SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 11, No. 4, (2025), pp. 333-349
Open Access. Online & Print

DOI: 10.5281/zenodo.11042528

DEEP LEARNING FOR TWO-PERSON MAE MAI MUAY THAI CLASSIFICATION: LEVERAGING NORMALIZED HUMAN POSE SEQUENCES AND CNN-LSTM

Thanirat Yoddamnern[1], Panomkhawn Riyamongkol[2*]

[1]Faculty of Engineering, Naresuan University, Phitsanulok, Thailand. Email: thaniraty60@nu.ac.th, Orcid ID: https://orcid.org/0009-0003-8973-1298
[2]Faculty of Engineering, Naresuan University, Phitsanulok, Thailand. Email: panomkhawnr@nu.ac.th, Orcid ID: https://orcid.org/0000-0001-5320-6246

Received: 11/11/2025
Accepted: 18/11/2025

Corresponding Author: Panomkhawn Riyamongkol (panomkhawnr@nu.ac.th)

ABSTRACT

Muay Thai postures are the distinctive stances and movements used in traditional Thai boxing, concentrating on balance, strength, and fluid transitions between offensive and defensive techniques. Therefore, this study presents a novel system for detecting and categorizing Muay Thai postures through video processing and deep learning methods, which are skilled at examining the coordinated movements of two people fighting on the stage. The procedure involves identifying human formations and evaluating incorporated positions during duplicative movements, and organizing the postures operating a hybrid CNN and LSTM model. Moreover, to improve learning efficiency with joint position sequence data, the MinMaxScaler technique is used for data normalization. The type is based on an extensive dataset of 590 videos featuring various backgrounds, divided into 80% for training and 20% for testing. The resulting model completes an overall type accuracy of 84%, providing a strong basis for future applications in solely and paired of Muay Thai analysis. This approach offers potential advantages in sports science, competitor training, and motion analysis, allowing real-time posture recognition and performance evaluation. It helps coaches and practitioners identify incorrect techniques, reduce injury risks, and increase the accuracy of battle movements. In addition, future developments could include real-time integration into training applications, support for 3D pose analysis, and expanded datasets to improve the model's accuracy and validity across various conditions and practitioners.

KEYWORDS: Human Pose Estimation, Muay Thai Posture, CNN-LSTM, Joint Detection, Action Recognition.

1. INTRODUCTION

Human Activity Recognition (HAR) is an important phenomenon in both the study of computer concepts and computer-based machine learning, as well as its wide use in the healthcare industry, security surveillance, body training, and human-computer interaction. The combination of deep learning models, namely Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM), assists in making the methods of Human Activity Recognition (HAR) more reliable and flexible. Recent advancements in the field have led to the development of models with more complex spatiotemporal patterns based on observable and detector data, allowing for more accurate and robust recognition of human actions in various settings. An integrated system with wearable activity cameras and discriminant classifications using LSTM neural structured learning has enhanced HAR implementation by integrating handcrafted context with in-depth material modeling to improve recognition using time-series sensor data. This approach demonstrates the effectiveness of combining statistical models with deep learning models in HAR systems involving mobile and wearable devices (Md Zia Uddin et al., 2020). Utilizing skeletal data and motion sensor data (gyroscopes and accelerometers) has improved the recognition of human movement patterns by identifying non-basic movement routines. This includes organized material modeling and a dual-chain framework consisting of CNN to capture structural elements and LSTM to model the material framework. Incorporating gyroscope data improved the model's receptivity to rotational and emotional body movements, therefore increasing its precision and reliability, specifically in real-time application procedures (Zheng et al., 2019), a spatiotemporal attention model for skeleton-based action recognition uses joint-wise attention instruments to emphasize important joints and time structures within movement sequences, therefore improving accuracy and strength in human activity recognition assignments (Song et al., 2017). In recent years, wearable sensing technologies have gained significant application in human motion recognition, specifically for accurately collecting detailed gestures and joint-level movements. A wearable glove-based approach was first used with flex sensors embedded in the fingers to detect hand gestures with high accuracy. This technique operated a deep learning measure, integrating Convolutional Neural Networks (CNNs) to categorize sensor-derived movements. The flex sensors showed direct measurements of finger bending angles, allowing the procedure to recognize fine finger motions in real-time effectively, which demonstrates benefits in fields such as potential validity, sign language interpretation, and physical rehabilitation, where accurate recognition of intricate hand postures is required (Lee et al., 2020). The flex sensors are presented in Figure 1.

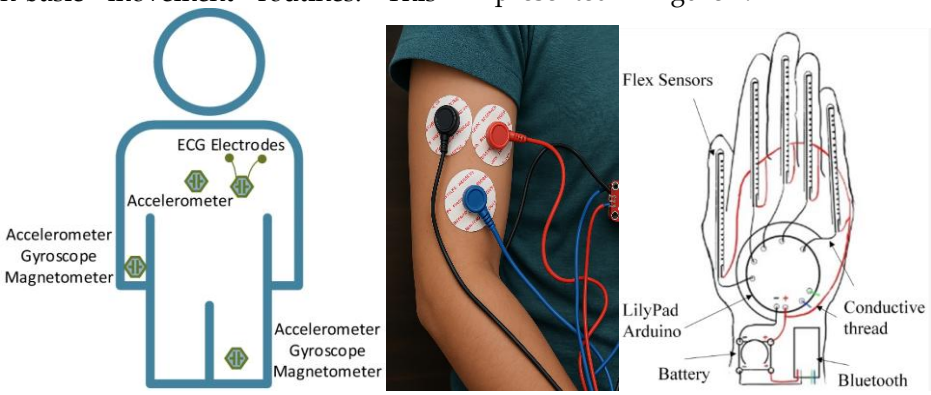


Figure 1: The Flex Sensor Placements on The Human Body.

However, these advances have relied on constrained settings with homogeneous or plain white backgrounds to improve skeleton extraction and reduce model variance in various existing studies. A deep learning line for Baduanjin movement classification using videos recorded under strictly controlled conditions with white backgrounds to ensure high-commitment joint detection (Yang et al., 2023). Presented in Figures 2 and 3.

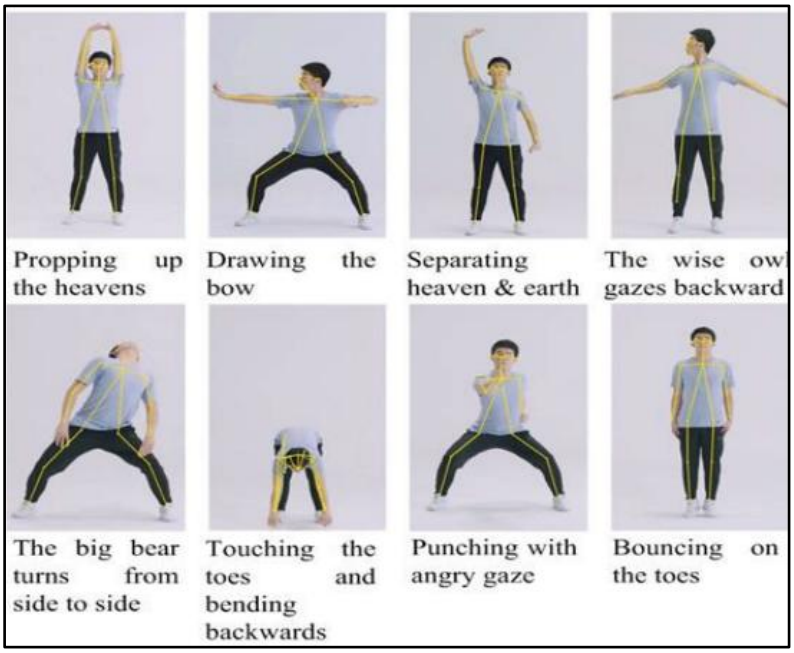


Figure 2: Baduanjin Movements Using Skeletal Data from Videos, The Dataset with A White Background.

In addition, various studies on Muay Thai action recognition have conducted curated datasets with uniform backgrounds to mitigate occlusions and improve pose tracking accuracy (Chatpun et al. 2023).

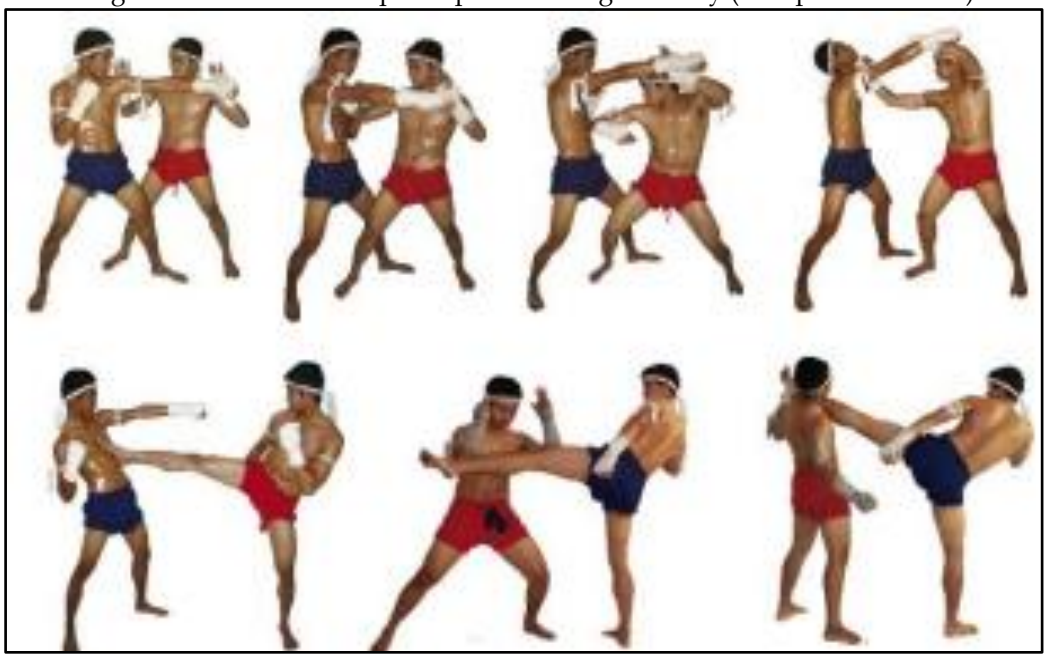


Figure 3: Thai Boxing Techniques Dataset for Deep Learning Action Recognition, The Dataset with A White Background.

To overcome the challenges posed by constrained environments, this study shows a method for recognizing Mae Mai Muay Thai postures under various background conditions. Integrating human pose estimation with a CNN-LSTM framework to achieve strong recognition performance applicable in real-world environments beyond controlled laboratory settings. This study contributes to the advancement of cultural context, scalable, and adaptive HAR systems. Moreover, recent progress in the HAR domain focuses on the growing use of hybrid models that combine Convolutional Neural Networks (CNNs) with temporal mechanisms like Long Short-Term Memory (LSTM) networks. Furthermore, these architectures are also outstanding for spatial structures and temporal dynamics in

video data, proving highly effective for applications such as sports performance analysis, human-computer interaction, and smart surveillance.

Table 1: Comparative Studies on Pose-Based Action Recognition.

Study	Techniques	Application Domain	Dataset
Baduanjin Action Recognition	Skeleton-based CNN + LSTM	Healthcare / Traditional exercise recognition	Custom dataset of Baduanjin practitioners
Skeleton-Based Action Recognition	Graph Convolutional Network (GCN), ST-GCN	Human Activity Recognition (HAR), Fitness	NTU RGB+D, Kinetics Skeleton
Attention-Based LSTM with Dilated CNN	Dilated CNN + Attention-based LSTM	Human Action Recognition, Surveillance	UCF101, Kinetics-400
Pose-Based CNN (P-CNN)	Pose-based CNN with hand-crafted motion features	Action recognition in videos	JHMDB, MPII Cooking, Hollywood2
Two-Stream LSTM Framework	CNN + LSTM (Pose Stream + Motion Stream)	Action Recognition from skeletal data and motion	NTU RGB+D

The 2023 video data on Baduanjin exercise consisted of video data recorded on a white background and used as a model that combined CNNs and LSTMs to pinpoint the various postures in the exercise. It must have been advantageous because of the clear background and strict chains of skeletons that could be studied through pose estimation (Yang et al. 2023). The spatiotemporal LSTM model, which applied the 3D body joint skeletal data of NTU RGB+D dataset to recognize all actions, found the LSTMs to be effective in modeling the dynamic movements of the body joints in HAR (Liu et al. 2017). These further concepts integrate attention conclusions and enlarged convoluted layers within the framework and conduct to prove that the model successfully understands small-scale nuanced details of local motion and big worldly scale motion measures in the UCF11, UCF Sports, and J-HMDB datasets (Saha et al. 2021). Additionally, in 2015, a pose-based convolutional descriptor (P-CNN), which utilized extracted body joint positions, presented by this approach offered a foundation regarding using pose recognition within deep learning centers with the examination of motion recognition on information packages such as JHMDB and MPII Cooking (Cheron et al. 2015). Additionally, a two-stream LSTM design was employed, which included occasional video (RGB) data as well as the optical flow data to improve the representation of motion. This methodology used numerous past methods on different data (Li et al. 2017). Human pose estimation has also been applied widely in many areas including activity recognition, healthcare surveillance, human-computer interaction (HCI), and sports performance analysis. Besides this, the advancement of the deep learning models, primarily the Convolutional Neural Networks (CNNs) and Recurrent Neural Networks such as the Long short-term memory (LSTM) have enhanced the accuracy and dependability of the pose estimation systems. It contains a summary of the current research with deep learning-based pose estimation that has been performed by order of research irrelevance as study title, applied model and roles of the model, as well as the datasets applied.

Table 2: Summary Of Related Work Using Pose Estimation Techniques.

Study	Techniques	Application Domain	Dataset
OpenPose	CNN + Part Affinity Fields (PAF)	Real-time multi-person pose estimation	COCO, MPII
AlphaPose	Region-based CNN + Pose Refine Machine	Multi-person pose estimation in crowded scenes	COCO, CrowdPose
HRNet	High-Resolution CNN	Human action and gesture recognition	COCO, MPII
Fall Detection with Pose Estimation	CNN + LSTM	Healthcare monitoring, elderly fall detection	UR Fall Detection
Yoga Pose Classification	MediaPipe + CNN	Physical exercise, pose correction	Custom Dataset
Temporal CNN for 3D Pose	Temporal CNN + 2D Keypoints	Monocular 3D pose estimation	Human3.6M
ST-GCN	Spatio-Temporal Graph CNN	Skeleton-based action recognition	NTU RGB+D
EfficientPose	EfficientNet + Pose Regression	Real-time systems, robotics	COCO
Improved 3D Pose Estimation (Zhou et al.)	CNN + Camera Calibration	Monocular 3D pose reconstruction	MPI-INF-3DHP, 3DPW
Cross-View Fusion Network	CNN + View Fusion	Multi-view 3D pose estimation	Human3.6M, MPII

The concept of Part Affinity Fields (PAF), which involves the connection of body parts, is regarded as one of the most influential OpenPose frameworks in real-time multi-person pose estimation that can be used in applications that reveal body parts even in a cluttered environment.

This was largely tested on the COCO and MPII datasets and formed a fundamental ground for future studies on the subject matter will look into (Du et al. 2015). AlphaPose also advances the real-time pose detection technique by adding a region CNN and a Pose Refine Machine (PRM), to enhance the ability to handle occlusions and a binding part. It works well in crowded datasets and has been tested on challenging datasets (e.g., CrowdPose and COCO) (Shahroudy et al. 2016). In addition, the High-Resolution Network (HRNet) suggests a paradigm shift because it holds high-resolution representations across the network, which is beneficial for keypoint adaptation in some details. It reports competitive performance using both COCO and MPII, in scenarios with single people (Zhang et al. 2019).

In the medical field, fall detection systems that use pose estimation with temporal modeling (CNN + LSTM) are very sensitive and specific in detecting potentially dangerous situations, particularly for elderly care. The techniques typically use the UR Fall Detection dataset to train and test these models (Cheron et al. 2015).

The recognition systems presented by yoga, based on the use of a lightweight set of MediaPipe and CNN classifiers, are intended to deliver real-time performance feedback for exercise correction. The systems are usually designed based on bespoke data gathered under controlled conditions (Song et al. 2016). Temporal CNNs for human pose estimation are centered on the application of binary series 2D keypoints to individually develop 3D poses over time.

A famous example is training such models using the Human3.6M dataset and validating them on data markedly larger in terms of what is covered in the dataset when viewed and acted upon (Wang et al. 2016). Moreover, the idea of spatial and temporal relationships in skeletal data is utilized by Spatio-Temporal Graph Convolutional Networks (ST-GCN), which represent joints as nodes in a graph. It is a practical architecture that focuses on action recognition problems and has worked extremely competently on the NTU RGB+D dataset (Li et al. 2015).

EfficientPose is a human pose estimation system that scales well and is better at both speed and accuracy by relying on the EfficientNet backbone, and demonstrates good performance in practice, such as robotics or augmented reality, and on the COCO benchmark. Monocular 3D pose improvements are improved with CNN-produced image features and camera pose calibration to estimate precise 3D pose determination without using multi-cameras using datasets such as MPI-INF-3DHP and 3DPW.

Besides, Cross-View Fusion Networks enhance the quality and accuracy of the 3D pose estimation by using the fact that the information of many camera perspectives is taken into consideration and such even applies to a situation such as sports analytics where various perspectives can be provided (Yan et al. 2018; Shi et al. 2019; Liu et al. 2020).

This study aims to provide a case of an innovative way of identifying Mae Mai Muay Thai postures and observing them in different background conditions through human pose estimation with a CNN-LSTM-based system. The model is suggested to ensure good performance outside controlled conditions. Additionally, classic Mae Mai Muay Thai training often relies on the instruction of experts, the location of training, and the specifics of observation, which may hinder participation in training, particularly in small or isolated communities.

The availability of financial resources to ensure the quality of trainers also limits access to this cultural heritage. Therefore, the study can contribute to a culturally responsible, scalable, and versatile human activity recognition (HAR) architecture that promotes better agent actions in diverse environmental settings and provides greater and fairer access to traditional martial arts through digital technologies.

2. LITERATURE REVIEW

Redmon et al. proposed a real-time processing technique called You Only Look Once (YOLO), which is based on GoogleNet and uses the unique properties of neural networks to segment an image into parts and predict the region and probability of each segment. The bounding box around an object is determined by the predicted probability.

The YOLO algorithm divides the image into a grid of N x N bins, with each bin containing only one predicted object, by applying the non-maxima suppression algorithm. This detection of one object per grid cell is independent of the number of frames and is not limited to the number of frames, but only the confident frames are counted. YOLO typically uses ImageNet for pre-training and then uses the target detection dataset for training.

In addition, many improvements have been made to the YOLO architecture, resulting in YOLOv2 and YOLOv3, which can improve detection accuracy while maintaining very high object detection speed. For example, in YOLOv3, a variant of the Darknet architecture is used, and 53 layers are trained on the ImageNet dataset, which is the Pascal VOC dataset.

Therefore, YOLOv3 outperforms most algorithms. Most of the detections are real-time.

Adding connections and up-sampling allows the detection of objects with up to 3 different aspect ratios, making YOLOv3 smarter and able to detect small objects, but it may have a slower processing speed than previous versions, due to the YOLO architecture (C. Sungur et al. 2019). Deval (2022), described the following indexes and values related to object detection performance:

Intersection over Union (IoU), is an index used to evaluate the overlap between the ground truth box and the predicted box, 29 analyzing whether the detection is correct (True Positive) or incorrect (False Positive), (S. Shah et al. 2023). Benoughidene et al. (2024), used CNN with LSTM to process the UCF Sports dataset using Feature Extraction technique along with analysis of posture changes through time sequence, resulting in an accuracy of 94.3%, which shows that the model can accurately capture the rhythm of athletes' movements.

Vinaya & Geeta (2023) used CNN-LSTM with sports datasets, such as Cricket and Soccer, by adjusting the LSTM structure to suit the data sequence length, resulting in an accuracy of 91.5% in terms of model development. Genc et al. (2024), proposed the use of Fine-tuned CNN with LSTM to classify athletes' postures from the KTHm dataset and experimented with Bidirectional LSTM to be able to analyze both past and future directions, which increased the accuracy to 93.6%, especially for fast-changing postures.

Yang et al. (2023) also showed that combining skeleton and CNN data can effectively train LSTM models, achieving 96.7% accuracy on Wushu, Baduanjin, and other Chinese physical exercise datasets in the dimension of sensor data applications. Lee et al. (2022) trained a CNN-LSTM model for the classification of basketball IMU sensor poses with dimensionality reduction. They combined it with Bidirectional LSTM, achieving a classification accuracy of 92.4%. This indicates the adaptability of the model to both image and wearable sensor data. Other works have also experimented with methods to improve model performance.

For example, the Attention Mechanism in LSTM can be used to focus on significant areas of the image, such as joints and the center of motion. Zhang et al. (2023) achieved an accuracy of 93.8% using this approach. Additionally, Dilated CNN can be used to reduce the number of parameters without sacrificing the field of view. Hassan et al. (2023) addressed background and lighting conditions by employing a thermal camera and a CNN-LSTM model to classify boxing poses based on infrared camera data, achieving an accuracy of 95.2%.

This study demonstrates resilience to environmental conditions. 32. Singh et al. (2025), concluded that the combination of CNN-LSTM is still the best approach for classifying sequential sports behaviors, with an emphasis on data preparation, such as, aligning the skeleton sequence (alignment), normalizing the data with MinMaxScaler or StandardScaler, which can reduce the effects of body position inconsistencies. Yilmaz et al. (2024) tested the CNN-LSTM model on an automated football training system and showed that it can provide accurate recommendations on a frame-by-frame level in real time.

3. RESEARCH OBJECTIVES

1) To develop a model for estimating Muay Thai postures using video from Muay Thai posture dance using an image processing method.
2) To develop a model for predicting Muay Thai postures using the vector map technique instead of the traditional feature map.

4. CONCEPTUAL FRAMEWORK

1. The video data must have a resolution of not less than 960 x 540 pixels but not more than 1,980 x 1,080 pixels.
2. The image of the person must be seen in full, including the head, body, hands, arms, and legs.
3. The Muay Thai dance postures consist of not less than 5 postures out of a total of 15 postures, which are: 1) Fish-tooth alternating pose, 2) Bird-breaking nest pose, 3) Lord Java throwing a spear, 4) Inao stabbing a dagger, 5) Mount Sumeru lifting pose, 6) Old Man holding a sheath pose, 7) Mon-ying pose, 8) Putting a ball toss pose, 9) Crocodile slashing its tail pose, 10) Aira breaking its trunk pose, 11) Naga twisting its tail pose, 12) Virulhak Klap pose, 13) Dab Chawala pose, 14) Khun Yak catching a monkey pose, and 15) Erawan breaking his neck pose, respectively.
4. Use the pose estimation principle and develop an algorithm to predict poses, namely YoloPose, and further apply it.
5. There must be 2 people to analyze the poses.

5. RESEARCH HYPOTHESIS

The Muay Thai posture prediction model can accurately predict postures at least 80% of the time.

6. METHODS

6.1. Sample Preparation

The data collection process includes videos demonstrating Mae Mai Muay Thai techniques from offensive and defensive angles. These demonstrations were performed by two professional Muay Thai fighters which are Chuthong Kiatchatchai and Payakdong Kiatchatchai, who both compete in the under-115-pound weight class, to ensure accuracy and authenticity of the recorded movements. In addition, the sample includes university students from Kamphaeng Phet Rajabhat University in the under-112-pound and under-175-pound weight classes. This diversity in participants was intended to enrich the dataset and ensure a broader representation of body types and movement variations, enhancing the generalizability of the model.

(a) (b)

Figure 4: Mae Mai Muay Thai stance sample group (a) Professional boxer b) Amateur Boxer.

6.2. Data Collection

Data collection includes videos showcasing Mae Mai Muay Thai techniques from both offensive and Data augmentation is a technique for expanding the amount of data available by inserting slightly modified copies of existing data or creating new artificial data from existing data. Once a machine learning model is trained, it acts as a normalizer and helps minimize overfitting. Examples include adjusting the angle of the image, zooming in, out, rotating, flipping, and translating it from the x-axis to the y-axis.

(a) (b) (c)

Figure 1: Data Augmentation (A) Original (B) Flip (C) Zoom 75%.

6.3. Feature Map to Vector Dataset

The videos are then converted into images using keyframes, and human detection is performed to extract joint skeletons. The joint skeleton data is recorded in CSV format for further analysis and system development.

Figure 6: Concept Of Feature Map to Vector Dataset.

From Figure 6, the video input serves as the primary data source for capturing dynamic human motion. Each video is first decomposed into individual frames using OpenCV, a widely adopted open-source computer vision library. This frame-by-frame extraction enables precise temporal analysis of human activities. Following frame acquisition, human detection and joint skeleton estimation are performed using the YOLO-pose model, a real-time, deep learning-based object detection framework optimized for identifying key body landmarks. YOLO-pose efficiently locates and maps critical joint positions of individuals within each frame, providing a detailed skeletal representation essential for downstream analysis. Subsequently, the positional data of detected joints, specifically for person 1 and person 2, are extracted in terms of X and Y axis coordinates. This structured joint data is systematically converted and stored in CSV format using Pandas in conjunction with OpenCV. The resulting CSV files provide a well-organized dataset that can be readily utilized for training and evaluating action recognition models in subsequent stages of the research pipeline.

6.4. Architecture for CNN-Bilstm Model

Deep learning has many different architectures designed to handle different types of data. CNNs and LSTMs each have outstanding advantages, and their integration in a CNN-LSTM model combines these strengths to enhance data analysis, especially in video and sequential image tasks. This architecture is a hybrid deep learning framework that combines the spatial feature extraction capabilities of Convolutional Neural Networks with the temporal sequence learning strengths of Long Short-Term Memory networks. It is highly effective for processing visual sequences, such as video frames or human pose trajectories. In this framework, CNN layers identify and compress spatial information from each frame, reducing dimensionality and easing the computational burden on the subsequent LSTM layers. Moreover, the LSTM component models the temporal relationships across the frames, allowing the system to better understand motion dynamics and classify actions with increased precision. However, to support efficient model training and convergence, the pose coordinate data are normalized using the MinMaxScaler tool from sklearn. preprocessing, which scales all joint positions to a standardized range between 0 and 1. This normalization encourages addressing disparities in numerical values across different joints. Furthermore, interactions including two individuals,

such as paired techniques in martial arts, the synchronization of their joint sequences is critical. Temporal positions arrangement ensures that sets of pose data reflect coordinated movements, which is important for accurately interpreting interaction-based actions. In addition, these preprocessing steps, normalization, and synchronization are also important for increasing the CNN-LSTM model's strength, reducing overfitting, and improving generalization performance.

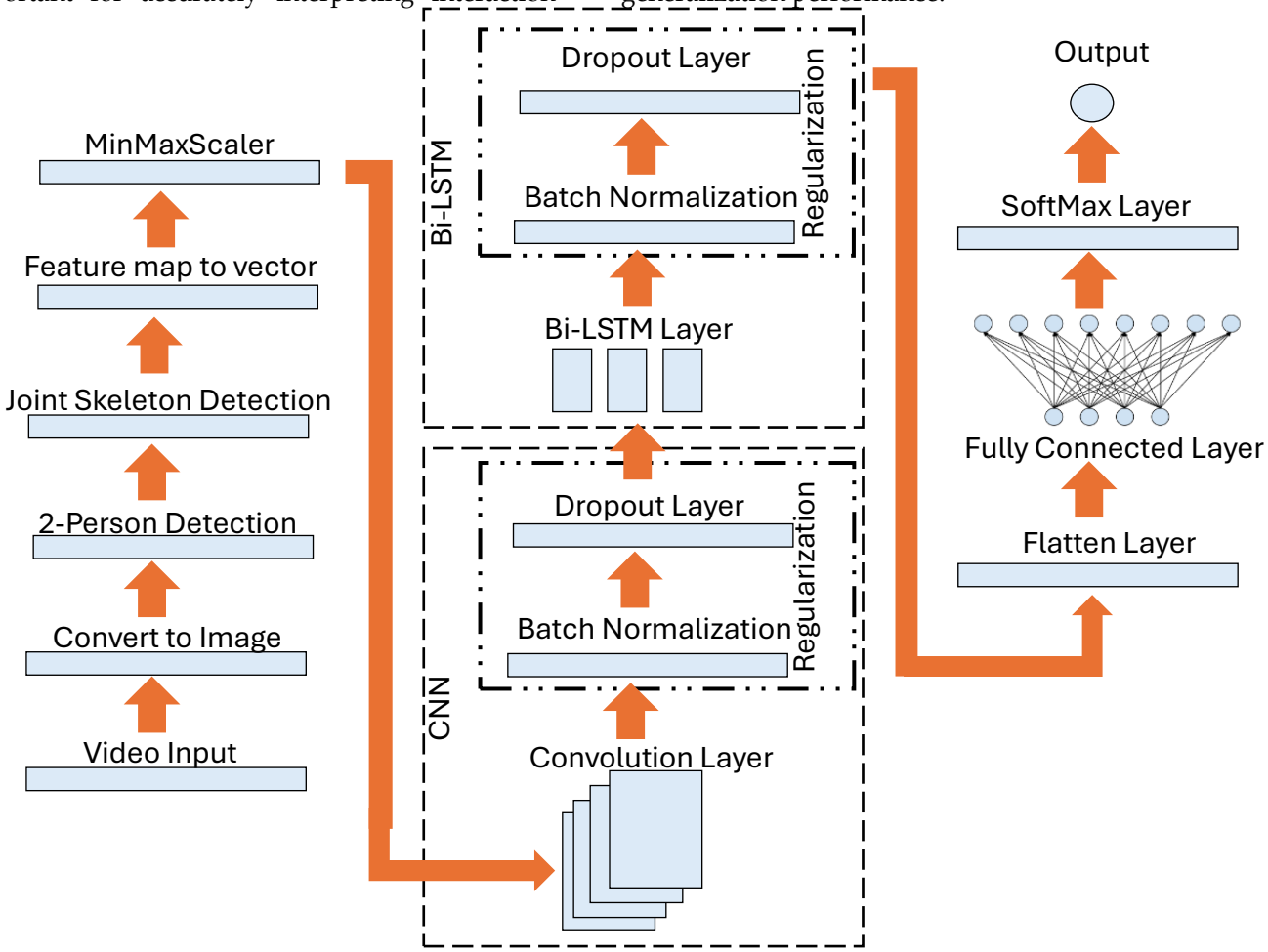


Figure 7: Methodology Of the Proposed Mae Mai Muay Thai Postures Recognition Using CNN-LSTM.

Figure 7 is a system overview diagram showing the operation of the CNN-BiLSTM model by converting video data into images and detecting only two people in the image, ignoring the background using YOLO Detection. When the people in the image are detected, the joint positions of the two people are found using YOLO Pose Estimation, and the values are saved in a data table in CSV file format for easy use in OpenCV, Pandas, Sklearn, Numpy, torch, and Tensorflow libraries. The next step is to enter the CNN network that has undergone 3 layers of regularization and send it to the Bi-LSTM network that has undergone 3 more layers of regularization. Then it will enter the Flatten Layer, Fully Connected Layer, and SoftMax Layer processes in order. It is a Deep Learning model creation using both CNN and Bi-LSTM for classifying sequence data. It uses techniques such as Batch Normalization and Dropout to improve efficiency and prevent overfitting. The brief instructions are as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where:

x is the raw value of the feature.

x_{min} is the minimum value of the feature in the dataset.

x_{max} is the maximum value of the feature in the dataset.

x' is the value obtained after normalizing to the range [0, 1].

2) Conv1D: 1D convolution layer for processing sequence data.

$$y_t = \sum_{i=0}^{k-1} x_{t+i} \cdot w_i + b \quad (2)$$

3) MaxPooling1D: 1-dimensional pooling layer for data reduction.

$$y_t = max(x_t, x_{t+1}, \dots, x_{t+k-1}) \quad (3)$$

Where:

y_t is the output value at position t.

x is the input vector of the sequence.

x_t+i is the input value at the position that the filter covers in the sequence (i = 0 to k−1).

w_i is the weight of the kernel at position i.

b is the bias term.

k is the size of the kernel (filter size).

4) Bidirectional: Bidirectional LSTM layer for processing both front-to-back and back-to-front sequence data.

$$Output_t = [\overrightarrow{h_t} \cdot \overleftarrow{h_t}] \quad (4)$$

Where:

$\overrightarrow{ht}$ is the hidden state from the forward LSTM.

$\overleftarrow{ht}$ is the hidden state from the backward LSTM.

5) Dropout: Dropout layer to prevent overfitting.

$$\tilde{x}i = r_i \cdot x_i \quad (5)$$

Where:

x_i is the input value of the i neuron

r_i is a random variable from the distribution (randomly 1 or 0)

$\tilde{x}i$ is the value passed to the next layer (some are 0 because they were dropped)

6) BatchNormalization: Batch Normalization layer to improve training stability.

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} xi \quad (6)$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (xi - \mu B)^2 \quad (7)$$

Where:

xi is the input of any layer in the model.

m is the number of samples in a batch.

μ_B is the average of the inputs in the batch.

σ_B^2 is the variance of the inputs in the batch.

$$\hat{x} = \frac{xi - \mu B}{\sqrt{\sigma^2 + \varepsilon}} \quad (8)$$

Where:

$\hat{x}i$ is the input adjusted to have mean = 0 and std = 1

ε is a small constant to prevent division by zero

7) Flatten Layer: Reshapes a multi-dimensional input tensor into a 1D vector.

$$Flatten(X) = x \in Rd1 \cdot d2 \cdot \dots \cdot dn \quad (9)$$

Where:

Input: A tensor (multi-dimensional array), such as output from a convolutional layer.

Operation: Concatenates all dimensions except the batch size into one single vector.

Output: A 1D vector (flattened), which can be fed into Dense layers.

8) Dense: Fully Connected layer

$$y = f(Wx + b) \quad (10)$$

Where:

x is the input vector from the previous layer

W is the weight matrix of size $[n_{output} \times n_{input}]$

B is the bias vector of size $[n_{output}]$

f is the activation function such as ReLU, sigmoid,softmax
y is the output vector
9) Softmax : Often placed in the last layer of a neural network to produce a probability output.

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (11)$$

Where:
x_i is logit or the output value from the last layer before activation (of class i)
n is number of all classes
$f(x_i) \in (0,1)$ is probability of each class

5. Evaluation

The image classification performance is evaluated using the following measurements:

1) Precision is the total actual accuracy. 2) Recall is the accuracy of non-actual items. 3) Accuracy is the total actual accuracy, and 4) F1score is the average of Precision and Recall.

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (15)$$

7. RESULTS

From the original video, videos of different lengths were cut into short videos to be used as the main keyframes for each pose. The dataset was divided into 80% videos for system learning (Train), totaling 472 videos, divided into Salap-Funpla: 53 samples, I-Nao-Thaeng-Krit: 44 samples, Yor-Khao-Phra-Su-Main: 51 samples, Pak-Luk-Thoi: 49 samples, Hak-Nguang-Aiyara: 54 samples, Nakha-Bit-Hang: 54 samples, Dap-Chawala: 104 samples, and Hak-Kho-Erawan: 63 samples. 20% of the videos were for testing the accuracy (Test), totaling 118 videos.

Table 3: Data Of mea Mai Muay Thai Posture.

Label	Mea Mai Muay Thai posture
01	Salap-Funpla
04	I-Nao-Thaeng-Krit
05	Yor-Khao-Phra-Su-Main
08	Pak-Luk-Thoi
10	Hak-Nguang-Aiyara
11	Nakha-Bit-Hang
13	Dap-Chawala
15	Hak-Kho-Erawan

The model will first find people in the image because the image does not have a white background, so there are many objects in the image that may not be people.

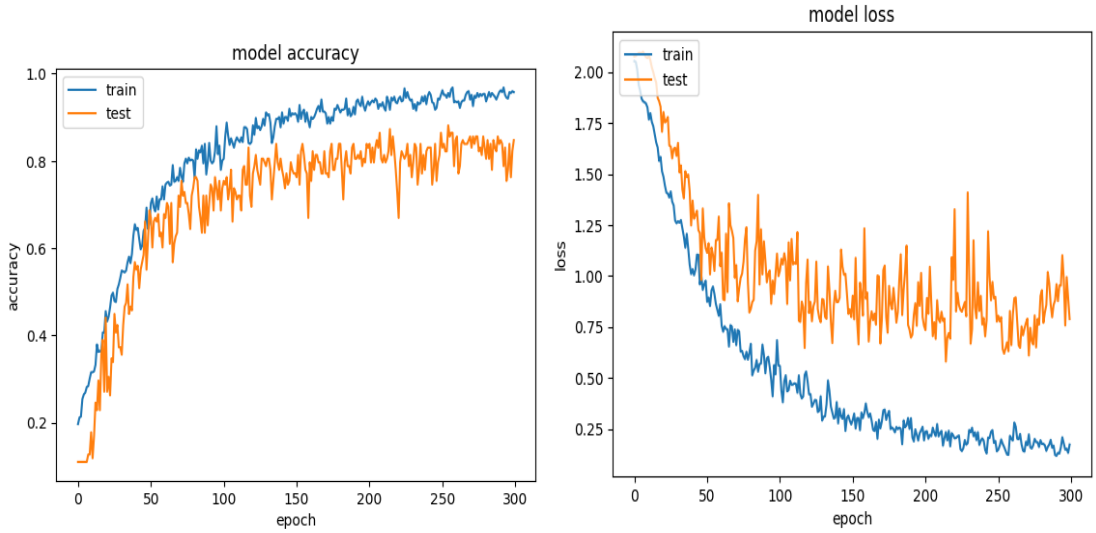
Then, it will find human joints using the Yolov8n-pose.pt algorithm that has been pre-trained by the main developer, with a total of 17 joints specified, resulting in new images that can be easily and quickly converted to vectors.

Table 4: Performance Evaluation of The Deep Learning Models.

DL Model	Accuracy
CNN	0.7627
LSTM	0.7905
CNN-LSTM (non MinMaxScaler)	0.7373
CNN-LSTM (MinMaxScaler)	0.8383

From Table 3, the performance comparison of different deep learning models, the CNN-LSTM model using MinMaxScaler normalization, 0.8383 as the best performance, shows that feature scaling is

an efficient method to enhance model performance. At most, the performance of the model using LSTMs alone was pretty good, achieving an accuracy of 0.7905. Whereas the CNN model achieved 0.7627 for accuracy, the CNN-LSTM model without MinMaxScaler had the worst accuracy of 0.7373, indicating that not normalizing the data hurt learning efficiency, and therefore accuracy.



(a) (b)

Figure 8: Accuracy And Loss of The Proposed Model During Training and Testing (A) Accuracy of The Proposed Model (B) Loss of The Proposed Model.

Figure 8 shows the comparative analysis of training (blue line) and testing (orange line) accuracy and loss for the model equal 300 epochs, highlighting the model's learning progress and generalization capabilities. During the initial phase (0 to 50 epochs), both the training and testing accuracy increase rapidly. This indicates that the model is effectively learning from the data. In the middle phase (50 to 200 epochs), the training accuracy continues to rise, reaching approximately 90-95%. The testing accuracy fluctuates but generally increases to around 80-90%. A divergence between the training and testing accuracy begins to emerge, potentially signaling the onset of overfitting. In the late phase (200 to 300 epochs), the training accuracy stabilizes at a high level (~95%), while the testing accuracy exhibits oscillations, averaging around ~85%. The model demonstrates signs of overfitting, as the training accuracy significantly surpasses the testing accuracy.

Table 5: Evaluation Of Model Performance

MMMT Posture	Precision	Recall	F1-score
Salap-Funpla	0.9000	0.9000	0.9000
I-Nao-Thaeng-Krit	0.8571	0.6667	0.7500
Yor-Khao-Phra-Su-Main	0.8571	0.9231	0.8889
Pak-Luk-Thoi	0.7500	0.8571	0.8000
Hak-Nguang-Aiyara	0.8889	0.6667	0.7619
Nakha-Bit-Hang	0.8571	0.9231	0.8889
Dap-Chawala	0.7778	1.0000	0.8750
Hak-Kho-Erawan	0.9130	0.8750	0.8936
Accuracy			**0.8475**
Macro Avg	**0.8501**	**0.8515**	**0.8448**
Weighted Avg	**0.8532**	**0.8475**	**0.8445**

Using the percentage Confusion Matrix allows a clearer analysis of the Class Imbalance problem and can effectively identify classes where the model learns well and where it has shortcomings.

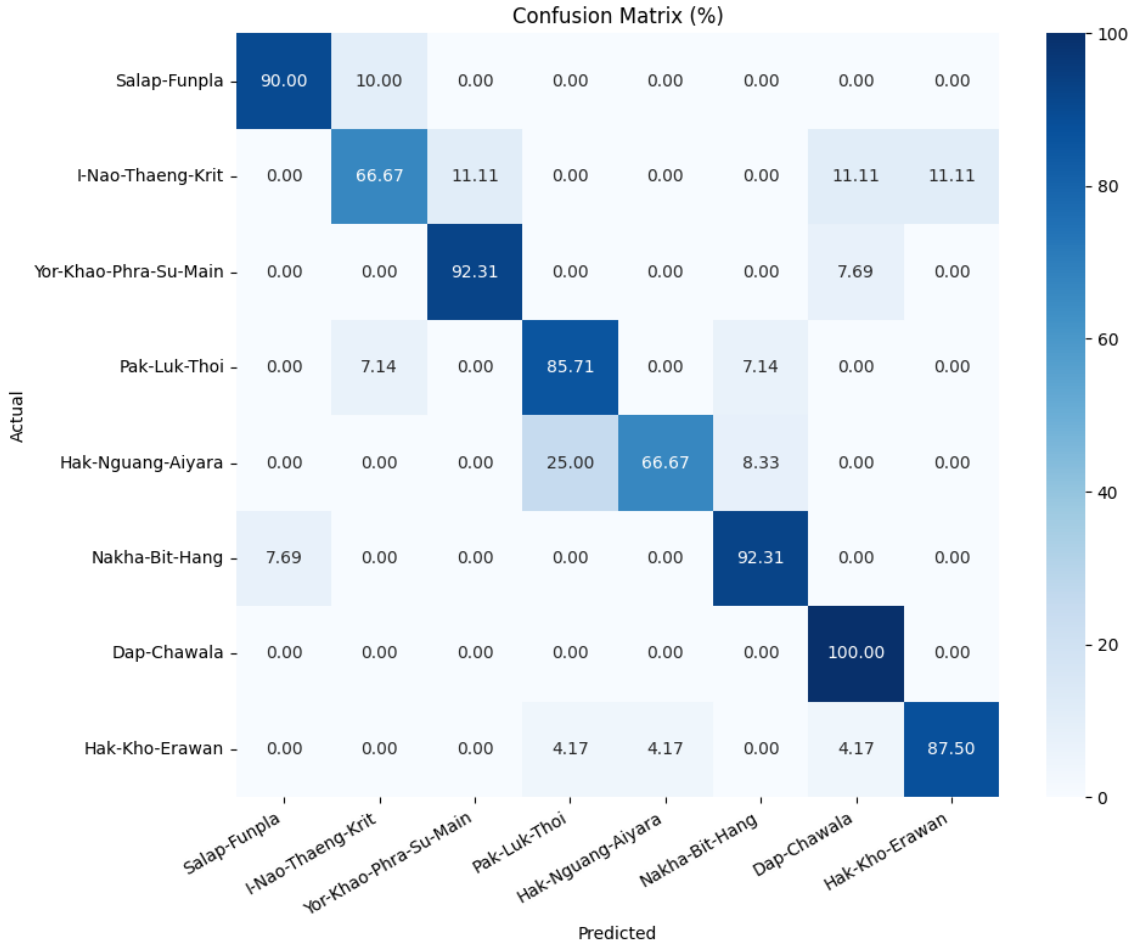


Figure 9: Confusion Matrix by Percentage of Data.

From Figure 9, the percentage Confusion Matrix is a useful way to make the model results more understandable, allowing a fair comparison of the accuracy of each class. The classes that predict more than 90% accurately are seen as having 5 classes in dark blue color: Salap-Funpla, Yor-Khao-Phra-Su-Main, Nakha-Bit-Hang, and Dap-Chawala.

8. DISCUSSION

The CNN-LSTM model without MinMaxScaler normalization performed the most destructively, achieving a precision of only 73.73%. This highlights the adverse effect of considering natural, unstructured information in deep learning systems that handle the configuration of information. By making the input scattered everywhere, this may interfere with the process through which the model learns and hence produces erratic updates of its internal gradients and subsequently a weaker training process. Such an unsatisfactory outcome also indicates that, in the absence of normalization, the model has difficulty in making sense of and using what it has learned, especially in new rows where the acceleration or size of the movement might be heterogeneous. Thus, while the combination of CNNs and LSTMs can still be a valuable suggestion for understanding visual aspects and movement sequences, the utilization of tools like MinMaxScaler should be considered to streamline the process. The presence of such outcomes justifies how critical it is to make the right choice of structure and ensure that the data preparation steps align with it if you want to achieve top performance in tasks, including the ability to recognize actions based on pose sequences. The joints can be connected to form complete skeletons, rather than relying solely on the raw joint coordinates of two people for posture prediction. These skeleton structures are possibly used as input data to indicate the overall movement posture. This method may produce more precise results, as it collects the coordinated motion between both players. However, when applied in real-time techniques, this approach is possibly overused regarding the complexity of controlling complete skeletal data from multiple persons collectively. In addition, the pose estimation system sometimes

detects a third person when only two individuals are present. This often occurs due to overlapping skeleton lines, such as when the right leg of the person on the left overlaps with other parts of the image, leading to false detections. Therefore, it is crucial to limit data collection and processing to only two individuals per frame to minimize errors during training. This restriction ensures cleaner, more accurate data and helps the model learn significant joint associations without being complicated by incorrect or excessive skeleton detections.

9. CONCLUSION

From the experimental results in both person and joint detection, including Muay Thai pose classification, it was found that the models used have satisfactory performance, especially the YOLOv8n model, which, although not the highest accuracy, has outstanding processing speed, making it suitable for tasks that require real-time speed. In addition, choosing YOLOv8n-Pose.pt, which is a pre-trained model, reduces the burden of training a new model and can be applied for practical use immediately. Furthermore, the model accuracy and loss graphs show clear signs of overfitting, with the training set's accuracy reaching ~95% while the test set's accuracy is only ~84%. While this is still acceptable, it suggests that further improvements could be made, such as using regularization techniques or increasing the data diversity. This study presents an innovative method for recognizing synchronized Muay Thai techniques by analyzing joint data from two individuals. An advantage of this is resilience to varying background conditions; the model is specifically designed to disregard background elements, such as colors, allowing it to function effectively in uncontrolled environments. However, conventional systems that concentrate on a single subject capture the interactive dynamics between two performers within the same scene, allowing a better understanding of paired movements, which are essential in Muay Thai. Furthermore, the system aims to enhance coordination among individuals, which is a crucial element in detecting traditional Muay Thai stances correctly, and synchronized offensive as well as defensive patterns. Additionally, the architecture is capable of learning complex spatiotemporal dependencies by treating the two participants as a unit. The preprocessing stage involves MinMax scaling of joint coordinates, which normalizes the input values to a similar range across different body types and varying camera distances. This improves model stability, accelerates convergence rates, and provides consistent data to support better learning.

Acknowledgments: The performers featured in the video data used for this research are professional Muay Thai fighters, Chuthong Kiatchatchai and Payakdong Kiatchatchai, whose expertise ensured the accuracy and authenticity of the movements captured. Their contribution was essential in providing high-quality data for model training and evaluation.

Funding Information: This research and publication were supported by a scholarship from Kamphaeng Phet Rajabhat University, Thailand.

REFERENCES

A. Ellouze, N. Kadri, A. Alaerjan, and M. Ksantini, "Combined CNN-LSTM Deep Learning Algorithms for Recognizing Human Physical Activities in Large and Distributed Manners: A Recommendation System," Computers, Materials and Continua, vol. 79, no. 1, pp. 351–372, 2024, doi: 10.32604/cmc.2024.048061.

A. Polo-Rodriguez, A. Montoro-Lendinez, M. Espinilla, and J. Medina-Quero, "Classifying Sport-Related Human Activity from Thermal Vision Sensors Using CNN and LSTM," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH, 2022, pp. 38–48. doi: 10.1007/978-3-031-13321 3_4.

B. Abdelhalim and T. Faiza, "Automatic Sports Video Classification Using CNN-LSTM Approach Automatic Sports Video Classification Using CNN-LSTM Approach ⋆," 2023. [Online]. Available: https://www.researchgate.net/publication/381740998

B. A. Khan and J.-W. Jung, "Deep Learning-Based Human Activity Recognition Using Dilated CNN and LSTM on Video Sequences of Various Actions Dataset," Apr. 15, 2025. doi: 10.20944/preprints202504. 1120.v1.

C. Sungur and A. Durdu, "Real-Time Diseases Detection of Grape and Grape Leaves using Faster R-CNN and SSD MobileNet Architectures," 2019. [Online]. Available:

https://www.researchgate.net/publication/334987612
Chatpun, S., & Supmak, S. (2023). Thai Boxing Techniques Dataset for Deep Learning Action Recognition. CIE Transactions. https://www.cie-dc.com/uploads/1/3/2/9/132987652/mt3.pdf
Cheron, G., Laptev, I., & Schmid, C. (2015). P-CNN: Pose-based CNN features for action recognition. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 3218–3226. https://doi.org/10.1109/ICCV.2015.367
Chéron, G. et al. (2015). P-CNN: Pose-based CNN features for action recognition. arXivpreprint. https://arxiv.org/abs/1506.03607
Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1110–1118. https://doi.org/10.1109/CVPR.2015.7298714
E. Genc, M. E. Yildirim, and Y. B. Salman, "Human activity recognition with fine-tuned CNN-LSTM," Journal of Electrical Engineering, vol. 75, no. 1, pp. 8–13, Feb. 2024, doi: 10.2478/jee-2024-0002.
E. Koşar and B. Barshan, "A new CNN-LSTM architecture for activity recognition employing wearable motion sensor data: Enabling diverse feature extraction," Eng Appl Artif Intell, vol. 124. p. 106529, Sep. 2023, doi: 10.1016/J.ENGAPPAI.2023.106529
H. Zhou, Y. Zhao, Y. Liu, S. Lu, X. An, and Q. Liu, "Multi-Sensor Data Fusion and CNN LSTM Model for Human Activity Recognition System," Sensors, vol. 23, no. 10, May 2023, doi: 10.3390/s23104750. 70
Lee, J., Kim, J., Park, J., & Kim, J. (2020). Wearable Glove-Based Hand Gesture Recognition Using Deep Learning. Journal of Mechanical Science and Technology, 34(2), 887–892. https://doi.org/10.1007/s12541-020-00467-w
Li, C. et al. (2017). Two-stream LSTM: A deep fusion framework for human action recognition. arXiv preprint. https://arxiv.org/abs/1704.01194
Li, C., Zhong, Q., Xie, D., & Pu, S. (2019). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 786–792. https://doi.org/10.24963/ijcai.2019/111
Liu, J., Shahroudy, A., Xu, D., Kot, A. C., & Wang, G. (2020). Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(3), 625–638. https://doi.org/10.1109/TPAMI.2018.2876325
Liu, J. et al. (2017). Skeleton-based action recognition using spatio-temporal LSTM. arXiv preprint. https://arxiv.org/abs/1707.02356
Md Zia Uddin & Ahmet Soylu. (2020). Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning. Scientific Reports. https://www.nature.com/articles/s41598-021-95947-y
Narkhede, A.H. Human Activity Recognition Based on Multimodal Body Sensing; Master's Thesis, San Jose State University: San Jose, CA, USA, 2019.
S. Shah and J. Tembhurne, "Object detection using convolutional neural networks and transformer-based models: a review," Journal of Electrical Systems and Information Technology, vol. 10, no. 1, Nov. 2023, doi: 10.1186/s43067-023-00123-z.
S. J. Dutta, T. Boongoen, and R. Zwiggelaar, "Human activity recognition: A review of deep learning-based methods," Jan. 01, 2025, John Wiley and Sons Inc. doi: 10.1049/cvi2.70003.
Saha, S. et al. (2021). Attention-based Bi-LSTM with dilated CNN for human action recognition. Future Generation Computer Systems, Elsevier.
Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1010–1019. https://doi.org/10.1109/CVPR.2016.104
Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12026–12035. https://doi.org/10.1109/CVPR.2019.01230
Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2017). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. AAAI Conference on Artificial Intelligence, 4263–4270. https://doi.org/10.1609/aaai.v31i1.11290
Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2017). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. AAAI Conference on Artificial Intelligence, 4263–4270.

https://doi.org/10.1609/aaai.v31i1.11290
V. R. M and G. C. Mara, "Journal of Advanced Zoology Human Activity Recognition Using CNN and Lstm Deep Learning Algorithms", [Online]. Available: https://jazindia.com
Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. European Conference on Computer Vision (ECCV), 20–36. https://doi.org/10.1007/978-3-319-46484-8_2
Yang, Z., et al. (2023). Deep Learning-Based Classification of Baduanjin Movements Using Skeletal Data From Videos. IEEE Journal of Translational Engineering in Health and Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10332470
Yang, X. et al. (2023). Baduanjin posture recognition using pose estimation and CNN-LSTM. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37435544/
Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. AAAI Conference on Artificial Intelligence, 7444–7452. https://doi.org/10.1609/aaai.v32i1.11779
Zhang, S., Liu, Z., Xiao, J., Shen, D., & Han, J. (2019). Attention-based LSTM with dilated CNN for human action recognition. Neurocomputing, 328, 89–100. https://doi.org/10.1016/j.neucom.2018.10.068
Zheng, W., Zeng, W., Wang, L., Wang, Q., & Tian, Q. (2019). Structured Temporal Pyramid Recurrent Neural Network for Video-Based Human Action Recognition. Sensors, 19(11), 2562. https://doi.org/10.3390/s19112562

BIOGRAPHIES OF AUTHORS

PANOMKHAWN RIYAMONGKOL is Associate Professor at Faculty of Engineering, Naresuan University, Phitsanulok City, Thailand. She received Master's degree in Electrical and Computer Engineering from the University of Miami, USA in 1999 and her Ph.D. in Electrical and Computer Engineering from the University of Miami, USA in 2003. She is currently a lecturer at the Department of Electrical and Computer Engineering, Faculty of Engineering, Naresuan University. Associate Professor Dr. Riyamongkol has published more than 20 journal articles, focusing on various fields such as image processing, computer vision, artificial intelligence, and video surveillance. She can be contacted at email: panomkhawnr@nu.ac.th.

THANIRAT YODDAMNERN is a Ph.D. student in Computer Engineering at Faculty of Engineering, Naresuan University, Phitsanulok City, Thailand. He is an Assistant Professor in Information Technology and works as a lecturer at the Computer Technology Program, Kamphaeng Phet Rajabhat University, Thailand. He holds a Master's degree in Electrical Engineering from King Mongkut's University of Technology North Bangkok, Bangkok, Thailand, in 2009. He is interested in research in image

		processing, embedded systems, IoT, and Human Activity Recognition. He can be contacted at email: Thanirat@kpru.ac.th.