การรู้จำภาษามือไทยท่าเคลื่อนไหวด้วยโครงข่ายประสาทเทียมแบบวนกลับ
Dynamic Thai Sign Language Recognition using Recurrent Neural Network

พิพัฒน์พงศ์ ธรรมสิทธิ์ (Pipatpong Thammasit)* และชัยนันท์ สมพงษ์ (Chaiyanan Sompong)*

Received: January 2, 2024
Revised: October 1, 2024
Accepted: October 10, 2024

* ผู้นิพนธ์ประสานงาน: ชัยนันท์ สมพงษ์ (Chaiyanan Sompong) อีเมล: chaiyanan@snru.ac.th

DOI:10.14416/j.it.2026.v1.005

บทคัดย่อ

ภาษามือคือการสื่อสารด้วยการแสดงสัญลักษณ์ท่าทางมือ ซึ่งมีการแสดงท่าทางได้ตั้งแต่ระดับหัวจนถึงระดับเอว พร้อมทั้งยังมีการแสดงออกทางสีหน้าเพื่อสื่อถึงอารมณ์ของผู้พูด โดยมีงานวิจัยที่พยายามในการรู้จำภาษามือแบบมีการเคลื่อนไหวด้วยวิธีการเรียนรู้ของเครื่อง แต่ด้วยรูปแบบภาษามือแบบเคลื่อนไหวเป็นข้อมูลต่อเนื่องเชิงเวลา นอกจากนี้ตำแหน่งของมือและการแสดงออกสีหน้าเป็นองค์ประกอบที่จะทำให้การสื่อสารภาษามือมีความสมบูรณ์ ดังนั้นการพัฒนาวิธีการการรู้จำภาษามือจึงเป็นงานที่ยังคงท้าทาย งานวิจัยนี้มีวัตถุประสงค์ในการพัฒนาตัวแบบการรู้จำภาษามือไทยด้วยวิธีโครงข่ายประสาทเทียมแบบวนกลับ โดยข้อมูลนำเข้าเป็นคีย์พอยท์ที่สกัดเอาจุดเด่นของผู้สื่อสารภาษามือด้วยไลบารี MediaPipe ซึ่งประกอบไปด้วยข้อมูลสามชุด ได้แก่ มือทั้งสองข้าง ใบหน้า และการแสดงท่าทางที่เป็นพิกัด (x, y, z) รวม 1,662 คีย์พอยท์ จากนั้นนำชุดข้อมูลไปเรียนรู้ด้วยโครงข่ายประสาทเทียมแบบวนกลับสามแบบ ได้แก่ ประเภท ได้แก่ 1) หน่วยความจำระยะสั้นยาว (Long Short-Term Memory: LSTM) 2) ความจำระยะสั้น-ยาวแบบ 2 ทิศทาง (BiLSTM) 3) และหน่วยเกทแบบวนกลับ (Gated Recurrent Unit: GRU) ชุดข้อมูลที่ใช้ในการทดลองเป็นวิดีโอภาษามือไทยจากอาสาสมัครที่เป็นล่ามภาษามือ และผู้บกพร่องทางการได้ยิน ทั้งหมด 10 คำ จำนวน 1,000 วิดีโอ ผลการทดลองแสดงให้เห็นถึงความแม่นยำของวิธีการที่นำเสนอที่ 99% ด้วยโครงข่ายประสาทเทียมแบบวนกลับ แบบความจำระยะสั้น-ยาว และเกทแบบวนกลับ

คำสำคัญ: การเรียนรู้เชิงลึก หน่วยความจำระยะสั้นยาว หน่วยเกทแบบวนกลับ มีเดียไพพ์แลนมาร์ก

Abstract

Sign Language is a communication using a hand gesture that can pose on head to waist along with a facial emotion. There are numerous articles attempting to recognize dynamic sign language using machine learning. However, the dynamic sign language is a temporal continuous data. In addition, the positions of the hands and facial emotion are components that contribute to the completeness of sign language communication. Therefore, a sign language recognition methodology development is still challenge. This research aims to develop a Thai sign language recognition approach using recurrence neural network (RNN). The MediaPipe library applies to landmark extraction consisting of the hands, face and posture using by coordinate (x, y, z) totally 1,662 keypoint for RNN input. After that, these keypoints are learned by RNN technique consisting of long short-term Memory (LSTM) 2) gated recurrent unit (GRU) and 3) bi-direction LSTM (BiLSTM). The dataset consists of 10 words of Thai sign Language totally 1,000 videos that are established by volunteers sign language interpreters and hearing impaired. The experiment result demonstrates that an accuracy of the proposed method at 99% by LSTM and GRU.

Keywords: Deep learning, Long short-term Memory, Gated recurrent unit, MediaPipe's landmark.

* *สาขาวิชาคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสกลนคร*
* *Computer Department, Faculty of Science and Technology, Sakonnahkon Rajabhat University.*

1. บทนำ

ภาษามือคือภาษาสัญลักษณ์ที่ใช้สื่อสารสำหรับผู้พิการทางการได้ยิน โดยมีการสื่อสารด้วยสัญลักษณ์ท่าทางมือ การแสดงสีหน้า และกิริยาท่าทางในการประกอบในการสื่อความหมาย [1], [2] เพื่อถ่ายทอดอารมณ์แทนการพูด ภาษามือของแต่ละชาติมีความหมายแต่ต่างกันเช่นเดียวกับภาษาพูด ซึ่งแตกต่างกันตามขนบธรรมเนียม ประเพณี วัฒนธรรม และลักษณะภูมิศาสตร์ เช่น ภาษามือจีน ภาษามืออเมริกัน และภาษามือไทย เป็นต้น ซึ่งภาษามือเป็นภาษาที่ผู้พิการทางการได้ยินตกลง และยอมรับกันแล้วว่าเป็นภาษาหนึ่งสำหรับการติดต่อสื่อความหมายระหว่างผู้พิการทางการได้ยิน รวมทั้งกับคนปกติด้วย

การเรียนรู้ของเครื่อง (Machine Learning) เป็นส่วนย่อยหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence: AI) ซึ่งเป็นการเรียนรู้จากข้อมูลที่เคยเกิดขึ้นก่อนเพื่อหารูปแบบของข้อมูลที่นำไปสู่การจำแนก (Classification) การทำนาย (Prediction) การวินิจฉัย (Diagnosis) และการจัดกลุ่มข้อมูล (Cluster) ซึ่งเป็นเครื่องมือที่ช่วยแก้ปัญหาให้กับมนุษย์เมื่อเจอข้อมูลที่ไม่ทราบมาก่อน เช่น โครงข่ายประสาทเทียม (Neural Network) และการเรียนรู้เชิงลึก (Deep Learning) เป็นเทคนิคหนึ่งที่ให้ประสิทธิภาพการจำแนกข้อมูลในระดับสูงมาก โดยมีข้อได้เปรียบกว่าเทคนิคอื่นในด้านข้อมูลนำเข้าที่ไม่ต้องคำนึงถึงคุณภาพและปริมาณอย่างไรก็ตามเทคนิคนี้จำเป็นที่จะต้องการปริมาณข้อมูลจำนวนมากในการเรียนรู้

ข้อมูลชิงเวลาคือข้อมูลที่ส่งออกมาตามช่วงเวลาอย่างต่อเนื่อง เช่น เสียงพูด ภาพวิดีโอ เซนเซอร์วัดความเอียง/ความเร่งในอุปกรณ์พกพาต่าง ๆ หรือการสื่อสารในภาษาธรรมชาติ ดังนั้นการประมวลผลข้อมูลแบบนี้จะมีคุณลักษณะที่ข้อมูลในเวลาก่อนหน้าจะส่งผลกระทบกับข้อมูลลำดับถัดไป สำหรับการเรียนรู้เชิงลึกกับข้อมูลเชิงเวลาส่วนมากจะใช้การเรียนรู้แบบวนกลับ (Recurrent Neural Networks: RNN) เป็นการนำผลลัพธ์ที่ได้จากการคำนวณที่เวลา t-1 กลับมาใช้เป็นข้อมูลขาเข้าอีกครั้งกับข้อมูลที่เวลา t โดยที่งานวิจัยส่วนมากจะนำวิธีการนี้ไปพยากรณ์เหตุการณ์ที่กำลังจะเกิด เช่น การสร้างโมเดลในการจำแนกเหตุการณ์อุบัติเหตุจากภาพวิดีโอกล้องหน้ารถด้วย Convolution Neural Network (CNN) ร่วมกับ RNN [3] ซึ่งให้ความถูกต้องถึงร้อยละ 92 ในขณะที่การติดตามคนพร้อมกันหลายคน (Multi-people Tracking) จากภาพวิดีโอ [4] CNN ถูกนำมาใช้ในการค้นหาคนในแต่ละเฟรมของภาพวิดีโอเคลื่อน เพื่อแก้ปัญหาความสัมพันธ์ของข้อมูล (Data Association Problem) ในแต่ละช่วงเวลาหรือเฟรม ซึ่งจะส่งผลต่อประสิทธิภาพการติดตามตัวคน (People Tracking) จึงได้นำ RNN เข้ามาช่วยให้เกิดประสิทธิภาพที่ดีขึ้น จากงานวิจัยดังกล่าว จะเห็นได้ว่า RNN มีประสิทธิภาพในการจำแนกข้อมูลที่มีความสัมพันธ์เชิงเวลา

การเรียนรู้เชิงลึกถูกนำมาใช้ในการรู้จำภาษามือไทย โดยใช้เทคนิค CNN โดยการใช้ภาพในแต่ละเฟรมจากชุดข้อมูลวิดีโอภาษามือหรือมีการใช้กล้อง Kinect ที่สามารถจำแนกบุคคลด้วยภาพร่วมกับกล้องจับความร้อนแล้วนำภาพแต่ละเฟรม [5] เข้ามาเรียนรู้ท่าทางภาษามือไทย แม้ว่าผลการรู้จำจะได้รับความแม่นยำมากกว่าร้อยละ 90 แต่ใช้กับจำนวนท่าภาษามือไม่มากนัก และนอกจากนี้ข้อมูลที่อินพุตมีจำนวนมากการประมวลผลจึงมีความซับซ้อนมาก อีกทั้งยังมีปัญหาความสัมพันธ์ของข้อมูลในแต่ละเฟรม ซึ่งจะทำให้เกิดความผิดพลาดในขณะที่จำแนกภาษามือแบบต่อเนื่องหลายท่าติดต่อกัน

เพื่อแก้ปัญหาดังกล่าวได้นำเสนอการรู้จำภาษามือที่เป็นภาษาไทยแบบเคลื่อนไหวโดยการใช้ MidiaPipe [6], [7] สกัดคีย์พอยท์ (Key Point) ที่เป็นจุดสังเกตของมือ (Hand Landmark) จากวิดีโอท่าทางภาษามือ และใช้แลนมาร์กของมือจำนวน 42 จุด เป็นข้อมูลนำเข้าเพื่อสร้างโมเดลสำหรับการรู้จำท่าทางภาษามือ โดยใช้ RNN ทั้ง 3 ประเภท ได้แก่ 1) หน่วยความจำระยะสั้นยาว (Long Short-Term Memory: LSTM) 2) ความจำระยะสั้น-ยาวแบบ 2 ทิศทาง (BiLSTM) 3) และหน่วยเกทแบบวนกลับ (Gated Recurrent Unit: GRU) ซึ่งให้แม่นยำมากกว่า 90% เมื่อเปรียบเทียบกับการรู้จำภาษามืออังกฤษ [8] ที่นำ MediaPipe สกัดจุดเด่นของร่างกาย ใบหน้า และมือ โดยมีคีย์พอยท์ที่ไม่รวมใบหน้า 258 จุด และแบบรวมคีย์พอยท์บนใบหน้าด้วยรวม 1,662 จุด เป็นอินพุตให้กับ RNN กับท่าทางภาษามือที่ใช้ 10 ท่า บนชุดข้อมูล DSL-10-Dataset ซึ่งให้ความแม่นยำถึง 100% ด้วยโมเดล GRU

จากประสิทธิภาพของ RNN ที่นำมาใช้ในการรู้จำภาษามืออังกฤษ ดังนั้นบทความนี้ จึงนำเทคนิคการรู้จำด้วย RNN และการใช้คีย์พอยท์ด้วย MediaPipe เพื่อพัฒนาระบบการรู้จำท่าทางภาษามือไทยแบบเรียลไทม์ โดยใช้การแสดงท่าทางของร่างกาย กับการแสดงออกทางสีหน้า เพื่อนำไปสู่การพัฒนา

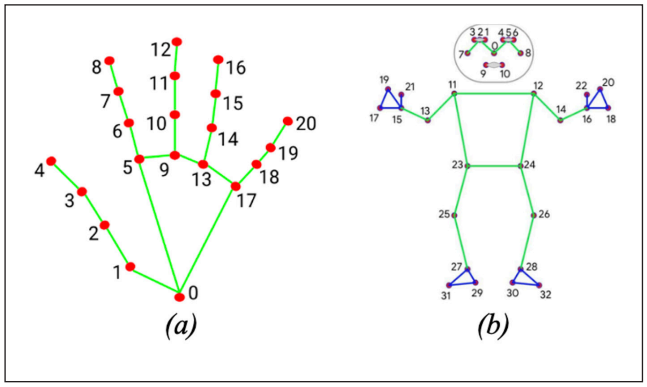
ให้สามารถใช้งานได้จริง โดยผู้วิจัยได้ใช้ชุดข้อมูลในการทดสอบจากอาสาสมัครของศูนย์บริการสนับสนุนนักศึกษาพิการระดับอุดมศึกษา มหาวิทยาลัยราชภัฏสกลนคร (Disability Support Services: DSS) มีท่าทางภาษามือ 10 ท่า แบบเคลื่อนไหว จากที่กล่าวมาผู้วิจัยได้จะพัฒนาระบบการรู้จำภาษามือไทยและท่าทางด้วยเทคนิคโครงค่ายประสาทเทียมแบบวนกลับ (RNN) ที่เป็นคำที่ใช้ในชีวิตประจำวัน เพื่อใช้ในการแปลภาษามือไทยของผู้พิการทำให้สามารถเข้าใจความหมายที่ต้องการจะสื่อได้

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 มีเดียไพพ์ (MediaPipe)

เพื่อขั้นในการพัฒนาระบบการรู้จำภาษามือ ผู้วิจัยจึงนำเทคโนโลยีของ MediaPipe ที่ใช้สำหรับการตรวจจับท่าทาง (Pose) มือ (Hand) และใบหน้า (Face) ของมนุษย์ได้ในเวลาเดียวกัน ซึ่งจะทำการสร้างคีย์พอยท์ของจุดสำคัญบนร่างกายหรือแลนด์มาร์ก (Landmark) ดังภาพที่ 1(a) แสดงจุดแลนด์มาร์กบนมือจำนวน 21 จุดและภาพที่ 1(b) แสดงจุดแลนด์มาร์กของท่าทางของมนุษย์จำนวน 33 จุด

มีงานวิจัยจำนวนมากที่ใช้ MediaPipe เป็นข้อมูลนำเข้าสำหรับการรู้จำท่าทางแทนการใช้ภาพหรือเซนเซอร์ต่าง ๆ เช่น การติดตามอัตราการเต้นของหัวใจและอัตราการหายใจแบบเรียลไทม์ผ่านวิดีโอ [9] การรู้จำอารมณ์จากสีหน้ากับภาพวิดีโอระยะไกล [10] หรือการรู้จำภาษามือ [6] - [8], [11] - [14] เป็นต้น ซึ่งจากผลการทดสอบพบว่าการใช้ MediaPipe สกัดจุดสำคัญของร่างกายเป็นอินพุตเข้าสู่การรู้จำด้วยเรียนรู้ของเครื่องจักร ให้ความแม่นยำในระดับที่มากกว่า 90%



ภาพที่ 1 *การตรวจจับท่าทางด้วย MediaPipe (a) แสดงจุดสำคัญบนมือ (b) แสดงจุดสำคัญของท่าทางมนุษย์ [11]*

2.2 โครงข่ายประสาทเทียมแบบวนกลับ (RNN)

RNN ถูกนำมาใช้ในการรู้จำกับอินพุตที่มีลักษณะต่อเนื่องเชิงเวลา มีหลักการทำงานคือ นำผลลัพธ์ที่ได้จากการคำนวณย้อนกลับมาใช้เป็นข้อมูลขาเข้าอีกครั้ง ซึ่งมีประโยชน์อย่างมากในข้อมูลที่มีความต่อเนื่อง เช่น ข้อมูลเสียง ข้อความ หรือวิดีโอ เป็นต้น ส่วนใหญ่ RNN ถูกออกแบบมาเพื่อแก้ปัญหาสำหรับงานที่มีข้อมูลที่มีลำดับ เช่น การพยากรณ์การเกิดการระบาดของ COVID-19 [15] หรือ ใช้ในการทำนายข้อมูลวิดีโอด้วยการใช้ร่วมกับ CNN [3], [4] สกัดข้อมูล ณ ช่วงเวลานั้น ๆ แล้วส่งให้ RNN ประมวลผลต่อ เป็นต้น โดย RNN ใช้หลักการนำสถานะภายในของโมเดล กลับมาเป็นข้อมูลเข้าใหม่คู่กับข้อมูลเข้าแบบปกติ เรียกว่า สถานะซ่อน (Hidden State) หรือสถานะภายใน (Internal State) ช่วยให้โมเดลรู้จำรูปแบบของลำดับข้อมูลนำเข้าที่ขึ้นอยู่กับข้อมูลก่อนหน้าได้ [15] แสดงดังภาพที่ 2(a) ในแต่ละโหนดของ RNN จะมีข้อมูลเข้าสองอย่าง ได้แก่ 1) ข้อมูลเข้า ณ เวลา t (x_t) และ 2) ผลลัพธ์ที่ได้จากการคำนวณเวลาก่อนหน้า (h_{t-1}) ซึ่งทั้งสองข้อมูลจะถูกนำมารวมเข้าด้วยกันและออกผลลัพธ์มาเป็นสองทาง คือ y_t เป็นผลลัพธ์ที่ออกมาที่เวลา t ดังสมการที่ (1) โดยที่ W_y คือค่าน้ำหนักของสถานะปัจจุบัน h_t c คือค่าไบแอสสำหรับเอาท์พุต และ σ คือฟังก์ชันการถ่ายโอน เช่น ReLU หรือ tanh ในส่วนของ h_t เป็นข้อมูลที่จะเป็นข้อมูลขาเข้าในเวลา t+1 ในสมการที่ (2) โดยที่ W_x คือค่าน้ำหนักของอินพุตปัจจุบัน x_t และ W คือค่าน้ำหนักของอินพุตกับสถานะก่อนหน้า h_{t-1} ข้อดีของ RNN คือ มีการใช้ข้อมูลก่อนหน้าในการทำนายสิ่งที่อาจจะเกิดขึ้นในอนาคต ซึ่งหมายถึงอะไรที่เคยเกิดขึ้นในอดีตย่อมส่งผลต่อเหตุการณ์ที่จะเกิดขึ้นในอนาคตด้วย แม้ RNN จะมีข้อดีในการทำงานของข้อมูลที่มีความต่อเนื่อง แต่ข้อเสียของ RNN คือ สามารถดูย้อนกลับได้แค่เพียงในช่วงระยะเวลาสั้น ๆ เท่านั้น โดยปัญหาหลักของ RNN จะเกิดขึ้นการฝึกมีจำนวนที่ต้องการเรียนรู้ย้อนกลับไปในลำดับที่ก่อนหน้าที่ยาวนานมาก จะทำให้การคำนวณค่าเกรเดียนท์เริ่มน้อยลงและเข้าใกล้ศูนย์จนเกิดปัญหาการสูญเสียของเกรเดียนท์ (Vanishing Gradient Problem) ที่นำมาใช้ในการปรับน้ำหนัก ซึ่งอาจเกิดได้จาก ซึ่งปัญหานี้ถูกแก้ไขโดยใช้เกทแบบวนกลับ (Gated Recurrent Unit: GRU) และหน่วยความจำระยะสั้นยาว (Long Short-Term Memory: LSTM)

$$y_t = \sigma(h_t W_y + c) \quad (1)$$

$$h_t = \sigma(x_t W_x + h_{t-1} W) \quad (2)$$

จากปัญหาการสูญเสียของค่าเกรเดียนท์ LSTM [15], [16] จึงได้ถูกนำเสนอเพื่อปรับปรุงการทำงานของ RNN ให้สามารถคำนวณค่าสถานะโดยมีการเรียนรู้ย้อนกลับได้ยาวนานขึ้นมีโครงสร้างที่มีส่วนประกอบ 3 ส่วน ได้แก่ 1) ประตูลืม (Forget Gate) ดังสมการที่ (3) ทำหน้าที่ในการตัดสินใจว่าควรลืมข้อมูลในสถานะปัจจุบัน 2) ประตูอินพุต (Input Gate) ดังสมการที่ (4) และ (5) ทำหน้าที่ในการรับอินพุตและคำนวณข้อมูลใหม่ และ 3) ประตูเอาท์พุต (Output Gate) ดังสมการที่ (6) - (8) จะคำนวณค่าเอาท์พุต (C_t และ h_t) และที่จะส่งไปยังสถานะต่อไป ซึ่งโดยแสดงโครงสร้างของ LSTM ดังภาพที่ 2(b) และการคำนวณในประตูต่าง ๆ

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (5)$$

$$C_t = f_f \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (6)$$

$$y_t = \sigma(W_y \cdot [h_{t-1}, x_t] + b_y) \quad (7)$$

$$h_t = y_t \cdot tanh(C_t) \quad (8)$$

ในขณะที่ LSTM มีโครงสร้างภายในที่ซับซ้อนเนื่องจากต้องผ่านการคำนวณทั้ง 3 ประตู ทำให้ใช้เวลาในการฝึกฝนโมเดลนานขึ้น จึงมีการนำเสนอวิธี GRU [16] - [18] โดยปรับปรุง LSTM ให้เหลือเพียง 2 ประตู ดังภาพที่ 2(c) ได้แก่ 1) ประตูรีเซต (Reset Gate) ดังสมการที่ (9) ทำหน้าที่คล้ายประตูลืมของ LSTM โดยตัดสินใจที่จะลืมข้อมูลเก่าจากสถานะก่อนหน้าผ่านฟังก์ชันการถ่ายโอน (σ) และ 2) ประตูอัพเดต (Update Gate) ดังสมการที่ (10) - (12) ทำหน้าที่คำนวณว่าจะมีการปรับปรุงข้อมูลใหม่ (h_t) กับข้อมูลเก่าจากสถานะก่อนหน้า ($\tilde{h}_t$)

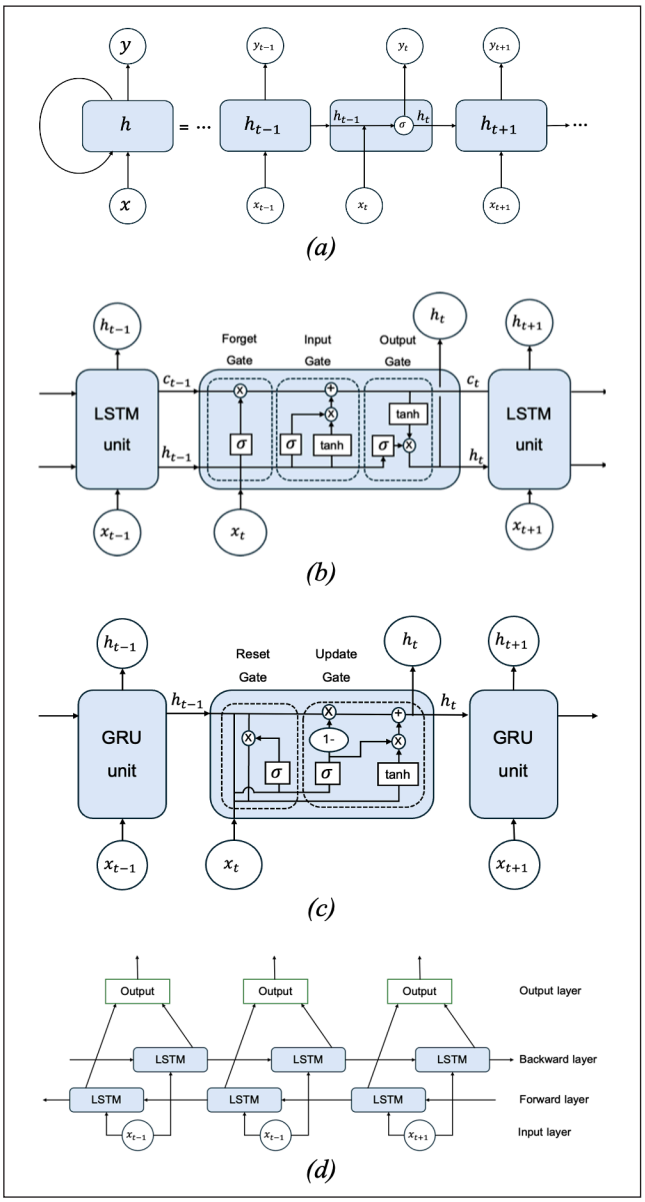
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (9)$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (10)$$

$$\tilde{h}_t = tanh(W_h \cdot [r_t * h_{t-1}, x_t]) \quad (11)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (12)$$

จากโครงสร้างของ LSTM จะเห็นได้ว่าสามารถเรียนรู้ข้อมูลที่ต้องการในการทำนายสิ่งที่จะเกิดขึ้น จึงมีการนำเสนอวิธี BiLSTM [19] เพื่อจับความสัมพันธ์ข้อมูลที่จะเกิดในอนาคต (Forward LSTM) กับข้อมูลที่เคยเกิดขึ้นในอดีตโดยใช้ (Backward LSTM) โดยมีโครงสร้างดังภาพที่ 2(d) ซึ่งวิธีการนี้มีความเหมาะสมกับข้อมูลที่ต้องการบริบทก่อนหน้าและบริบทที่กำลังจะเกิดขึ้น เช่น ข้อมูลภาษาสำหรับการสร้างข้อความถามตอบ ข้อมูลจากเซนเซอร์สำหรับการพยากรณ์พฤติกรรมมนุษย์ หรือข้อมูลภาพวิดีโอสำหรับการทำนายพฤติกรรม เป็นต้น



ภาพที่ 2 *การทำงานของโครงข่ายประสาทเทียมแบบวนกลับ a) RNN แบบทั่วไป b) LSTM c) GRU และ d) BiLSTM*

2.3 งานวิจัยที่เกี่ยวข้อง

การรู้จำภาษามือมีหลายงานวิจัยที่สามารถจดจำคำต่าง ๆ ไม่ว่าจะเป็นคำที่ใช้ท่าทางมือที่หยุดนิ่ง เช่น ตัวเลข หรือตัวอักษร [6], [11] เป็นต้น หรือภาษามือที่มีการเคลื่อนไหว (Dynamic) ซึ่งเป็นคำที่แสดงถึงกิริยาที่ใช้ในชีวิตประจำวัน [5], [7] - [8], [14] โดยเป็นคำง่าย ๆ เช่น สวัสดี ขอบคุณ หรือ ไม่สบาย เป็นต้น

การรู้จำภาษามือไทย [20] นำเสนอวิธีการประมวลผลภาพดิจิทัลในการรู้จำภาษามือที่เป็นตัวอักษรภาษาอังกฤษ โดยใช้วิธีการแยกสีผิวเพื่อค้นหารูปร่างของมือแล้วนำวิธีการจับคู่รูปแบบของสัญลักษณ์ภาษามือ (Template Matching) และใช้ค่าสัมประสิทธิสหสัมพันธ์ (Correlation Coefficient) ในการตัดสินว่าเป็นสัญลักษณ์ใด โดยผลการทดสอบพบว่ามีความถูกต้องที่ 76.15% ในส่วนของการใช้อุปกรณ์เสริม [5] ได้นำเสนอรู้จำภาษามือจำนวนสี่ท่าโดยการใช้กล้อง Kinect ในการสกัดท่าร่างภาษามือแบบสามมิติ โดยมีระดับความลึกจากเซนเซอร์วัดระยะของกล้อง Kinect และภาพสองมิติที่แปลงเป็นภาพขาวดำ แล้วนำภาพมาแบ่งส่วนของออกเป็นเก้าส่วนรวมกับมิติความลึกอีก 7 ส่วน และขนาดความก้าวสูงของท่าภาษามืออีก 2 ส่วน รวม 19 มิติ เพื่อนำเข้าสู่ Backpropagation ANN ผลลัพธ์ที่ได้มีความแม่นยำที่ 84% มาช่วยในขณะที่ [14] ใช้เพียงแค่ภาพในระนาบสองมิติและนำวิธีการ CNN ในการรู้จำภาษามือไทยโดยผลการทดสอบกับภาษามือจำนวนสามท่าให้ความแม่นยำที่ 93% จะเห็นว่าการเรียนรู้เชิงลึกหลาย ๆ วิธีมีประสิทธิภาพในการรู้จำค่อนข้างสูง แต่กรณีข้อมูลที่มีลักษณะเชิงเวลาแล้ว การนำข้อมูลในช่วงเวลาก่อนหน้ามารวมกับข้อมูลปัจจุบันจะสามารเพิ่มความแม่นยำของตัวแบบการเรียนรู้เชิงลึกได้ [3] - [4] จึงนำเสนอวิธีการนำ CNN ในการสกัดข้อมูล ณ ช่วงเวลานั้น ๆ แล้วส่งให้ RNN ทำการรู้จำกิจกรรมที่เกิดขึ้นในภาพรวมของหนึ่งช่วงเวลา

เพื่อลดความซับซ้อนและจำนวนอินพุต MediaPipe ถูกนำมาสกัดจุดแลนมาร์กของส่วนต่าง ๆ ของร่างกายแบบสามมิติ โดย [11] ได้นำเสนอภาษามืออัสสัมในการนับเลข 1 ถึง 9 โดยการใช้ MediaPipe สกัดจุดแลนมาร์กของมือร่วมกับ RNN ผลลัพธ์ที่ได้มีความแม่นยำถึง 99% ขณะที่ [13] นำเสนอการแปลภาษามือที่เป็นตัวอักษรภาษาไทย ภาษาอังกฤษและตัวเลข รวมถึงคำในภาษาอังกฤษง่าย ๆ เช่น Left, Forward หรือ Like เป็นต้น ซึ่งใช้วิธีการสกัดจุดแลนมาร์กด้วย MediaPipe และใช้กฎในการระบุตำแหน่งของจุดแลนด์มาร์กในแกน (x, y, z) เป็นการรู้จำภาษามือซึ่งทดสอบให้ความถูกต้องที่ 95.39% แต่พบว่าจำเป็นต้องว่าตำแหน่งมือให้ถูกต้องตามกฎที่ตั้งไว้ด้วย ในส่วนของภาษามืออังกฤษ [8] นำเสนอวิธีการรู้จำภาษามือแบบเคลื่อนไหว โดยการสกัดลักษณะเด่นของมือ ลำตัว และใบหน้าด้วยไลบารี MediaPipe โดยทำการเปรียบเทียบระหว่างใช้ใบหน้าที่มีข้อมูล 1,662 มิติ กับไม่ใช้ที่มีข้อมูล 258 มิติ จากภาษามือแบบเคลื่อนไหว 10 ท่า เข้าสู่การรู้จำด้วย RNN พบว่าให้ความแม่นยำถึง 99% ในทั้งสองกรณี ขณะที่ภาษามือไทย [6] ใช้วิธีการเดียวกันกับ [8] ในการรู้จำภาษามือท่านิ่ง เป็นตัวเลข 1 ถึง 5 ให้ความแม่นยำที่ 97% ด้วย RNN แบบ LSTM ในส่วนของ [7] มีการพัฒนาให้สามารถรู้จำภาษามือไทยจำนวน 10 ท่า ด้วยการใช้แลนด์มาร์กร่วมกับ RNN โดยเป็นท่านิ่ง 3 ท่า (ตัวเลข) และท่าเคลื่อนไหว 7 ท่า โดยมีความแม่นยำที่ 91% ด้วย RNN แบบ BiLSTM ซึ่งเมื่อเปรียบเทียบงานวิจัยต่าง ๆ ที่เกี่ยวข้องสามารถแสดงให้เห็นถึงประสิทธิภาพของวิธีการที่แตกต่างกันได้ดังตารางที่ 1 จากผลการทดสอบจะสังเกตได้ว่าการลดมิติข้อมูลในการนำเข้า RNN ด้วย MediaPipe ช่วยให้การรู้จำภาษามือแบบเคลื่อนไหวมีประสิทธิภาพที่สูงขึ้น ดังนั้นวิธีการนี้จึงถูกนำมาใช้ในการรู้จำภาษามือไทยท่าเคลื่อนไหวในงานวิจัยน

ตารางที่ 1 *ตารางเปรียบเทียบงานวิจัยที่เกี่ยวข้อง*

Authors	Sign Language	Approach	Accuracy (%)
Tatiyavoranun [5]	ไทย 4 ท่า เคลื่อนไหว	Kinect + Backpropagtion ANN	84
Chaikaew [6]	ไทย ตัวเลข 5 ท่านิ่ง	MediaPipe + RNN	97
Damrongekarun [7]	ไทย 7 ท่า เคลื่อนไหว และตัวเลข 3 ท่านิ่ง	MediaPipe + RNN	91
Gerges [8]	อังกฤษ 10 ท่า เคลื่อนไหว	MediaPipe + RNN	100
Bora [11]	อัสสัมตัวเลข 9 ท่านิ่ง	MediaPipe + RNN	99
Gedkhaw [14]	ไทย 3 ท่า เคลื่อนไหว	CNN	93

3. วิธีการดำเนินงานวิจัย

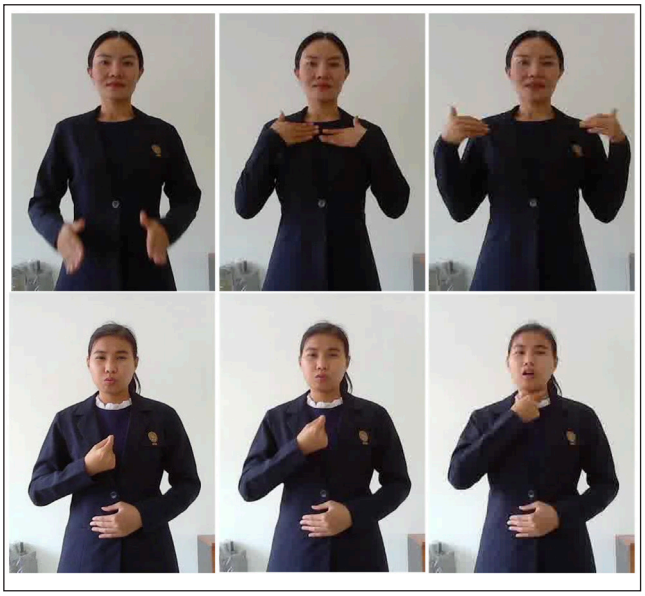
การพัฒนาระบบการรู้จำภาษามือไทยมีขั้นตอนหลักอยู่ 3 ขั้นตอน ได้แก่ 1) การเตรียมข้อมูล 2) การสร้างโมเดลการรู้จำภาษามือด้วย RNN 3) การทดสอบและวัดประสิทธิภาพ แสดงภาพรวมของงานวิจัยดังภาพที่ 3

ตารางที่ 2 *คำศัพท์ภาษามือและความหมาย*

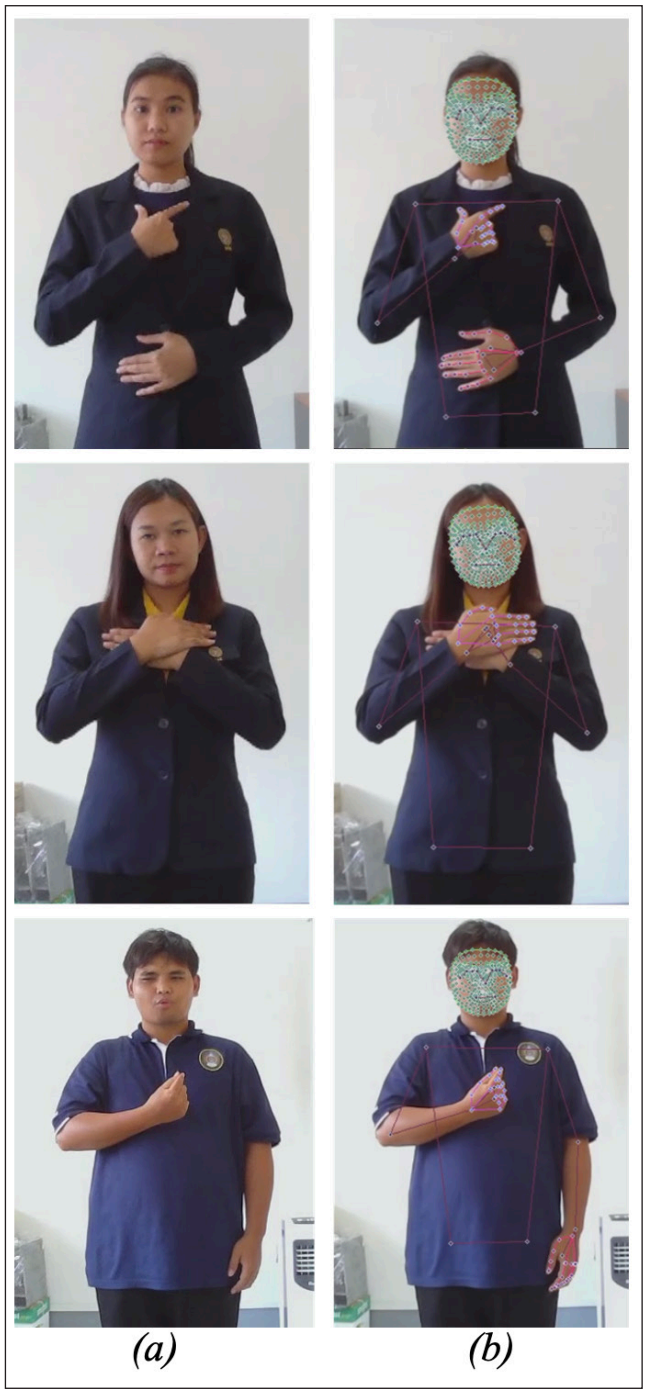
ภาษาไทย	ภาษาอังกฤษ	ความหมาย
ขอบคุณ	Thank You	กล่าวแสดงความรู้สึกถึงบุญคุณหรือกล่าวเมื่อได้รับความช่วยเหลือ
ขอโทษ	Sorry	ขออภัยเมื่อได้ทำผิดพลาดอย่างใดอย่างหนึ่ง
ไม่เป็นไร	That is OK	คำแสดงความรู้สึกที่ไม่ได้ถือโทษหรือโกรธเคืองใด ๆ เพื่อให้ผู้ฟังรู้สึกดีขึ้นหรือไม่ต้องรู้สึกผิด
สบายดี	Fine	สภาวะปกติของทั้งร่างกายและจิตใจ ร่างกายไม่เจ็บป่วย รวมทั้งอารมณ์ดี มีความสุข ไม่มีอะไรให้กังวล
ชอบ	Like	พอใจ แสดงอาการพึงพอใจ
รัก	Love	มีใจผูกพันอย่างมาก
ไม่สบาย	Sick	สภาวะที่ร่างกายและจิตใจไม่ปกติ หรือเกิดอาการป่วย
สวัสดี	Hello	ใช้สำหรับการทักทายผู้คน
ฉัน	I (Am)	ใช้สำหรับการเรียกแทนตัวเอง
คุณ	You	ใช้สำหรับเรียกแทนผู้ที่เราพูดด้วย

3.1 การเตรียมชุดข้อมูลภาษามือไทย

การเตรียมข้อมูลสำหรับการสร้างระบบรู้จำท่าทางภาษามือไทย (Thai Sign Language: TSL) ในงานวิจัยนี้เก็บข้อมูลเป็นวิดีโอภาษามือไทยจำนวน 10 คำ (TSL10) โดยคำที่ใช้ในชีวิตประจำวัน และเพื่อใช้ในการเปรียบเทียบวิธีการกับ [6] และ [7] ผู้วิจัยจึงได้กำหนดคำศัพท์ที่ใช้การรู้จำได้แก่ ขอบคุณ ขอโทษ ไม่เป็นไร สบายดี ชอบ รัก ไม่สบาย สวัสดี ฉัน และคุณ ดังตารางที่ 2 และแสดงตัวอย่างของท่าเคลื่อนไหวภาษามือไทยแสดงดังภาพที่ 3 โดยแถวที่หนึ่งเป็นการเคลื่อนไหวต่อเนื่องของคำว่าสบายดีที่มีการเคลื่อนไหวช่วงท่อนบนของร่างกาย และแถวที่สองคือคำว่าหิวที่มือทั้งสองข้างอยู่ต่างตำแหน่งกัน อีกทั้งยังมีการแสดงออกทางสีหน้าอีกด้วย โดยชุดข้อมูลที่ในในบทความนี้แบ่งเป็นสองชุด คือ 1) ชุดข้อมูล TSL10-Train สำหรับการสร้างโมเดล จำนวน 900 คลิป จากอาสาสมัครเจ้าหน้าที่ผู้เชี่ยวชาญภาษามือประจำศูนย์ DSS และนักศึกษาบกพร่องทางการได้ยิน จำนวน 5 คน และ 2) ชุดข้อมูล TSL10-Test โดยนักศึกษาบกพร่องทางการได้ยิน จำนวน 5 คน รวม 100 คลิป โดยที่ความยาวของวิดีโอภาษามือ TSL10-dataset อยู่ที่ 1 วินาที ขนาด 30 เฟรมต่อวินาที ความละเอียดที่ 640x480 จุดภาพเท่ากันทั้ง 1,000 คลิป เพื่อเป็นข้อมูลในโหนดนำเข้าที่เท่ากันทุกคลิป หลังจากนั้นนำวิดีโอไปทำการสกัดหาจุดแลนด์มาร์กบริเวณใบหน้า มือ และร่างกายด้วย MediaPipe รวม 543 จุด ดังตัวอย่างในภาพที่ 4 ซึ่งแต่ละจุดแลนด์มาร์กถูกจัดเก็บในรูปแบบคีย์พอยท์ x, y และ z ดังนั้นจากจุดแลนด์มาร์กของบริเวณที่สนใจจะได้รับชุดข้อมูลทั้งหมด 1,662 มิติข้อมูล และแต่ละวิดีโอจะมีข้อมูลประมาณ 30 แถว



ภาพที่ 3 *ตัวอย่างท่าเคลื่อนไหวภาษามือไทย TLS10 Dataset แถวแรกคือ "สบายดี" และแถวที่สองคือ "หิว"*



ภาพที่ 4 *ตัวอย่าง TLS10 Dataset (a) ภาพวิดีโอต้นฉบับ (b) ภาพการสกัดคีย์พอยท์ด้วย MediaPipe*

3.2 การสร้างแบบจำลองการรู้จำด้วย RNN

จากข้อเสียของ RNN ที่ย้อนกลับเอาอินพุตจากสถานะในเวลาก่อนหน้าได้ในระยะสั้น ๆ ในบทความนี้จึงได้นำ RNN ทั้งหมด 3 โมเดลได้แก่ LSTM, GRU, BiSLTM ที่แก้ไขปัญหาดังกล่าวนำมาใช้ในการรู้จำภาษามือไทยแบบเรียลไทม์ โดยมีพารามิเตอร์ได้แก่ Number of Nodes คือ จำนวนของ Input Node ในบทความนี้กำหนดขั้นต่ำไว้ 64 จนถึง 256 Activation คือตัวฟังก์ชันที่ใช้ในการรับผลรวมจากการประมวลผลทั้งหมดจากทุก Input Node เข้ามาพิจารณาตามกลไกการคำนวณของ Activation Function นั้น ๆ แล้วส่งต่อไปเป็น Output ซึ่งในบทความนี้กำหนดเป็น ReLU และ Softmax และ Optimizer คือ อัลกอริทึมการเพิ่มประสิทธิภาพทำหน้าที่เป็นกลไกการปรับปรุงค่าน้ำหนักของตัวแปรต้นต่าง ๆ รวมถึงค่าความคลาดเคลื่อน (Bias) ในบทความนี้ได้เลือกใช้ได้แก่ Adagrad, Adamax, Adam or RMSprop ดังตารางที่ 3

ตารางที่ 3 *พารามิเตอร์ของ RNN ที่ใช้ในการทดลอง*

Parameters	Value
RNN Model	LSTM, GRU, BiLSTM
Number of Nodes	Between (64, 256)
Activation	'Relu' or 'Softmax'
Optimizer	'Adagrad', 'Adamax', 'Adam' or 'RMSprop'

โครงสร้างของ RNNi โดยที่ i = {LSTM, GRU, BiLSTM} แสดงในภาพที่ 5 โดยที่ 3 ชั้นแรกเป็น RNN โมเดล เนื่องจากชุดข้อมูล TSL10 ถูกสร้างจากอาสาสมัครจำนวนน้อย ซึ่งอาจเกิดปัญหาโอเวอร์ฟิตติ้ง (Overfitting) จึงทำการเพิ่มชั้น Dropout ลงไประหว่างชั้นของ RNN โดย Dropout layer จะสุ่มตัดการเชื่อมต่อระหว่างโหนดในชั้นก่อนหน้ากับชั้นต่อไป โดยตัดการเชื่อมต่อเหล่านี้ด้วยการกำหนดค่าเป็นศูนย์ (zero) โดยสุ่มตัดบางโหนดออกจากการคำนวณในแต่ละรอบการฝึกฝน การทำ Dropout จะช่วยให้โมเดลสามารถเรียนรู้และสร้างรูปแบบที่เหมาะสมกับข้อมูลได้ดีขึ้น โดยไม่เกิดการเรียนรู้ที่ผิดพลาดจาก overfitting และ 3 ชั้นสุดท้ายเป็นชั้นของ Dense ซึ่งจะทำหน้าที่ในการการปรับค่าน้ำหนัก (weight) และค่าไบแอส (bias) ของโหนดในแต่ละชั้น ซึ่งช่วยให้โมเดลสามารถเรียนรู้และสร้างรูปแบบที่ซับซ้อนได้มากขึ้น

3.3 การวัดประสิทธิภาพ

การวัดประสิทธิภาพของการรู้จำภาษามือไทยจะแบ่งชุดข้อมูล ออกเป็น 2 ส่วน คือ 1) TSL10-Train จากวิดีโอทั้งหมด 1,000 คลิป โดยแบ่งเป็นข้อมูลโดยการสุ่มสำหรับการฝึกต่อการทดสอบเป็นอัตราส่วน 60:40, 70:30 และ 80:20 พร้อมทั้งใช้ 5-fold สำหรับการทดสอบโมเดล และ 2) TSL10-Test

จำนวน 100 คลิป ซึ่งเป็นวิดีโอจากนักศึกษาผู้บกพร่องทางการได้ยิน จำนวน 5 คน

เครื่องมือสำหรับวัดประสิทธิภาพของโมเดล ในบทความนี้ใช้ Confusion Matrix เพื่อหาค่าความแม่นยำ (Accuracy) ของการทำนายของโมเดล โดยสมการ (13)

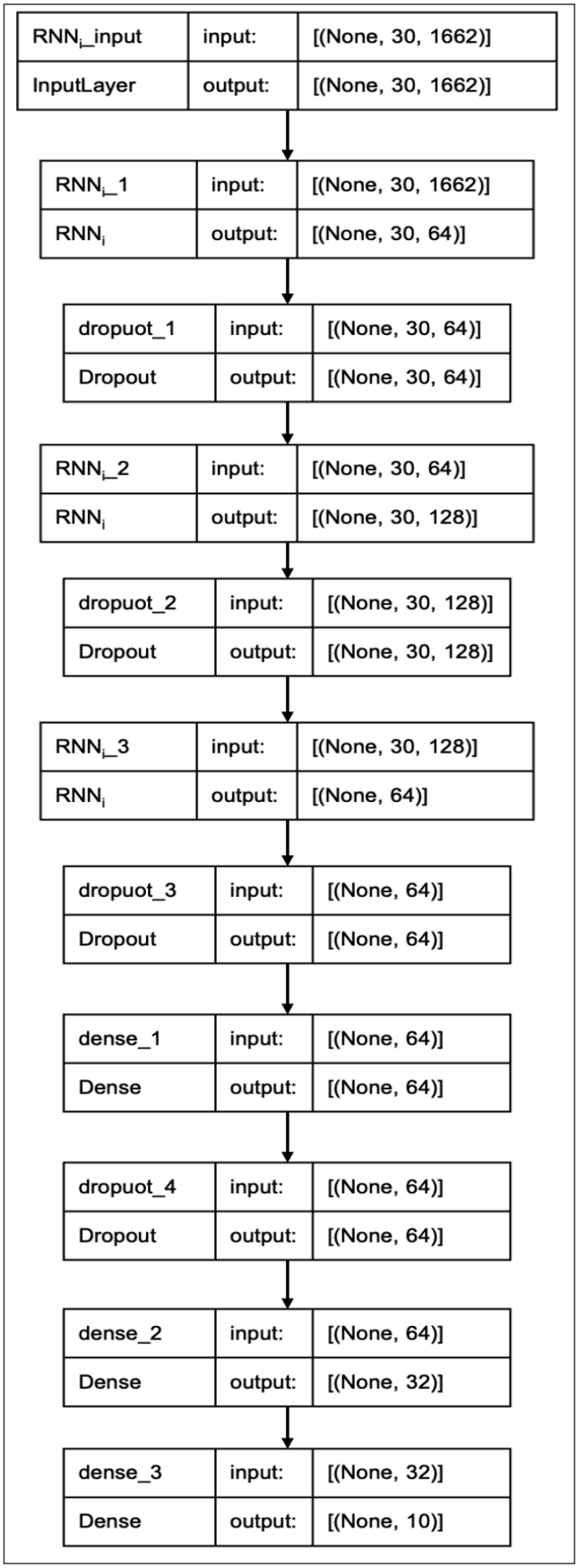
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \quad (13)$$

โดยที่ TP คือส่วนที่โมเดลทำนายว่าจริง และเป็นจริง TN คือส่วนที่โมเดลทำนายว่าเท็จ และเป็นเท็จ FP คือส่วนที่โมเดลทำนายว่าจริงแต่เป็นเท็จ และ FN ส่วนที่โมเดลทำนายว่าเท็จแต่เป็นจริง

4. ผลการดำเนินงาน

4.1 ชุดข้อมูล TSL10-Train

ผลการทดสอบการรู้จำท่าทางภาษามือไทยด้วย RNN บนชุดข้อมูล TSL10-Train จำนวน 900 คลิป ทำการทดสอบหาประสิทธิภาพโดยแบ่งวิดีโอเป็นอัตราส่วนข้อมูลฝึกต่อข้อมูลทดสอบ (Train : Test) เท่ากับ 60:40 70:30 และ 80:20 ตามลำดับ โดยมีกำหนดค่าอัตราการเรียนรู้ (Epochs) ของ LSTM, GRU และ BiLSTM เท่ากับ 500 รอบ โดยผลการทดสอบหาค่าความแม่นยำ (Accuracy) และ ค่าการสูญเสีย (Loss) แสดงในตารางที่ 4 โดยมีค่าความแม่นยำของ LSTM, GRU และ BiLSTM เฉลี่ยที่ 83.15% 87.13% และ 48.51% มีค่าการสูญเสียที่ 0.2283, 0.2268 และ 2.1673 ตามลำดับ และทำการทดสอบด้วย 5-fold cross validation ซึ่งแสดงผลในตารางที่ 5 ซึ่งพบว่าผลการทดสอบแบบ 5-fold ให้ผลลัพธ์ที่ดีที่สุดที่ โดยที่ LSTM, GRU และ BiLSTM ให้ความแม่นยำ 92.35%, 91.36% และ 86.23% มีค่าการสูญเสียที่ 0.1006, 0.3829 และ 0.2185 ตามลำดับ โดยในภาพที่ 6 แสดงผลการฝึกโมเดลทั้ง 3 แบบของ TSL10-Train ด้วย 5-fold แสดงให้เห็นค่าความแม่นยำเทียบกับรอบอัตราการเรียนรู้ (ภาพที่ 6 (a)) และค่าการสูญเสียเทียบกับรอบอัตราการเรียนรู้ (ภาพที่ 6 (b)) จากผลการทดสอบจะเห็นได้ว่าโมเดล BiLSTM มีประสิทธิภาพต่ำที่สุด เนื่องจาก BiLSTM นั้นมีความต้องการจำนวนข้อมูลสำหรับฝึกมากกว่า LSTM และ GRU เป็นผลจากการแบ่งข้อมูลในแต่ละ fold พบปัญหาที่ข้อมูลในแต่ละคลาสไม่เท่ากัน (Imbalance data) โดยเฉพาะใน 1-fold และ 4-fold ที่ให้ค่าความแม่นยำต่ำกว่าทุกการทดลอง



ภาพที่ 5 *โครงสร้าง RNN*

ตารางที่ 4 *ผลการวัดประสิทธิภาพชุดข้อมูล TSL10-Train แบบแบ่งอัตราส่วน Train : Test*

Train : Test	LSTM		GRU		BiLSTM	
	Accuracy(%)	Loss	Accuracy(%)	Loss	Accuracy(%)	Loss
60:40	81.83	0.2299	88.99	0.2400	50.99	3.1336
70:30	88.00	0.1956	90.14	0.1885	36.42	2.2460
80:20	79.62	0.2596	82.25	0.2521	58.12	1.1224
Average	83.15 ± 3.5461	0.2283 ± 0.0261	**87.13 ± 3.4801**	0.2268 ± 0.0275	48.51 ± 9.03	2.1673 ± 0.8229

ตารางที่ 5 *ผลการวัดประสิทธิภาพชุดข้อมูล TSL10-Train แบบ 5-fold*

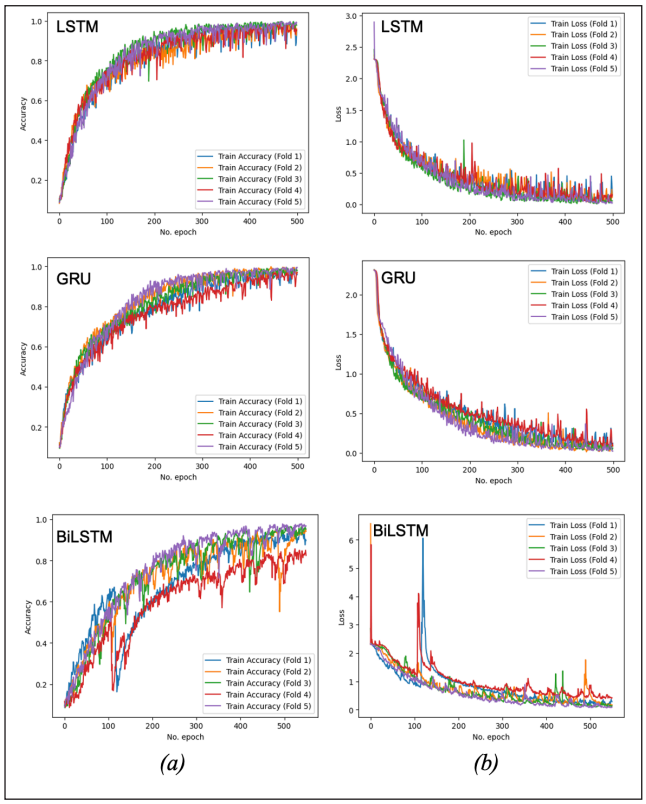
k-fold	LSTM		GRU		BiLSTM	
	Accuracy(%)	Loss	Accuracy(%)	Loss	Accuracy(%)	Loss
1-fold	93.89	0.1519	90.56	0.1215	88.89	0.2823
2-fold	87.78	0.1163	90.00	0.0984	92.77	0.1690
3-fold	94.45	0.0562	92.78	0.6095	88.89	0.1345
4-fold	92.78	0.1528	90.56	0.8560	69.45	0.4169
5-fold	92.78	0.0260	92.78	0.2291	91.11	0.0898
Average	**92.34 ± 2.36**	0.1006 ± 0.0512	91.36 ± 1.19	0.3829 ± 0.2993	86.23 ± 8.51	0.2185 ± 0.1179

4.2 ชุดข้อมูล TSL10-Test

การทดสอบกับ TSL10-Test จำนวน 100 คลิปโดยใช้อาสาสมัครที่ไม่ได้อยู่ใน TSL10-Train จำนวน 5 คน ผลการทดสอบแสดงในภาพที่ 7 ซึ่งเห็นได้ว่าทั้ง 10 คำมีความแม่นยำในระดับ 98.00% ขึ้นไป โดยคำว่า "สบายดี" "สวัสดี" "ป่วย" และ "ขอบคุณ" ที่มีความแม่นยำสูงในทั้ง 3 โมเดล ในขณะเดียวกันคำว่า "หิว" เป็นคำที่ให้ความถูกต้องต่ำสุด โดยเฉพาะใน BiLSTM ซึ่งผลการทดสอบกับชุดข้อมูล TSL10-Test ให้ผลลัพธ์อยู่ที่ 99.00% 99.00% และ 98.00% ด้วยวิธี LSTM, GRU และ BiLSTM ตามลำดับ โดยผลการทดลองแสดงในตารางที่ 6 ในภาพที่ 8 แถวแรกแสดงตัวอย่างของผลการทดสอบที่ให้ผลที่ถูกต้องกับคำว่า "สบายดี" และ "รัก" ด้วย GRU และในภาพที่ 8 แถวที่สองจากซ้ายไปขวาแสดงตัวอย่างของผลการทดสอบกับคำว่า "หิว" ที่ทำนายผิดพลาดเป็น "ฉัน" และ "คุณ" ตามลำดับด้วยโมเดล BiLSTM

4.3 การเปรียบเทียบผลการทดลอง

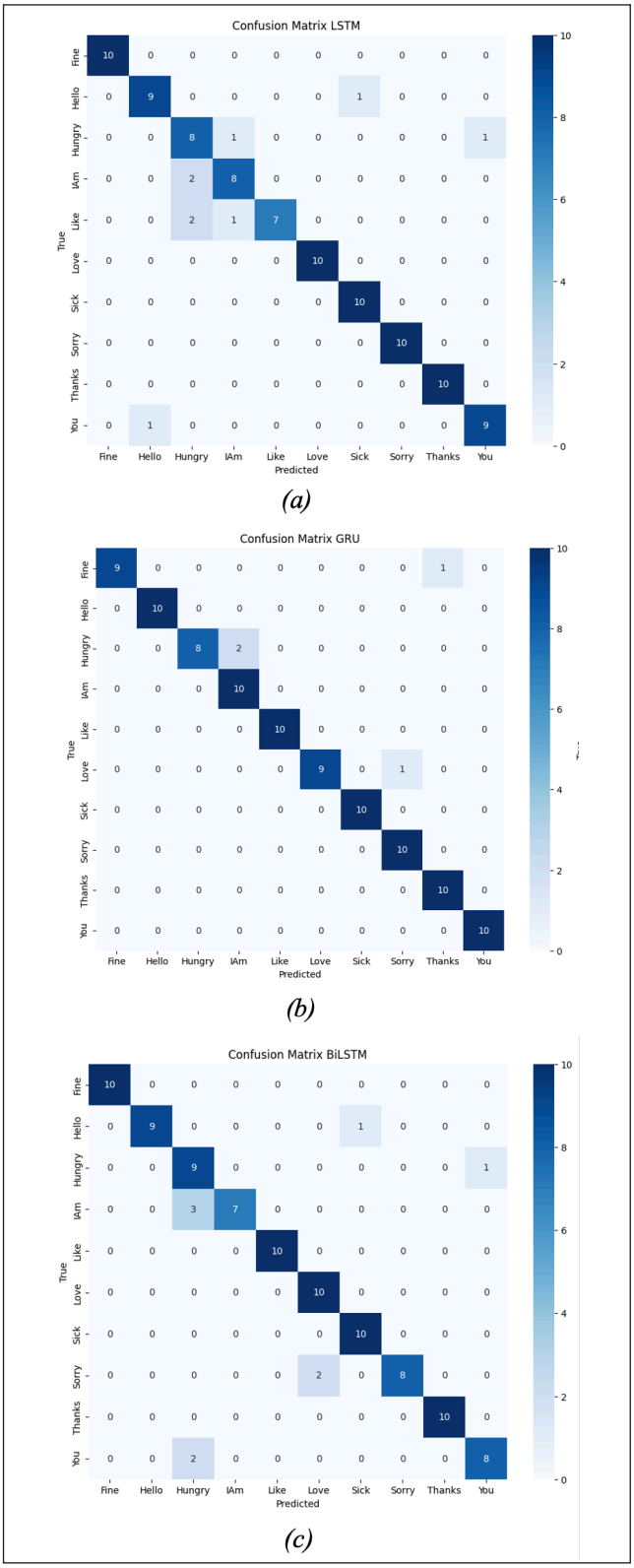
ผลการทดลองการรู้จำภาษามือด้วย RNN กับชุดข้อมูล TSL10 ทั้งสองชุด โดยใช้อัลกอริทึมแบบ LSTM, GRU และ BiLSTM พบว่าวิธี LSTM และ GRU ให้ผลที่ใกล้เคียงกัน เนื่องจากตัวแบบจำลองของทั้งสองวิธีมีลักษณะการทำงานคล้ายกันและเหมาะสมกับข้อมูลที่เป็นลำดับ โดยมีการออกแบบมาให้มีการดักจับข้อมูลที่เป็นลำดับในระยะยาว (Long-Term Data) แต่ด้วยโครงสร้างของ LSTM และ GRU มีโครงสร้างที่คล้ายกัน แต่ GRU มีความซับซ้อนน้อยกว่าทำให้ได้ความแม่นยำมีความใกล้เคียงกัน แต่ค่าความสูญเสียของ GRU มีค่ามากกว่า ซึ่งมีความเป็นไปได้ที่วิธีดังกล่าวจะมีความแม่นยำสูงถ้ามีการเพิ่มอัตราการเรียนรู้โดยผลการทดสอบค่าความแม่นยำเฉลี่ยของ LSTM และ GRU ใกล้เคียงกันที่ 92.34 % และ 91.36% ตามลำดับ ในขณะที่วิธี BiLSTM เป็นวิธีที่ใช้การประมวลผลแบบ LSTM แบบสองทางเป็นวิธีที่เหมาะสมกับ



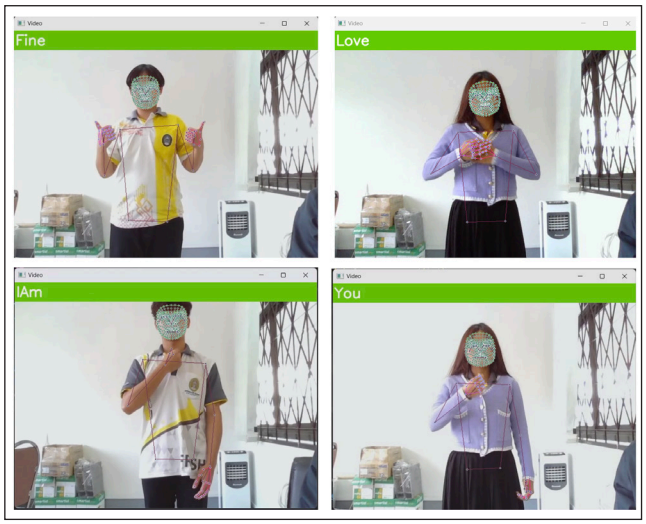
ภาพที่ 6 *กราฟแสดงผลการทดสอบข้อมูล TSL10-Train ที่อัตราส่วน 5-fold (a) Accuracy และ (b) Loss*

การทำนายสิ่งที่จะเกิดขึ้น (Forward LSTM) โดยอาศัยบริบท (Backward LSTM) ซึ่งไม่เหมาะสมกับข้อมูลภาพวิดีโอภาษามือที่สนใจแค่สิ่งที่เกิดขึ้นข้างหน้าด้านเดียวทำให้ผลลัพธ์ออกมาต่ำกว่าวิธี LSTM และ GRU ในทุกการทดลอง นอกจากนี้สังเกตได้ว่าในการทดลองของการแบ่งข้อมูล Train : Test ด้วยวิธี BiLSTM ให้ผลลัพธ์ที่ต่ำผิดปกติ เนื่องจากใช้วิธีการสุ่มในการสร้างชุดข้อมูลจึงทำให้เกิดปัญหาการแบ่งข้อมูลของคลาสที่ไม่เท่ากัน

การเปรียบเทียบผลการทดลองกับชุดข้อมูล TLS10-Test แบบท่าเคลื่อไหว 10 ท่า โดยเปรียบเทียบกับการรู้จำภาษามือแบบกับทั้งภาษามือไทยและภาษามือกับงานวิจัยอื่น ๆ ในตารางที่ 1 พบว่าการวิธี MediaPipe ร่วมกับ RNN (GRU, LSTM) ให้ประสิทธิภาพที่ดีกว่าวิธีการอื่น ๆ สอดคล้องกับวิธีการที่นำเสนอทดสอบกับภาษามือท่าเคลื่อนไหวจากชุดข้อมูล TSL10-Test จำนวน 10 ท่า ให้ความแม่นยำถึง 99% ด้วย RNN แบบ LSTM และ GRU ซึ่งเมื่อเทียบกับงานวิจัย [7] ที่ใช้วิธีการเดียวกัน พบว่าการกำหนดโครงสร้างของ RNN มีความแตกต่างกัน โดยวิธีการที่นำเสนอใช้ Dropout เพื่อลดการเกิด Overfitting หลังชั้นของ RNN และ Dropout ที่ 4 เพื่อสุ่มปิดการทำงานบางโหนดหลังชั้นของ Dense ก่อนออก Output ซึ่งเหมาะสมกับชุดข้อมูล TSL10 ที่มีขนาดเล็ก



ภาพที่ 7 *ผลการทดสอบชุดข้อมูล TSL10-Test (a) LSTM, (b) GRU และ (c) BiLSTM*



ภาพที่ 8 *ตัวอย่างผลการทดลอง*

ตารางที่ 6 *ผลการทดสอบความแม่นยำบน TSL10-Test*

Words	LSTM(%)	GRU(%)	BiLSTM(%)
สบายดี	100	100	100
สวัสดี	100	100	99
หิว	95	98	94
ฉัน	97	98	97
ชอบ	99	100	100
รัก	100	100	98
ป่วย	100	100	99
ขอโทษ	100	100	98
ขอบคุณ	100	100	100
คุณ	99	100	97
รวม	99	99	98

5. สรุป

งานวิจัยนี้มีวัตถุประสงค์ในการพัฒนาตัวแบบการรู้จำภาษามือไทยแบบเคลื่อนไหวด้วยวิธีโครงข่ายประสาทเทียมแบบวนกลับ โดยข้อมูลนำเข้าเป็นคีย์พอยท์ที่สกัดเอาจุดเด่นของผู้สื่อสารภาษามือด้วยไลบารี MediaPipe ซึ่งเป็นแนวทางในการต่อยอดไปสู่การสร้างโปรแกรมหรือแอปพลิเคชันในการแปลภาษามือไทยของผู้พิการทางการได้ยินหรือการสื่อสารให้บุคคลทั่วไปสามารถเข้าใจในสิ่งที่ผู้พิการต้องการจะสื่อได้ โดยในงานวิจัยนี้ได้รวมตัวอย่างคำศัพท์ที่ใช้ในชีวิตประจำวันของผู้พิการทางการได้ยินหรือการสื่อสารจำนวน 10 คำศัพท์ สำหรับการทดสอบสร้างระบบรู้จำท่าทางภาษามือไทย โดยใช้โมเดล LSTM, GRU และ BiLSTM ในการทดลอง ผู้วิจัยได้สร้างชุดข้อมูลวิดีโอภาษามือ 10 คำ TSL10-dataset โดยผู้เชี่ยวชาญภาษามือ และผู้บกพร่องทางการได้ยิน จำนวน 1,000 คลิป ประกอบไปด้วย 10 ท่า จำนวน 900 คลิป สำหรับการฝึก (TSL10-Train dataset) และ 10 ท่า จำนวน 100 คลิป สำหรับการทดสอบ (TSL10-Test dataset) ผลการทดสอบบน TSL10-Train ให้ประสิทธิภาพสูงสุดของ LSTM, GRU และ BiLSTM 92.35%, 91.36%, และ 86.23% ตามลำดับ ขณะที่บนชุดข้อมูล TSL10-Test มีค่าความแม่นยำที่ 99%, 99%, และ 98% ตามลำดับ

การพัฒนาระบบการรู้จำภาษามือไทยด้วยวิธีการสกัดจุดเด่นของผู้สื่อสารภาษามือด้วย MediaPipe และการรู้จำด้วย RNN แบบ LSTM และ GRU ให้ผลการทดลองที่มีประสิทธิภาพสูงกับภาษามือท่านิ่งและท่าเคลื่อนไหว ในขณะที่โมเดลแบบ BiLSTM พบว่ารูปแบบของข้อมูลแบบลำดับของภาพวิดีโอไม่เหมาะสมกับกับโมเดล BiLSTM ที่ต้องการทำนายสิ่งที่จะเกิดขึ้นข้างหน้าโดยอาศัยการตัดสินใจจากบริบทก่อนหน้า นอกจากนี้ปัญหาที่พบคือจำนวนข้อมูลที่อินพุตเข้าไปในทั้งสองตัวแบบใช้ชุดข้อมูล 1 ท่าต่อ 1 วินาที หรือ 30 เฟรม ซึ่งการใช้ภาษามือในสถานการณ์จริงอาจช้าหรือเร็วตามเฉพาะบุคคล ทำให้ระบบยังมีประสิทธิภาพไม่ดีพอในการจำภาษามือแบบต่อเนื่องหลายคำทำให้ยังไม่เหมาะสมกับการพัฒนาแอปพลิเคชันแปลภาษามือ อีกทั้งการเรียงประโยคของผู้ใช้ภาษามือยังเป็นปัญหาที่จะแปลภาษามือได้ทั้งประโยค ในอนาคตผู้วิจัยจะพัฒนาระบบที่สามารถรู้จำภาษามือที่ต่อเนื่องกันเพื่อรองรับการพัฒนาแอปพลิเคชันการแปลภาษามือได้ให้มีความถูกต้องยิ่งขึ้น

6. เอกสารอ้างอิง

[1] W. Daengrueang, J. Kaewsritong, and W. Intakan. "The Development of Thai Sign Language Multimedia on Home Economics for Students with Hearing Impairment." *The Golden Teak : Humanity and Social Science Journal,* Vol. 28, No. 3, pp. 163 - 175, 2022.

[2] P. Kaewdee, M. Koodduderm, and W. Kiewkam. "Structural Analysis of Occupational Vocabulary in Thai Sign Language for Sign Language Interpreters." *Interdisciplinary Studies Journal,* Vol. 23, No. 2, pp. 2 - 17, July - December, 2023.

[3] R. Kanakala, J. Mohan, and K. Reddy. "Modelling a deep network using CNN and RNN for accident classification." *Measurement: Sensors,* Vol. 27, pp. 1 - 10, 2023.

[4] M. Babaee, Z. Li, and G. Rigoll. "A dual CNN–RNN for multiple people tracking." *Neurocomputing,* Vol. 368, pp. 69 - 83, 2019.

[5] C. Tatiyavoranun. "An Application of Artificial Neural Network for Thai Sign Language Recognition." *Engineering Transactions,* Vol. 23, No. 1, pp. 51 - 57, 2020.

[6] A. Chaikaew, K. Somkuan, and T. Yuyen. "Thai Sign Language Recognition: An Application of Deep Neural Network." *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical Computer and Telecommunication Engineering,* Chiang Rai, pp. 128 - 131, 2021.

[7] C. Damrongekarun, L. Pisitpipattana, S. Waijanya, and N. Promrit. "Development of Thai Sign Language Detection and Conversion System into Thai with Deep Learning." *KKU SCIENCE JOURNAL,* Vol. 51, No. 3, pp. 216 - 225, September - December, 2023.

[8] G. H. Samaan, A. R. Widie, A. K. Attia, A. M. Asaad, A. E. Kamel, S. O. Slim, M. S. Abdallah, and Y. Cho. "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition." *Electronics,* Vol. 11, No. 19, pp. 1 - 15, 2022.

[9] M. Alnaggar, A. I. Siam, M. Handosa, T. Medhat, and M.Z. Rashad. "Video-based real-time monitoring for heart rate and respiration rate." *Expert Systems with Applications,* Vol. 225, pp. 1 - 11, 2023.

[10] C. Bisogni, L. Cimmino, M. D. Marsico, F. Hao, and F. Narducci. " Emotion recognition at a distance: The robustness of machine learning based on hand-crafted facial features vs deep learning models." *Image and Vision Computing,* Vol. 136, pp. 1 - 15, 2023.

[11] J. Bora, S. Dehingia, A. Boruah, A. A. Chetia, and D. Gogoi. "Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning." *Procedia Computer Science,* Vol. 218, pp. 1384 - 1393, 2023.

[12] B. Sundar and T. Bagyammal. "American Sign Language Recognition for Alphabets Using MediaPipe and LSTM." *Procedia Computer Science,* Vol. 215, pp. 642 - 651, 2022.

[13] S. Khumwongsa and W. Yawai. "Smart Application for Thai and English Sign Language Translation." *Journal of Applied Informatics and Technology (JIT),* Vol. 5, No. 2, pp. 178 - 194, 2023.

[14] E. Gedkhaw. "The Performance of Thai Sign Language Recognition Using 2D Convolutional. Neural Networks." *The 13th NPRU National Academic Conference, Nakhon Pathom University,* Nakhon Pathom, Thailand, pp. 546 - 573, 2021.

[15] K.E. A. Kumar, D. V. Kalaga, Ch. M. S. Kumar, M. Kawaji, and T. M. Brenza. "Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells." *Chaos, Solitons & Fractals,* Vol. 146, pp, 1 - 12, 2021.

[16] S. M. Al-Selwi, M. F. Hassan, S. J. Abdulkadir, A. Muneer, E. H. Sumiea, A. Alqushaibi, and M. G. Ragab. "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review." *Journal of King Saud University - Computer and Information Sciences,* Vol. 36, No. 5, pp. 1 - 34, 2024.

[17] S. Khan and V. Kumar. "A novel hybrid GRU-CNN and residual bias (RB) based RB-GRU-CNN models for prediction of PTB Diagnostic ECG time series data." *Biomedical Signal Processing and Control,* Vol. 94, pp. 1 - 18, 2024.

[18] M. R. Ahmed, S. Islam, A.K.M. M. Islam, and S. Shatabda. "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition." *Expert Systems with Applications,* Vol. 218, pp. 1 - 21, 2023.

[19] W. Fan, J. Yao, S. Cui, Y. Wang, S. Xu, Y. Tan, F. Yang, and W. Wu. "Bi-LSTM/GRU-based anomaly diagnosis for virtual network function instance." *Computer Networks,* Vol. 249, pp. 1 - 16, 2024.

[20] K. Subyen, W. Samhansub, and J. Suraseing. "Sign language processing software." *RMUTSB Academic Journal,* Vol. 4, No.1, pp. 46 - 56, 2016.