป็นสัญญาณรบกวนที่เกิดจากการเกิดร่วมกันของสัญญาณรบกวนทั้ง 3 แบบที่กล่าวมา



รูปที่ 1. แสดงตัวอย่างสัญญาณรบกวนแต่ละแบบที่ระดับ 25% 50% และ 100%

แหล่งที่มา : "Arrhythmia ECG Noise Reduction by Ensemble Empirical Mode Decomposition" โดย Chang, K. M., 2010

3. Symbolic Aggregate Approximation (SAX)

Symbolic Aggregate Approximation [2, 3] เป็นวิธีการแปลงข้อมูลอนุกรมเวลาให้เป็นรูปแบบของกลุ่มตัวอักษรเพื่อลดมิติของข้อมูล โดยมีขั้นตอนดังนี้

3.1 การทำข้อมูลให้เป็นมาตรฐาน (Z-normalization)

กำหนดให้ อนุกรมเวลา $T=(t_1,t_2,\ldots,t_n)$ คือ ลำดับของค่าข้อมูลที่ถูกบันทึกตามช่วงเวลา โดยที่ $n \in N$ [9]

$$T_{norm}=\frac{T-\mu}{\sigma} \quad (1)$$

โดยที่ T คือ ค่าข้อมูลอนุกรมเวลา

μ คือ ค่าเฉลี่ยของอนุกรมเวลา

σ คือ ส่วนเบี่ยงเบนมาตรฐานของอนุกรมเวลา

3.2 การทำหน้าต่างบานเลื่อน (Sliding Windows)

การแบ่งอนุกรมเวลา $T=(t_1,t_2,\ldots,t_n)$ ที่มีความยาว n ออกเป็นหน้าต่างขนาด w คงที่ จะได้

$$\text{windows}(T,w)=\left\{\underbrace{S_{1;w}}_{(t_1,t_2,\ldots,t_w)},\ \underbrace{S_{2;w}}_{(t_2,t_3,\ldots,t_{w+1})},\ \ldots,S_{n-w+1;w}\right\} \quad (2)$$

โดยที่ $i=1,2,\ldots,n-w+1$

$S_{i;w}$ คือ หน้าต่างบานเลื่อนที่ i ความยาว w

i คือ ลำดับของหน้าต่างแต่ละบาน

ดังนั้น หน้าต่างสองบานที่ติดกันจะทับซ้อนกัน w ตำแหน่ง [6]

3.3 Piecewise Aggregate Approximation (PAA)

การลดมิติของข้อมูล (Dimension Reduction) ด้วย Piecewise Aggregate Approximation (PAA) เป็นวิธีการลดมิติของข้อมูลโดยการแบ่งข้อมูลออกเป็นส่วนย่อย ๆ ส่วนละเท่า ๆ กัน และคำนวณค่าเฉลี่ยของแต่ละส่วน เพื่อเป็นตัวแทนของข้อมูลในแต่ละส่วน

กำหนดให้ อนุกรมเวลามีความยาว w จุดเวลา จะแบ่งข้อมูลออกเป็น p ส่วน และหาค่าเฉลี่ยของข้อมูลในแต่ละส่วน [2]

$$\overline{c_i}=\frac{p}{w}\sum_{j=\frac{w}{p}(i-1)+1}^{\frac{w}{p}i} C_j \quad (3)$$

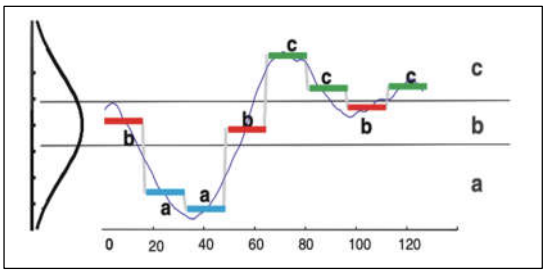
โดยที่ $i=1,\ldots,p$ และ $j=1,\ldots,w$

$\overline{C}_i$ คือ ค่าเฉลี่ยของอนุกรมเวลส่วนที่ i

C_j คือ ค่าของข้อมูลอนุกรมเวลาจุดที่ j

3.4 การแบ่งช่วงข้อมูล(Discretization) ด้วยวิธีแซ็ค(SAX)

การแบ่งช่วงข้อมูล (Discretization) ด้วยวิธีแซ็ค (SAX) เป็นการแปลงค่าเฉลี่ยของข้อมูลในแต่ละส่วนที่ได้จากขั้นตอน PAA ให้เป็นตัวอักษร โดยจะต้องกำหนดจำนวนตัวอักษร (alphabet size) ที่ใช้ และแบ่งจำนวนช่วงของข้อมูลให้เท่ากับจำนวนตัวอักษร โดยที่พื้นที่ใต้กราฟในแต่ละช่วงจะต้องมีขนาดเท่า ๆ กัน จะได้จุดแบ่ง (Breakpoint: β) แต่ละช่วง ด้วยค่ามาตรฐาน (Z- Score) โดยกำหนดความยาวเท่ากับ 8 และจำนวนตัวอักษรเท่ากับ 3 (A, B, C)



รูปที่ 2. แสดงตัวอย่างการแปลงข้อมูลอนุกรมเวลาโดยวิธีแซ็ค (SAX) แหล่งที่มา : "Experiencing SAX: a novel symbolic representation of time series" โดย Lin et al., 2007

3.5 การลดขนาดข้อมูลด้วย Numerosity Reduction

Numerosity Reduction จะนับรูปแบบของตัวอักษรที่ซ้ำกันในหน้าต่างบานเลื่อนที่ติดกันเพียงครั้งเดียวและจะนับซ้ำอีกครั้ง ก็ต่อเมื่อ พบรูปแบบของตัวอักษรเดิมที่เกิดขึ้นอีกในหน้าต่างลำดับอื่น ๆ ที่ไม่ติดกัน [2]

4. Bag of Symbolic Fourier Approximation Symbols (BOSS)

การประมาณฟูเรียร์เชิงสัญลักษณ์ Symbolic Fourier Approximation (SFA) [4-6] เป็นอีกวิธีหนึ่งในการแปลงข้อมูลอนุกรมเวลาให้เป็นรูปแบบตัวอักษร โดยมี 2 ขั้นตอนแรก คือ การทำข้อมูลให้เป็นมาตรฐาน (Z-Normalization) และการทำหน้าต่างบานเลื่อน (Sliding Windows) ตามที่ได้กล่าวมาในขั้นตอนของแซ็ค (SAX) และมีขั้นตอนต่อไปดังนี้

4.1 การประมาณค่าสัมประสิทธิ์การแปลงฟูเรียร์แบบไม่ต่อเนื่อง

การประมาณค่าการแปลงฟูเรียร์แบบไม่ต่อเนื่อง (Discrete Fourier Transform: DFT) เป็นการประมาณค่าสัมประสิทธิ์ของอนุกรมฟูเรียร์จากข้อมูลสัญญาณรายจุดหรือสัญญาณไม่ต่อเนื่อง (Discrete) รายคาบตั้งแต่ 0 ถึง N-1 ที่แบ่งมาจากสัญญาณต่อเนื่องเพื่อที่จะแปลงสัญญาณโดเมนเวลาให้เป็นโดเมนความถี่ [10]

$$X(m)=\sum_{n=0}^{N-1}x(n)e^{\frac{-j2\pi nm}{N}} \quad (4)$$

โดยที่ $n,m=0,1,...,N-1$ และ $j=\sqrt{-1}$

$x(n)$ คือ ลำดับของตัวเลขจากสัญญาณไม่ต่อเนื่อง

$X(m)$ คือ ผลการแปลงฟูเรียร์แบบไม่ต่อเนื่องของ $x(n)$

จากสมการของออยเลอร์ (Euler's Equation)

$$e^{j\theta}=\cos(\theta)-j\sin(\theta) \quad (5)$$

โดยที่ $\theta=\frac{2\pi nm}{N}$

ถ้าแทนสมการที่ (5) ลงในสมการที่ (4) จะได้

$$X(m)=\sum_{n=0}^{N-1}x(n)[\cos(\frac{2\pi nm}{N})-j\sin(\frac{2\pi nm}{N})] \quad (6)$$

จะได้ว่า $X(m)$ เป็นค่าสัมประสิทธิ์การแปลงฟูเรียร์ ซึ่งจะอยู่ในรูปของจำนวนเชิงซ้อน ประกอบด้วย (ส่วนจริง(m) , ส่วนจินตภาพ(m)) โดยที่ $m=0,1,2,...,N-1$

4.2 การทำควอนไทซ์ (Quantisation)

การทำควอนไทซ์ (Quantisation) เป็นการกำหนดช่วงของข้อมูลที่จะแปลงค่าประมาณสัมประสิทธิ์การแปลงฟูเรียร์แบบไม่ต่อเนื่อง (DFT) เป็นตัวอักษร (alphabet) ด้วยเทคนิค The Multiple Coefficient Binning (MCB) [4] โดยมีการเตรียมข้อมูลดังนี้

กำหนดให้ เมทริกซ์ A ที่มาจากขั้นตอนการแปลงฟูเรียร์แบบไม่ต่อเนื่อง (DFT) ของชุดข้อมูลเรียนรู้ (Training Data) แสดงได้สมการที่ (7)

$$A=\begin{pmatrix}\mathrm{DFT}(T_1)\\ \vdots\\ \mathrm{DFT}(T_i)\\ \vdots\\ \mathrm{DFT}(T_N)\end{pmatrix}=\begin{pmatrix}\mathrm{real}_{1,1}\ \mathrm{img}_{1,1}\ \ldots\ \mathrm{real}_{1,\frac{l}{2}}\ \mathrm{img}_{1,\frac{l}{2}}\\ \vdots\\ \mathrm{real}_{i,1}\ \mathrm{img}_{i,1}\ \ldots\ \mathrm{real}_{i,\frac{l}{2}}\ \mathrm{img}_{i,\frac{l}{2}}\\ \vdots\\ \mathrm{real}_{N,1}\ \mathrm{img}_{N,1}\ \ldots\ \mathrm{real}_{N,\frac{l}{2}}\ \mathrm{img}_{N,\frac{l}{2}}\end{pmatrix}=(C_1,C_2,..,C_l) \quad (7)$$

โดยที่ $i=1,...,N$ $k=1,...,\frac{l}{2}$ และ $j=1,...,l$

T_i คือ ชุดข้อมูลเรียนรู้ (Training Data) แถวที่ i

$real_{ik}$ คือ ค่าจริงของค่าสัมประสิทธิ์ฟูเรียร์ตัวที่ k ของข้อมูลเรียนรู้ตัวที่ i

img_{ik} คือ ค่าจินตภาพของค่าสัมประสิทธิ์ฟูเรียร์ตัวที่ k ของข้อมูลเรียนรู้ตัวที่ i

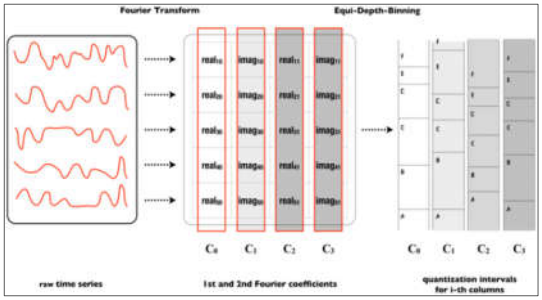
C_j คือ ค่าสัมประสิทธิ์ฟูเรียร์คอลัมน์ที่ j

The Multiple Coefficient Binning (MCB) และมีขั้นตอนดังนี้

ขั้นตอนที่ 1 เรียงค่าของข้อมูลในแต่ละคอลัมน์ C_j

ขั้นตอนที่ 2 กำหนดจำนวนตัวอักษรเท่ากับ a และแบ่งข้อมูลออกเป็น a ส่วน โดยที่จำนวนข้อมูลในแต่ละส่วนเท่า ๆ กัน (equi-depth binning bins) จะได้ จุดแบ่ง (breakpoints) สำหรับคอลัมน์ C_j

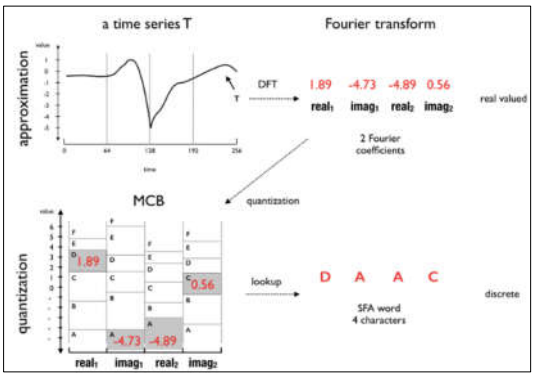
ขั้นตอนที่ 3 กำหนดลำดับตัวอักษรประจำตำแหน่งในแต่ละส่วนในคอลัมน์ C_j จะได้ตัวอย่างการทำ The Multiple Coefficient Binning (MCB) แสดงได้ดังรูปที่ 3.



รูปที่ 3. แสดงขั้นตอน The Multiple Coefficient Binning (MCB)
แหล่งที่มา: "Human Activity Recognition Based on Symbolic Representation Algorithms for Inertial Sensors"
โดย Wesllen Sousa Lima et al., 2018

4.3 Symbolic Fourier Approximation

การแปลงฟูเรียร์เชิงสัญลักษณ์ (Symbolic Fourier Approximation) เป็นการแปลงค่าประมาณสัมประสิทธิ์ฟูเรียร์ (DFT) ให้เป็นตัวอักษร ด้วย Multiple Coefficient Binning (MCB)



รูปที่ 4. แสดงตัวอย่าง Symbolic Fourier Approximation
แหล่งที่มา: "Bag-Of-SFA-Symbols in Vector Space (BOSS VS)" โดย Schäfer, 2015

5. แบบจำลองปริภูมิเวกเตอร์ (Vector Space Model : VSM)

แบบจำลองปริภูมิเวกเตอร์ (Vector Space Model: VSM) เป็นวิธีหนึ่งในการแทนเอกสารที่ไม่มีโครงสร้าง (Unstructured Text Document) ด้วยแบบจำลองปริภูมิเวกเตอร์ โดยกำหนดให้เอกสารแต่ละฉบับเปรียบเสมือนเวกเตอร์ของคำ ขนาดของเวกเตอร์ขึ้นอยู่กับจำนวนของคำที่ปรากฏอยู่ในเอกสารฉบับนั้น [11]

5.1 การหาความถี่ของคำและการผกผันความถี่ในเอกสาร (Term Frequency-Inverse Document Frequency: TF-IDF) [5]

$$tf_{t,T} = \begin{cases} 1+\log(B_T(t)) \ ; \ if \ B_T(t) > 0 \\ 0 \ ; \ otherwise \end{cases} \quad (8)$$

โดยที่ T คือ อนุกรมเวลา

$B_T(t)$ คือ ความถี่ของคำ (t) ของอนุกรมเวลา T

$$tf_{t,C} = \begin{cases} 1+\log(\sum_{T \in C} B_T(t)) \ ; \ if \ \sum_{T \in C} B_T(t) > 0 \\ 0 \ ; \ otherwise \end{cases} \quad (9)$$

โดยที่ $\sum_{T \in C} B_T(t)$ คือ ความถี่ของคำ (t) ของอนุกรมเวลา T ที่อยู่ในประเภท C

$$idf_{t,C} = \log(1 + \frac{|\ CLASSES\ |}{|\ \{C \mid T \in C \cap B_T(t) > 0\}\ |}) \quad (10)$$

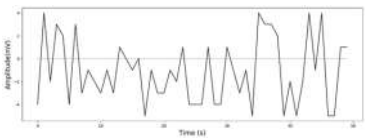
โดยที่ $|\ \{C \mid T \in C \cap B_T(t) > 0\}\ |$ คือ จำนวนประเภท C ที่พบคำ T

$|\ CLASSES\ |$ คือ จำนวนประเภท C ทั้งหมด

$tf \times idf(t,C)$ ใช้วัดความสำคัญของคำ (t) ในประเภท (C) แต่จะถูกลดทอนด้วยความสำคัญของคำ (t) ในประเภท (C) ทั้งหมด สามารถคำนวณได้ดังสมการที่ (10)

$$tf \times idf(t,C) = 1 + \log(\sum_{T \in C} B_T(t)) \times \log(1 + \frac{|\ CLASSES\ |}{|\ \{C \mid T \in C \cap B_T(t) > 0\ ||}) \quad (11)$$

(1) การแปลงข้อมูลให้เป็นมาตรฐาน (Z-normalization)



(2) การทำหน้าต่างบานเลื่อน (Sliding windows)

$$\text{windows}(T, w) = \left\{ \underbrace{S_{1;w}}_{(t_1,t_2,\ldots,t_w)}, \underbrace{S_{2;w}}_{(t_2,t_3,\ldots,t_{w+1})}, \ldots, S_{n-w+1;w} \right\}$$

(3) การแปลงฟูเรียร์เชิงสัญลักษณ์ด้วย Symbolic Aggregate Approximation (SAX) หรือ Bag of Symbolic Fourier Approximation: BOSS

$S_{1;w}$ = (aaccbabc, aaccbabc, ... baccbcaa, ccaccbaa)

$S_{2;w}$ = (baccbabc, acccccbc, ... bacccaba, bacccaba, ccaccbaa)

: $S_{n-w+1;w}$ = (baccbbaa, cbcabbcc, cbcabbcc, ... abccaacc, acccbabc, bbcbaaca)

(4) การลดขนาดข้อมูลด้วย Numerosity Reduction

$S_{1;w}$ = (aaccbabc, ... baccbcaa, ccaccbaa)

$S_{2;w}$ = (baccbabc, acccccbc, ... bacccaba, ccaccbaa)

:

$S_{n-w+1;w}$
= (baccbbaa, cbcabbcc, ... abccaacc, acccbabc, bbcbaaca)

(5) การนับความถี่ของรูปแบบตัวอักษรต่าง ๆ Bag Of SFA Symbol: BOSS หรือ Bag of Symbolic Fourier Approximation: BOSS

(6)

Window	aaccbabc	abcceaaa	bbcceaaa	…
1	0	10	3	…
2	1	0	2	…
:	:	:	:	…
n	…	…	…	…

รูปที่ 5. แสดงสรุปขั้นตอนของ SAX และ BOSS

5.2 ค่าความเหมือนโคไซน์ (Cosine similarity)

ค่าความเหมือนโคไซน์ (Cosine similarity) คือ การหาความคล้ายคลึงด้วยองศา เป็นการวัดความเหมือนของเวกเตอร์ 2 เวกเตอร์ว่าไปในทิศทางเดียวกันหรือไม่ โดยตัดขนาด (Magnitude) ของเวกเตอร์ออกไป [12]

$$similarity(Q,C) = \frac{\vec{Q} \cdot \vec{C}}{\| \vec{Q} \| \cdot \| \vec{C} \|} \quad (12)$$

$$similarity(Q,C) = \frac{\sum_{t \in Q} tf(t,Q) \cdot (tf \times idf(t,C) + 1)}{\sqrt{\sum_{t \in Q} (tf(t,Q))^2} \sqrt{\sum_{t \in C} (tf \times idf(t,Q))^2}} \quad (13)$$

$$label(Q) = \underset{C \in CLASSES}{\arg\max} (similarity(Q,C)) \quad (14)$$

โดยที่ $similarity(Q,C)$ คือ ค่าความเหมือนโคไซน์ของอนุกรมเวลา Qกับประเภท C

$label(Q)$ คือ ผลลัพธ์หรือประเภทของอนุกรมเวลา Q $\sum_{t \in Q} tf(t,Q)$ คือ ความถี่ของคำ t ใน Q

$tf \times idf(t,C)$ คือ ความถี่ของคำและการผกผันความถี่ (TF-IDF) ของคำในประเภท C

6. การจำลองสัญญาณรบกวนในคลื่นไฟฟ้าหัวใจ

ในงานวิจัยนี้จะใช้วิธีการจำลองสัญญาณรบกวนในคลื่นไฟฟ้าหัวใจจาก งานวิจัย K. M. Chang, 2010 [8]

คำจำกัดความในการจำลองสัญญาณรบกวน

คำจำกัดความที่ 1: Function คือ ฟังก์ชันในการจำลองของสัญญาณรบกวน

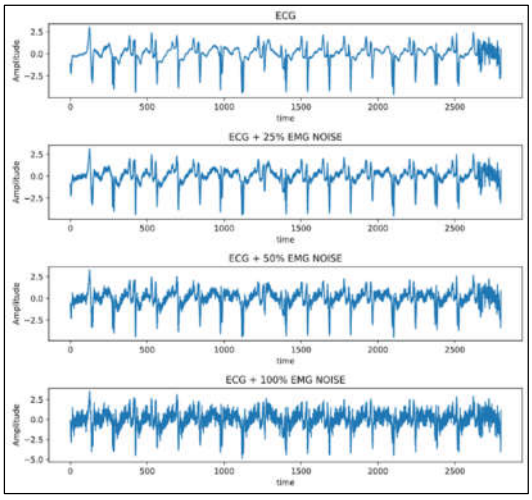
คำจำกัดความที่ 2: Vpp คือ ระยะขจัดตั้งแต่จุดสูงสุดถึงจุดต่ำสุดของคลื่นไฟฟ้าหัวใจปกติ

คำจำกัดความที่ 3: Frequency คือ ความถี่ของสัญญาณรบกวน

คำจำกัดความที่ 4: Reduced Ratio คือ อัตราส่วนสัญญาณต่อสัญญาณรบกวน

ตารางที่ 1. แสดงการจำลองสัญญาณรบกวน

Noise Type	Function	Vpp	Frequency	Reduced Ratio
EMG	Normal	5 mV	-	$\frac{1}{8}$
Powerline Interference (50Hz)	Sine	5 mV	50 Hz	$\frac{1}{4}$
Baseline Wander	Sine	5 mV	0.333 Hz	-
Composite	$0.5\ x\ (EMG\ +\ Powerline\ Interference) + Baseline$			



รูปที่ 6. แสดงตัวอย่างคลื่นไฟฟ้าหัวใจ 20 จังหวะ ที่ปกติและที่มีสัญญาณรบกวนแบบ EMG ระดับ 25% 50% และ 100%

7. เครื่องมือและวิธีการ

งานวิจัยนี้จะใช้ข้อมูล ECG5000 ซึ่งประกอบด้วยจังหวะการเต้นของหัวใจทั้งหมด 5,000 จังหวะ ความยาว 140 จุดเวลา และจัดประเภทการเต้นของหัวใจที่ปกติและผิดปกติไว้ทั้งหมด 5 ประเภท แต่ด้วยสัดส่วนข้อมูลประเภทที่ 2 3 และ 4 มีน้อยเกินไป ดังนั้นในงานวิจัยนี้จะใช้ข้อมูลในประเภท 0 (จังหวะการเต้นของหัวใจปกติ) และประเภท 1 (ภาวะหัวใจห้องล่างเต้นผิดจังหวะแบบพีวีซี R-ON-T) ทำให้เหลือข้อมูลทั้งหมด 4,686 จังหวะ และจำลองสัญญาณรบกวน 4 แบบ ได้แก่ 1) EMG Noise 2) Powerline Interference 3) Baseline Wander และ 4) Composite Noise และปรับระดับของสัญญาณรบกวนที่ 25% 50% 100% และไม่มีสัญญาณรบกวน เพื่อเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลคลื่นไฟฟ้าหัวใจระหว่าง SAXVSM และ BOSSVS

งานวิจัยนี้ใช้ข้อมูลทั้งหมด 4,686 จังหวะ แต่ละจังหวะมีความยาว 140 จุดเวลา ดังตารางที่ 2.

ตารางที่ 2. แสดงตัวอย่างข้อมูลคลื่นไฟฟ้าหัวใจ

No.	t_1	...	t_{140}	Class
1	-0.6566696	...	0.2257076	1
2	0.3973526	...	0.52316457	0
3	-1.862747	...	-0.6100227	0
...	...	...	...	...
4686	-1.7946848	...	-1.2108684	0

สำหรับชุดที่ใช้ศึกษา ประกอบด้วย ข้อมูลคลื่นไฟฟ้าหัวใจที่มีจำลองสัญญาณรบกวน 4 แบบ ได้แก่ 1) EMG Noise 2) Powerline Interference 3) Baseline Wander และ 4) Composite Noise แต่ละแบบปรับระดับของสัญญาณรบกวนไว้ที่ 25% 50% 100% จะได้ชุดข้อมูลที่ใช้ศึกษา ดังตารางที่ *3*.

ตารางที่ 3. แสดงชุดข้อมูลที่ใช้ศึกษา

ชุดข้อมูล	EMG Noise	Powerline Interference	Baseline Wander	Composite Noise
1	-	-	-	-
2	25%	-	-	-
3	50%	-	-	-
4	100%	-	-	-
7	-	25%	-	-
8	-	50%	-	-
9	-	100%	-	-
…	…	…	…	…
13	-	-	-	100%

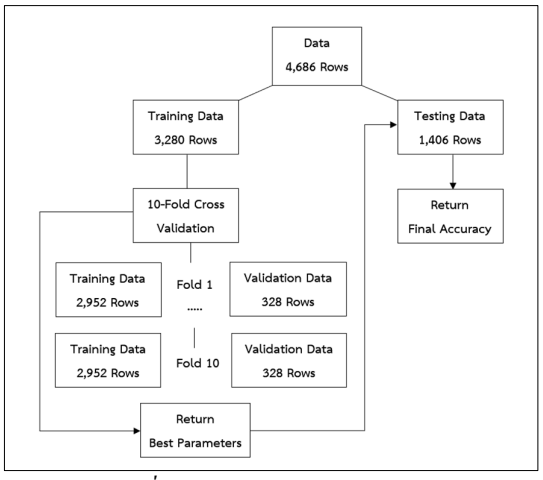
7.1 การแบ่งข้อมูล

จากข้อมูลทั้งหมดจะถูกแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลชุดเรียนรู้ (Training Data) สัดส่วน 70% (3,280 ตัวอย่าง) และข้อมูลชุดทดสอบ (Testing Data) สัดส่วน 30% (1,406 ตัวอย่าง) สำหรับข้อมูลชุดเรียนรู้ (Training Data) สัดส่วน 70% จะนำไปทำ 10-Folds Cross Validation เพื่อหาพารามิเตอร์ที่ดีที่สุด ส่วนข้อมูลชุดทดสอบ (Testing Data) สัดส่วน 30% จะนำไปทดสอบความแม่นยำ

7.2 การเลือกพารามิเตอร์

เริ่มจากการกำหนดขอบเขตล่างและขอบเขตบนสำหรับพารามิเตอร์แต่ละตัว จากนั้นจะสุ่มเลือกชุดของพารามิเตอร์ เพื่อหาค่าความแม่นยำโดยเฉลี่ยของพารามิเตอร์ชุดนั้น ๆ สำหรับ

SAX และ BOSS โดยจะใช้ชุดข้อมูลเรียนรู้ (Training Data) ทำ K-fold Cross Validation เพื่อแบ่งข้อมูลออกเป็นชุดข้อมูลเรียนรู้ (Training Data) และข้อมูลชุดตรวจสอบ (Validation Data)



รูปที่ 7. แสดงแผนภาพการแบ่งข้อมูล

ตารางที่ 4. แสดงพารามิเตอร์

พารามิเตอร์	ความหมาย
Window Size (w)	ขนาดหน้าต่างบานเลื่อน
Word Size (p)	ความยาวของตัวอักษร
Alphabet Size (a)	จำนวนตัวอักษร

สำหรับในแต่ละส่วน (fold) มีขั้นตอนย่อย ดังนี้

ขั้นตอนที่ 1 ใช้ชุดข้อมูลเรียนรู้ (Training Data) หาความถี่ของรูปแบบตัวอักษร (Bag of Patterns) โดยใช้วิธีของ SAX และ BOSS

ขั้นตอนที่ 2 สร้างเมทริกซ์ความถี่ของคำ (Term Frequency Matrix)

ขั้นตอนที่ 3 คำนวณความถี่ของคำและการผกผันความถี่ในเอกสาร (tf-idf) ของรูปแบบต่าง ๆ ในแต่ละประเภท (Class)

ขั้นตอนที่ 4 ใช้ชุดข้อมูลตรวจสอบ (Validation Data) หาความถี่ของรูปแบบตัวอักษร ตามขั้นตอนของ SAX และ BOSS

ขั้นตอนที่ 5 คำนวณค่าความเหมือนโคไซน์ (Cosine Similarity) ระหว่าง tf-idf จากขั้นตอนที่ 3) กับความถี่ของตัวอักษรในชุดข้อมูลตรวจสอบ (Validation Data) จากขั้นตอนที่ 4) หาค่าความเหมือนโคไซน์ (Cosine Similarity) เพื่อจำแนกประเภทข้อมูล (Class) ให้กับชุดข้อมูลนี้ ถ้าประเภทใดมีค่าความเหมือนโคไซน์ (Cosine Similarity) สูงที่สุด จะทำนายว่าเป็นประเภทนั้น (Predicted Class)

ขั้นตอนที่ 6 คำนวณค่าความแม่นยำ (Accuracy) ระหว่างค่าจริง (Actual Class) กับค่าทำนาย (Predicted Class) ของชุดข้อมูลตรวจสอบ (Validation Data)

ขั้นตอนที่ 7 หาค่าเฉลี่ยของค่าความแม่นยำทั้ง K รอบ สำหรับพารามิเตอร์ชุดนั้น ๆ และเลือกชุดของพารามิเตอร์ที่ดีที่สุดที่ให้ค่าเฉลี่ยค่าความแม่นยำสูงที่สุด เพื่อไปใช้ทดสอบประสิทธิภาพในข้อมูลชุดทดสอบ

7.3 การทดสอบประสิทธิภาพ

ใช้ชุดของพารามิเตอร์ที่ได้จากขั้นตอนที่ 7 ไปทดสอบกับชุดข้อมูลทดสอบ (Testing Data) และคำนวณค่าความแม่นยำ (Accuracy) ได้จาก *Confusion Matrix* สำหรับ SAX และ BOSS

ตารางที่ 5. แสดง Confusion Matrix

	ค่าทำนาย (Predicted Class)	
ค่าจริง (Actual Class)	**Positive**	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

โดยที่ True Positive (TP) คือ จำนวนข้อมูลที่ถูกทำนายว่า “จริง” และ มีค่าเป็น “จริง”

True Negative (TN) คือ จำนวนข้อมูลที่ถูกทำนายว่า “ไม่จริง” และ มีค่า “ไม่จริง”

False Positive (FP) คือ จำนวนข้อมูลที่ถูกทำนายว่า “จริง” แต่ มีค่าเป็น “ไม่จริง”

False Negative (FN) คือ จำนวนข้อมูลที่ถูกทำนายว่า “ไม่จริง” แต่ มีค่าเป็น “จริง”

ค่าความถูกต้อง (Accuracy) สามารถคำนวณได้ดังสมการที่ 15

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

ค่าความแม่นยำ (Precision) สามารถคำนวณได้ดังสมการที่ 16

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

ค่าความระลึก (Recall) สามารถคำนวณได้ดังสมการที่ 17

$$Recall = \frac{TP}{TP + FN} \quad (17)$$

คะแนน F1 (F1-Score) สามารถคำนวณได้ดังสมการที่ 18

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (18)$$

8. ผลการวิเคราะห์ข้อมูล

การเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลอนุกรมเวลาระหว่างวิธี SAXVSM และ BOSSVS สำหรับข้อมูลคลื่นไฟฟ้าหัวใจที่มีสัญญาณรบกวนแบบต่าง ๆ และที่ระดับต่างๆ เริ่มต้นจากการจำลองการเพิ่มสัญญาณรบกวนในคลื่นไฟฟ้าหัวใจและแบ่งข้อมูลออกเป็นชุดเรียนรู้และชุดทดสอบสัดส่วน 70% และ 30% ตามลำดับ และใช้ข้อมูลชุดเรียนรู้เพื่อหาพารามิเตอร์ที่ดีที่สุดด้วยการสุ่มชุดพารามิเตอร์ทั้งหมด 30 ชุดแล้วใช้ 10-Folds Cross Validation เลือกชุดพารามิเตอร์ที่ให้ค่าความแม่นยำสูงที่สุดไปทดสอบกับข้อมูลชุดทดสอบ (Testing Data) สามารถสรุปได้ว่า ทั้ง SAX และ BOSS มีประสิทธิภาพดีใกล้เคียงกัน

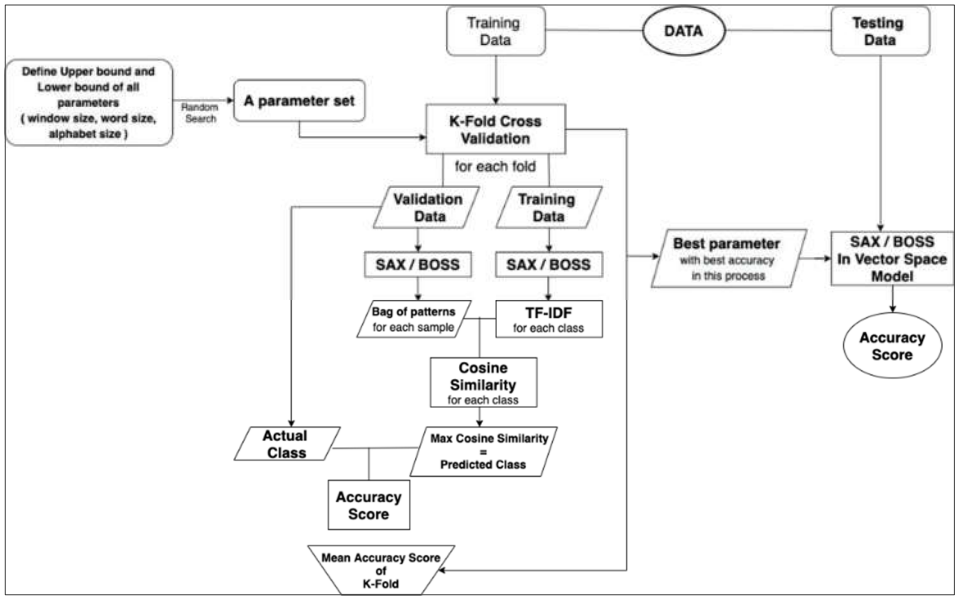
8.1 การเปรียบเทียบค่าความถูกต้อง (Accuracy)

สำหรับข้อมูลคลื่นไฟฟ้าหัวใจที่ไม่มีสัญญาณรบกวน ทั้งสองตัวแบบให้ค่าความถูกต้อง (Accuracy) อยู่ที่ประมาณ 99% แต่เมื่อเพิ่มสัญญาณรบกวนเข้าไปทำให้ค่าความถูกต้อง (Accuracy) ของทั้งสองตัวแบบลดลงเพียงเล็กน้อยแต่ยังคงทำงานได้ดีใกล้เคียงกัน มีค่าความถูกต้อง (Accuracy) ประมาณ 97-99%

The Accuracy Comparison

Noise Type	Level	Model: SAXVSM	Model: BOSSVS
None	0	0.99075	0.99075
baseline	25	0.98435	0.99075
	50	0.98791	0.99289
	100	0.99289	0.98009
composite	25	0.99004	0.99147
	50	0.99289	0.99004
	100	0.99218	0.98720
emg	25	0.99431	0.99004
	50	0.99147	0.99360
	100	0.98933	0.99218
powerline	25	0.98578	0.99147
	50	0.98933	0.98862
	100	0.97937	0.98933

รูปที่ 9. แสดง Heat Map เปรียบเทียบค่าถูกต้องระหว่าง SAXVSM และ BOSSVS



รูปที่ 8. แสดงกระบวนการทำงาน

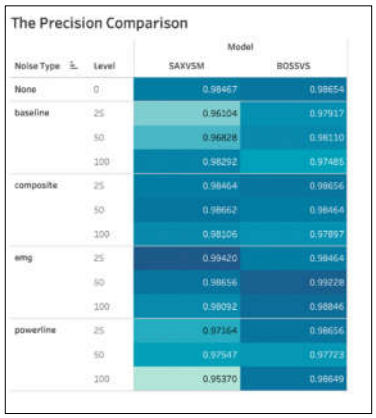
8.2 การเปรียบเทียบคะแนน F1 (F1-Score)

สำหรับข้อมูลคลื่นไฟฟ้าหัวใจที่ไม่มีสัญญาณรบกวน ทั้งสองตัวแบบให้คะแนน F1 (F1 Score) อยู่ที่ประมาณ 99% แต่เมื่อเพิ่มสัญญาณรบกวนเข้าไปทำให้คะแนน F1 (F1 Score) ทั้งสองตัวแบบลดลงเพียงเล็กน้อยแต่ยังคงทำงานได้ดีใกล้เคียงกัน มีคะแนน F1 (F1 Score) ประมาณ 97-99%

รูปที่ 10. แสดง Heat Map เปรียบเทียบคะแนน F1 ระหว่าง SAXVSM และ BOSSVS

8.3 การเปรียบเทียบค่าความแม่นยำ (Precision)

สำหรับข้อมูลคลื่นไฟฟ้าหัวใจที่ไม่มีสัญญาณรบกวน ทั้งสองตัวแบบให้ค่าความแม่นยำ (Precision) อยู่ที่ประมาณ 98% แต่เมื่อเพิ่มสัญญาณรบกวนเข้าไปทำให้ค่าความแม่นยำ (Precision) ทั้งสองตัวแบบลดลงเพียงเล็กน้อยแต่ยังคงทำงานได้ดีใกล้เคียงกัน มีค่าความแม่นยำ (Precision) ประมาณ 95-99%



รูปที่ 11. แสดง Heat Map เปรียบเทียบค่าความแม่นยำ ระหว่าง SAXVSM และ BOSSVS

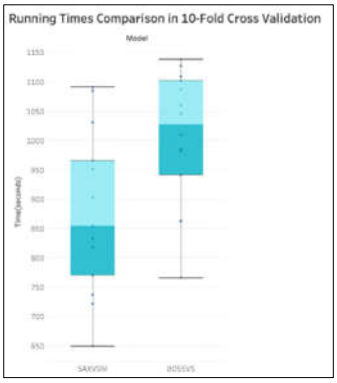
8.4 การเปรียบเทียบค่าความระลึก (Recall)

สำหรับข้อมูลคลื่นไฟฟ้าหัวใจที่ไม่มีสัญญาณรบกวน ทั้งสองตัวแบบให้ค่าความระลึก (Recall) อยู่ที่ประมาณ 99% แต่เมื่อเพิ่มสัญญาณรบกวนเข้าไปทำให้ค่าความระลึก (Recall) ทั้งสองตัวแบบลดลงเพียงเล็กน้อยแต่ยังคงทำงานได้ดีใกล้เคียงกัน มีค่าความระลึก (Recall) ประมาณ 97-100%

รูปที่ 12. แสดง Heat Map เปรียบเทียบค่าความระลึก ระหว่าง SAXVSM และ BOSSVS

8.5 การเปรียบเทียบเวลาในการสอนตัวแบบ

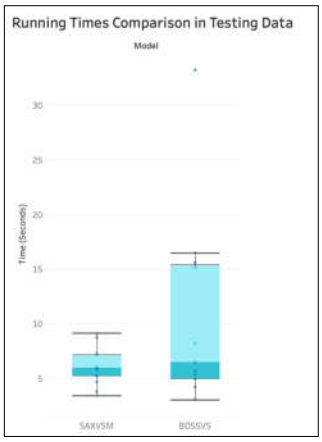
สำหรับการเปรียบเทียบเวลาที่ใช้ในขั้นตอนการสอนตัวแบบด้วย 10-Folds Cross Validation เมื่อทดสอบกับ SAXVSM และ BOSSVS เพื่อหาพารามิเตอร์ที่ดีที่สุดสำหรับข้อมูลแต่ละชุด จะเห็นว่า BOSSVS ใช้เวลาประมาณ 950-1,100 วินาที หรือ 15-18 นาที ต่อข้อมูล 1 ชุด ในขณะที่ SAXVSM ใช้เวลาประมาณ 650-950 วินาที หรือ 10-16 นาที ต่อข้อมูล 1 ชุด



รูปที่ 13. แสดง Boxplot เปรียบเทียบเวลาที่ใช้ในขั้นตอนการสอนตัวแบบ ระหว่าง SAXVSM และ BOSSVS

8.6 การเปรียบเทียบเวลาในการทดสอบตัวแบบ

สำหรับการเปรียบเทียบเวลาที่ใช้ในการทดสอบตัวแบบสำหรับ Testing Data ระหว่าง SAXVSM และ BOSSVS จะเห็นว่า BOSSVS ใช้เวลานานกว่า SAXVSM และมีการกระจายตัวมากกว่า อยู่ที่ประมาณ 5-15 วินาที ต่อข้อมูล 1 ชุด ส่วน SAXVSM ใช้เวลาอยู่ที่ประมาณ 5-8 วินาที ต่อข้อมูล 1 ชุด



รูปที่ 14. แสดง Box plot เปรียบเทียบเวลาการทดสอบตัวแบบระหว่าง SAXVSM และ BOSSVS

9. บทสรุปและการอภิปราย

การเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลอนุกรมเวลาระหว่าง SAXVSM และ BOSSVS โดยใช้ข้อมูลคลื่นไฟฟ้าฟ้วใจเป็นกรณีศึกษา และทำการเพิ่มสัญญาณรบกวนแบบ EMG Powerline Baseline และ Composite ที่ระดับ 25% 50% และ 100% สรุปได้ว่า โดยภาพรวมทั้ง 2 ตัวแบบมีประสิทธิภาพดีใกล้เคียงกัน ทั้งในแง่ของความถูกต้อง คะแนน F1 ค่าความแม่นยำ และค่าความระลึก โดยเฉลี่ยอยู่ที่ 98-99% สำหรับข้อมูลทั้ง 13 ชุด ทั้งข้อมูลคลื่นไฟฟ้าหัวใจที่ยังไม่ได้มีการเติมสัญญาณรบกวนหรือมีการเพิ่มสัญญาณรบกวนแบบต่าง ๆ อย่างไรก็ดี เมื่อเปรียบเทียบเวลาที่ใช้ในการประมวลผล เนื่องจาก SAXVSM มีกระบวนการที่ซับซ้อนน้อยกว่าจะใช้เวลาน้อยกว่า BOSSVS

สำหรับตัวแบบ SAXVSM เมื่อพิจารณาถึงค่าความถูกต้อง (Accuracy) คะแนน F1 (F1 Score) และค่าความแม่นยำ (Precision) ที่ทดสอบกับข้อมูลที่มีสัญญาณรบกวนแบบ Powerline ที่ระดับ 100% ทำให้ตัววัดประสิทธิภาพทั้ง 3 ตัวข้างต้นน้อยที่สุด ซึ่งสัญญาณรบกวนประเภทนี้เป็นคลื่นความถี่สูง เกิดจากการรบกวนของสนามแม่เหล็กไฟฟ้า ถูกจำลองมาจากฟังก์ชันคลื่นไซน์ (Sine Wave) ยิ่งที่ระดับสัญญาณรบกวนที่สูงขึ้น อาจมีส่วนทำให้ระยะขจัดที่แกว่งจากแนวเส้นฐาน (Amplitude) ผิดเพี้ยนไปจากรูปแบบของคลื่นไฟฟ้าหัวใจปกติ ทำให้ตัวแบบทำนายผิดพลาดได้ อย่างไรก็ดี เมื่อพิจารณาในส่วนของเวลาในการประมวลผลทั้งขั้นตอนการสอนและการทดสอบตัวแบบ SAXVSM รวดเร็วกว่า BOSSVS อย่างเห็นได้ชัด

สำหรับตัวแบบ BOSSVS เมื่อพิจารณาถึงค่าความถูกต้อง (Accuracy) คะแนน F1 (F1 Score) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ที่ทดสอบกับข้อมูลที่มีสัญญาณรบกวนแบบ Baseline ที่ระดับ 100% ทำให้ตัววัดประสิทธิภาพทั้ง 4 ตัวข้างต้นน้อยที่สุด ซึ่งสัญญาณรบกวนประเภทนี้เกิดขึ้นได้จากสาเหตุทางกายภาพ เช่น การเคลื่อนไหวของร่างกาย ความชื้นหรือเหงื่อที่ผิวหนัง การเคลื่อนที่ของอิเล็กโทรดหรือขั้วไฟฟ้าที่ติดบนผิวหนังและเชื่อมเข้ากับเครื่องบันทึกคลื่นไฟฟ้าหัวใจ ถึงแม้จะมีความถี่ต่ำแต่มักพบสัญญาณรบกวนประเภทนี้ได้ง่าย สัญญาณรบกวนประเภทนี้ถูกจำลองมาจากฟังก์ชันคลื่นไซน์ (Sine Wave) ยิ่งมีระดับสัญญาณรบกวนที่สูงขึ้น อาจมีส่วนทำให้ระยะขจัดที่แกว่งจากแนวเส้นฐาน (Amplitude) ผิดเพี้ยนไปจากรูปแบบของคลื่นไฟฟ้าหัวใจปกติ ทำให้ตัววัดประสิทธิภาพทั้ง 4 ค่าให้ผลออกมาน้อยที่สุด เมื่อเทียบกับสัญญาณรบกวนแบบอื่น ๆ เมื่อพิจารณาค่าความระลึกสำหรับสัญญาณรบกวนแบบ Baseline ที่ระดับ 50% พบว่าตัวแบบไม่มีการทำนายผิดพลาด แต่เมื่อระดับ 100% ค่าความระลึก ลดลงน้อยที่สุด เมื่อเทียบกับสัญญาณรบกวนแบบอื่น ๆ จึงเป็นข้อควรระวังหากตรวจจับสัญญาณรบกวนประเภทนี้ได้ ก็ควรใช้เทคนิคการกรองสัญญาณรบกวนประเภทนี้ออก เนื่องจากตัวแบบอาจจะยังทำงานได้ไม่ดีเท่าที่ควร

สำหรับเวลาที่ใช้ในการประมวลผลสำหรับตัวแบบ SAXVSM และ BOSSVS ในขั้นตอนการเลือกพารามิเตอร์โดยใช้ 10-Folds Cross Validation สำหรับข้อมูลในแต่ละชุด ยิ่งมีการเพิ่มสัญญาณรบกวนในระดับที่สูงขึ้นทั้ง SAXVSM และ BOSSVS จะใช้เวลาในการสอนตัวแบบน้อยลง เนื่องจากในขั้นตอนการสร้างเมทริกซ์น้ำหนักคำด้วยการหาความถี่ของคำและการผกผันความถี่ในเอกสาร (TF-IDF) ยิ่งข้อมูลที่มีระดับสัญญาณรบกวนสูงขึ้น ทำให้ตัวแบบเจอรูปแบบของสัญญาณรบกวนมาบดบังรูปแบบที่แท้จริงของคลื่นไฟฟ้าหัวใจ ทำให้ได้

จำนวนรูปแบบตัวอักษรระหว่างประเภทที่ปกติและประเภทผิดปกติซ้ำกันมากขึ้น และความหลากหลายของรูปแบบตัวอักษรน้อยลง จึงทำให้เมทริกซ์มีขนาดเล็กลง เมื่อไปคำนวณค่าความเหมือนโคไซน์ (Cosine Similarity) จึงทำได้รวดเร็วขึ้น

โดยสรุปจากการวิจัยนี้ ทั้งสองตัวแบบนี้ใช้เทคนิคการลดมิติของอนุกรมเวลาด้วยการแบ่งอนุกรมเวลาเป็นอนุกรมเวลาย่อย ๆ และพยายามแปลงให้เป็นลำดับของตัวอักษรที่สามารถอธิบายได้ด้วย Piecewise Aggregate Approximation (PAA) สำหรับ SAXและ Discrete Fourier Transform (DFT) สำหรับ BOSS จะได้ความสำคัญของรูปแบบตัวอักษรนั้น ๆ จึงต้องพิจารณาความถี่ของรูปแบบของอักษรนั้น แต่จะถูกลดทอนลงด้วยสัดส่วนของรูปแบบตัวอักษรนั้นที่อยู่ในประเภท (Class) อื่น ๆ ด้วย และหาค่าความเหมือนโคไซน์ เพื่อใช้ในการจำแนกประเภทข้อมูลต่อไป ผลการวิจัยพบว่าทั้ง 2 ตัวแบบเหมาะสำหรับการจำแนกประเภทข้อมูลอนุกรมเวลา และมีประสิทธิภาพดีใกล้เคียงกัน แต่ SAXVSM ใช้เวลาการประมวลผลน้อยกว่า

10. ข้อเสนอแนะ

งานวิจัยนี้ใช้ข้อมูลคลื่นไฟฟ้าหัวใจจากฐานข้อมูล Physionet เพื่อเปรียบเทียบประสิทธิภาพของตัวแบบจำแนกประเภทข้อมูลอนุกรมเวลา ซึ่งให้ผลการเปรียบเทียบประสิทธิภาพของทั้ง 2 วิธี ยังไม่ชัดเจนเท่าที่ควร อาจเพราะปริมาณข้อมูลที่นำมาศึกษาน้อยเกินไป ในการวิจัยครั้งต่อไปควรเพิ่มปริมาณข้อมูลที่ใช้ศึกษาให้มากขึ้น รวมทั้งศึกษาข้อมูลด้านอื่น ๆ เพิ่มเติมด้วย อีกทั้งงานวิจัยนี้ทำการเปรียบเทียบประสิทธิภาพของตัวแบบจำแนกประเภทข้อมูลอนุกรมเวลาด้วยวิธีการ 2 วิธี ได้แก่ SAXVSM และ BOSSVS เท่านั้น ในการวิจัยครั้งต่อไปอาจจะเพิ่มการเปรียบเทียบการจำแนกประเภทข้อมูลอนุกรมเวลาด้วยเทคนิคอื่น ๆ เพิ่มเติม

เอกสารอ้างอิง

[1] Thailand Online Hospital. "การตรวจคลื่นไฟฟ้าหัวใจ (Electrocardiography)," 2021. [Online]. Available: https://bit.ly/3EiRCUa. [Accessed April 15, 2021].

[2] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," *Data Mining and Knowledge Discovery,* vol. 15, no. 2, pp. 107-144, doi: 10.1007/s10618-007-0064-z, 2007.

[3] P. Senin and S. Malinchik, "Sax-vsm: Interpretable time series classification using sax and vector space model," in *2013 IEEE 13th international conference on data mining*, IEEE, pp. 1175-1180, 2013.

[4] P. Schäfer, "Bag-Of-SFA-Symbols in Vector Space (BOSS VS)," Zuse-Institut Berlin (ZIB), 2015.

[5] P. Schäfer, "Scalable time series similarity search for data analytics," doi: 10.18452/17338, 2015.

[6] P. Schäfer, "The BOSS is concerned with time series classification in the presence of noise," *Data Mining and Knowledge Discovery,* vol. 29, no. 6, pp. 1505-1530, doi: 10.1007/s10618-014-0377-7, 2014.

[7] Y. Kaya and H. Pehlivan, "Classification of premature ventricular contraction in ECG," *Int J Adv Comput Sci Appl,* vol. 6, no. 7, pp. 34-40, doi: 10.14569/IJACSA.2015.060706, 2015.

[8] K. M. Chang, "Arrhythmia ECG noise reduction by ensemble empirical mode decomposition," *Sensors (Basel),* vol. 10, no. 6, pp. 6063-80, doi: 10.3390/s100606063, 2010.

[9] P. Senin. "Z-normalization of time series," 2021. [Online]. Available: https://jmotif.github.io/saxvsm_site/morea/algorithm/znorm.html. [Accessed April 21, 2021].

[10] Souspace. "รู้จัก Discrete Fourier Transform และ Fast Fourier Transform," 2021. [Online]. Available: https://www.youtube.com/watch?v=a03-5mr5yn4. [Accessed April 21, 2021].

[11] V. V. Raghavan and S. M. Wong, "A critical analysis of vector space model for information retrieval," *Journal of the American Society for information Science,* vol. 37, no. 5, pp. 279-287, doi: 10.1002/(SICI)1097-4571(198609)37:5<279::AID-ASI1>3.0.CO;2-Q, 1986.

[12] N. Srikong. "Cosine similarity," 2021. [Online]. Available: https://medium.com/@srikong_n/cosine-similarity-f1f9a962ddc5 [Accessed April 21, 2021].