ICIC Express Letters
Volume **18**, Number **1**, January **2024**

 ISSN 1881-803X
pp. **69–77**

RECOGNITION OF EXERCISE ACTIVITY USING CNN AND LSTM BASED ON ACCELEROMETER DATA

Ponnipa Jantawong[1], Olarik Surinta[2], Anuchit Jitpattanakul[3,4] and Sakorn Mekruksavanich[1,*]

[1]Department of Computer Engineering
School of Information and Communication Technology
University of Phayao
19 Moo 2 Tambon Maeka, Amphur Muang, Phayao 56000, Thailand
ponnipa.jantawong@gmail.com; *Corresponding author: sakorn.me@up.ac.th

[2]Multi-Agent Intelligent Simulation Laboratory (MISL) Research Unit
Department of Information Technology
Faculty of Informatics
Mahasarakham University
Khamriang Sub-District, Kantarawichai District, Mahasarakham 44150, Thailand
olarik.s@msu.ac.th

[3]Department of Mathematics
Faculty of Applied Science
[4]Intelligent and Nonlinear Dynamic Innovations Research Center
Science and Technology Research Institute
King Mongkut's University of Technology North Bangkok
1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800, Thailand
anuchit.j@sci.kmutnb.ac.th

Received April 2023; accepted June 2023

Abstract. *Applying wearable sensors to recognize human activities has developed as an emergent topic in the field of research on artificial intelligence. This is because human activity recognition (HAR) implementations extend from intelligent and sophisticated healthcare applications to other areas like innovative home surveillance systems and exercise performance monitoring devices. Exercise activity recognition (EAR) is a subclass of HAR investigating complicated human movement sequences. Literature evaluations indicate that understanding multi-modal sensors of diverse data kinds has various obstacles. We investigated multi-modal EAR utilizing deep learning techniques utilizing sensor data from many body areas. Focusing on accelerometer data, we proposed the hybrid model with the combination of a deep convolutional neural network (CNN) and long short-term memory (LSTM) neural network (called CNN-LSTM) for effectively recognizing fitness activities. The trained deep learning classifier's accuracy, loss, and F1-score were determined using a public standard EAR dataset (MEx dataset) to assess the newly proposed classifier. We inferred from experimental findings that the proposed CNN-LSTM could classify exercise activities utilizing accelerometer data from object location with the most significant accuracy (97.23%) and F1-score (97.20%), surpassing existing baseline classifiers.*

Keywords: Exercise activity recognition, Deep learning, Wearable sensor, Accelerometer, Multi-modal HAR

1. **Introduction.** In the realm of human-computer interaction, there has been a significant amount of attention given to the human activity recognition (HAR) in recent decades [1]. HAR focuses on developing methods and techniques to identify and categorize human actions using sensor information without human intervention. HAR aims to enable

DOI: 10.24507/icicel.18.01.69

machines to understand and interpret human actions and movements in real-world environments. The three primary ways of recognizing human activities are vision, non-vision (sensor), and hybrid integration [2]. Vision-based methods commonly use video and depth cameras, and wearable camera technology has recently emerged as a promising option [3, 4, 5, 6, 7]. Sensor-based approaches can be categorized as wearable, ambient, or hybrid, depending on the position of the sensors [8, 9, 10]. Each approach has pros and cons, and in this study, we are constrained to wearable sensors.

Exercise activity recognition (EAR), one topic study of the HAR, refers to the recognition of physical activity performed intentionally to improve or maintain physical fitness, health, and overall well-being. This encompasses multiple application domains, including calisthenics, weight training, yoga, and sports [11, 12]. Inertial measurement units (IMUs) are the most prevalent data source in the published EAR research [13, 14, 15]. EAR commonly classifies numerous discrete labels based on sensor data streams. Typically, these identification techniques employ a manual feature extraction pipeline and a classification method such as k-Nearest Neighbors, Random Forest, Decision Trees, or Hidden Markov Model. EAR has two key applications: automatic logging of exercises to reduce manual data entry by trainers, and providing real-time access to unbiased records for coaches and doctors [16]. This study is restricted to physical activities, specifically physical exercises, which can be defined as any activity that improves or maintains an individual's health and fitness.

Initially, the categorization of sensor data for EAR consisted of three steps: data gathering, pre-processing, and classification. The pre-processing workflow comprises two components: segmentation to produce data instances from receiving sensor streaming data and feature extraction. Windowing is the most used form of segmentation, in which a fixed temporal window is utilized to divide the sensor stream of data into data instances [17]. Subsequently, each data instance is turned into features using a predefined set of time-domain, frequency-domain, and spatial feature extraction methods. Nevertheless, manual construction and the selection of features is laborious, yet these approaches have obtained great effectiveness with sparse information for customized EAR applications.

Current deep learning (DL) methods have merged feature extraction and classification processes, where the learning of features is constrained by iterative optimization [18, 19, 20]. As with earlier techniques, windowing is utilized to obtain data instances; alternatively, feature transformation techniques are also implemented. These feature transformation approaches comprise the transformation of time-series data in the frequency domain [21]. While DL approaches (convolutional neural network (CNN) and long short-term memory (LSTM)) are state-of-the-art in HAR investigations [22], these are rarely addressed in EAR [23]. For example, the researchers [13] utilize recurrent structures to recognize the shoulder rehabilitation process from wrist-worn IMU data streams with an accuracy of 88.9%; their dataset is not available to the public. Furthermore, their approaches cannot be reassigned to other exercise contexts due to a complete lack of sensors that recognize motion patterns from body parts other than the wrist. In direct opposition, the publicly available MEx dataset [23] earlier revealed single detection capability on physiotherapy exercise identification employing four sensors; two accelerometers placed on the wrist and thigh, a pressure mat, and a depth camera with F1-score of 63.35%, 90.15%, 74.08%, and 87.20%, respectively. These findings indicate that more than a single sensor is required to identify various exercises accurately. In this perspective, we emphasize the necessity for multi-modal learning methodologies to boost efficiency in classification techniques. This necessitates fusion structures and techniques, such as attention, to merge heterogeneous multi-modal sensor data.

In this study, we focus on further EAR research by providing innovative DL methods and emphasizing the importance of designing EAR solutions that are unobtrusive and straightforward to implement in the actual world. We propose a novel hybrid deep learning

approach called CNN-LSTM for EAR. This model combines a deep CNN and LSTM neural network for effectively identifying fitness movements. The experimental findings indicate that the proposed combination model achieved significant results with accuracy and F1-score measures.

The remainder of the article is structured as follows. Section 2 describes the CNN-LSTM model utilized in this study in depth. Section 3 demonstrates our experimental outcomes using a publicly available benchmark dataset. In addition, this section compares the proposed and baseline deep learning models' performance. Section 4 summarizes this work and outlines potential future activities.

2. **The Proposed Sensor-Based EAR Framework.** In this study, a sensor-based workflow for EAR is employed, which consists of five primary stages: obtaining data, preparing data, segmenting data, developing a model, and refining the model, as depicted in Figure 1.

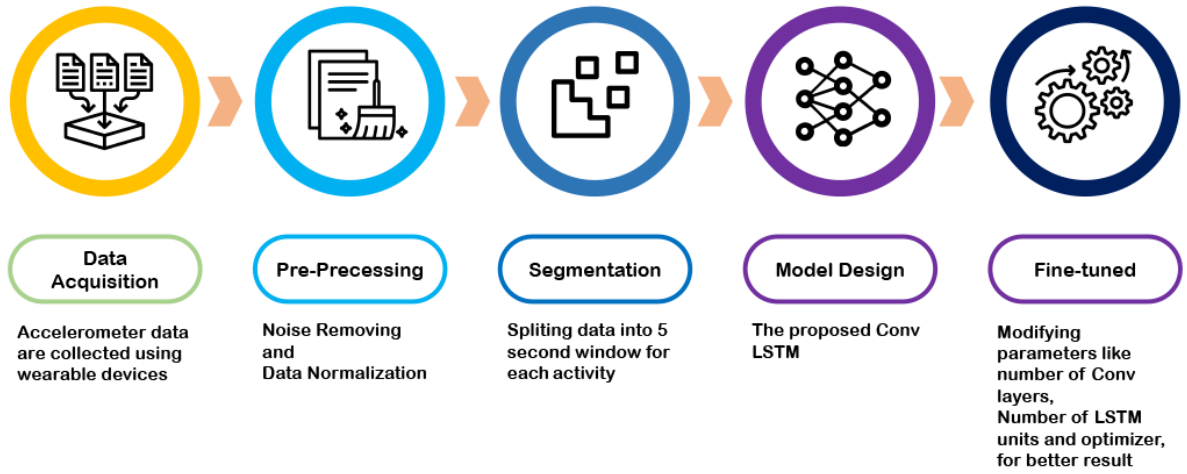

FIGURE 1. The EAR workflow based on wearable sensors used in this work

2.1. **Multi-modal exercise dataset.** This study utilized the publicly available Multi-modal Exercise dataset (abbreviated MEx dataset) for EAR [23]. The necessity to recognize and evaluate the quality of exercise performance in participants with musculoskeletal disorders (MSD) encouraged the compilation of this dataset. The MEx dataset chose seven activities commonly prescribed for MSD individuals by physiotherapists and gathered information with four sensors, as indicated in Table 1: a pressure mat, a depth camera, and two accelerometers. The dataset incorporates three data types, numerical time series, video, and pressure sensor data, offering fascinating research problems for EAR and exercise quality evaluation.

TABLE 1. Details of sensors operated for gathering exercise activities in the MEx dataset

Sensor	Product	Details
Depth Camera Sensor	Obbrec Astra Depth Camera	Frame rate = 15 fps Frame resolution = 320×240
Pressure Sensor	Sensing Tex Pressure Mat	Frame rate = 75 fps Frame resolution = 32×16
Accelerometer	Axivity AX2 3-Axis Logging Accelerometer	Sample rate = 100 Hz Accelerometer range = ± 8 g

In the dataset, two accelerometers will be attached to the wrist and thigh of each of the 30 participants; the pressure mat will be utilized as an exercise mat on which the recipient will execute activities; and the depth camera will be positioned above the participant, capturing an aerial perspective. Throughout the remainder of this work, the accelerometer on the thigh, the wrist, the pressure mat, and the depth camera will be designated as ACT, ACW, PM, and DC, respectively.

2.2. **Data pre-processing.** Raw sensor data were adjusted in data pre-processing: noise reduction and data normalization. In our investigation, we used averaged smoothing filters over all three dimensions of the accelerometer sensor to remove noise from the data. Next, the sensor data are standardized, which supports resolving the model learning difficulty by determinating all data values into a close range. As a consequence, gradient descents could converge more rapidly. The normalized data were then separated employing 5-second fixed-width sliding windows with a 50% overlap.

2.3. **The proposed CNN-LSTM model.** CNN is particularly successful in extracting and learning the characteristics of one-dimensional sequence data, such as univariate time series data, when necessary [24]. In addition, it is feasible to deploy the CNN model in the hybrid arrangement in conjunction with an LSTM backend. CNN analyzes the input subsequences, which are then transmitted sequentially to the LSTM model for further comprehension.

The hybrid proposed model is known as the CNN-LSTM model (shown in Figure 2), and its design extracts characteristics from input data using CNN layers, while the LSTM element enables sequence forecasting. The CNN-LSTM model can interpret subsequences produced from the main sequence in the format of blocks by first extracting the significant characteristics from each block, followed by LSTM interpretation of those characteristics.

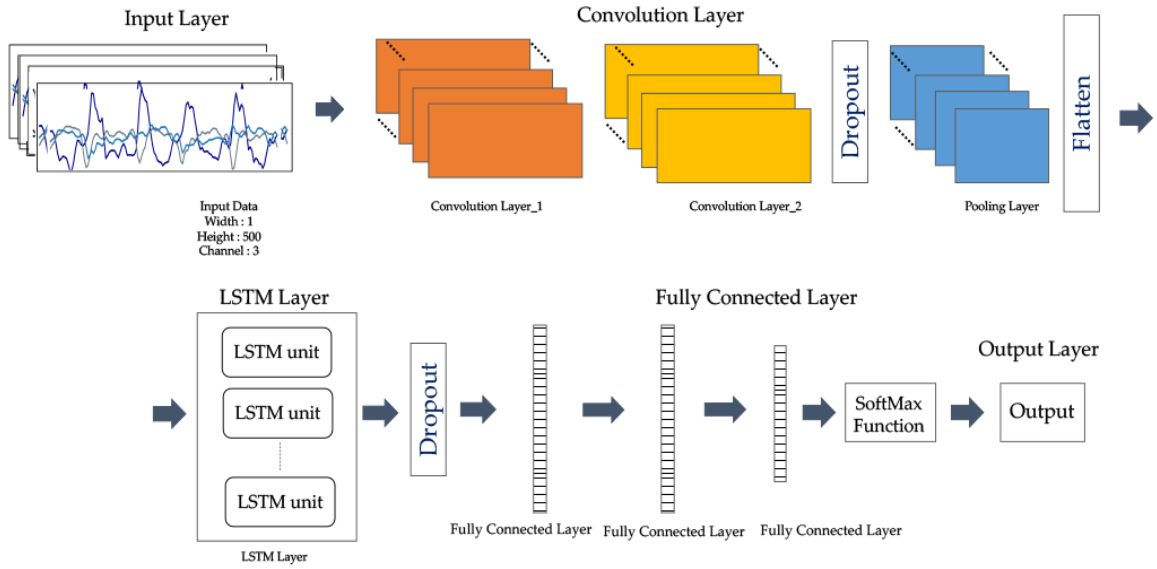

FIGURE 2. Architecture of CNN-LSTM model

This study utilizes a CNN layer to identify local features in time series data once pre-processing has been completed. The one-dimensional CNN is an effective tool for time series analysis as its convolution kernel operates in a single direction, making it possible to extract previously undiscovered data features in the time domain. The LSTM model is then used to receive the extracted features from the encoder (CNN) and incorporate them as input. Throughout the training process, the different gates of the LSTM network and the training data are continually adjusted so that the LSTM model can identify the relationships between the input and output sequence.

The proposed method employs an encoder-decoder model. Unlike other deep learning models such as LSTM and CNN, the CNN-LSTM model does not directly produce a vector series. Instead, it comprises two parts: an encoder that reads and encodes the input sequence and a decoder that reads the encoded input sequence and predicts the output sequence one step at a time. A straightforward but efficient CNN architecture is used for the encoder, consisting of two convolutional layers and a max-pooling layer. The input sequence is read by the first convolutional layer and projected onto feature maps, which are then processed by the second layer to amplify significant features. The kernel size for reading input sequences is two-time steps, with 64 feature maps in the convolutional layer. The max-pooling layer selects the top 1/4 signal values to simplify the feature maps.

The flattened feature maps generated by the pooling layer are input for the decoder model, consisting of an LSTM hidden layer with 100 units. This layer decoded the input sequence and output a vector for each unit that captures its features. The input sequence representation is duplicated for each time step in the output sequence and provided to the LSTM decoder to generate the output sequence. A fully connected layer is applied before the final output layer to interpret each time step in the output sequence. This approach ensures that each step in the output sequence receives the same layers, allowing the LSTM decoder to determine the context needed for each step and interpret them independently while still using the same weights. Hyperparameters such as filter number, kernel size, pool size, and dropout ratio were determined by Bayesian optimization. All details of CNN-LSTM hyperparameters are listed in Table 2.

Table 2. Hyperparameters details of the proposed CNN-LSTM network in this work

Stage	Hyperparameters		Values
Architecture	Convolution-1	Kernel Size	3
		Stride	1
		Filters	64
	Convolution-2	Kernel Size	3
		Stride	1
		Filters	64
	Dropout-1		0.5
	Maxpooling		2
	LSTM Neuron		100
	Dropout-2		0.5
	Dense		100
Training	Loss Function		Cross-entropy
	Optimizer		Adam
	Batch Size		64
	Number of Epoches		200

2.4. **Performance measurement indicators.** The proposed deep learning model is evaluated using a 5-fold cross-validation procedure that calculates four quality assessment indicators: accuracy, precision, recall, and F1-score. The equations for the four measures are as follows in mathematics:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_N + F_P} \tag{1}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{2}$$

$$Recall = \frac{T_P + T_N}{T_P + F_N} \tag{3}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Typically, HAR is measured by the following four indicators of effectiveness. The identification is a true positive (T_P) for the group under consideration, but for all other groups, it is a true negative (T_N). It is possible for sensor data that belongs to one group to be incorrectly identified as pertaining to another, resulting in a false positive (F_P). Conversely, a false negative (F_N) identification can occur if input from an activity sensor that belongs to a different group is incorrectly labeled as pertaining to the same group.

3. **Experiments and Results.** Including the proposed multi-resolution CNN, this part evaluates three foundational deep learning models (CNN, LSTM, and CNN-LSTM) by describing the experimental setup and demonstrating the findings.

3.1. **Experimental environment.** To carry out the experiments in this research, the Google Colab Pro platform and a Tesla-V100 were utilized. The Python application used in the study was created by implementing various libraries, including Python 3.6.9, TensorFlow 2.2.0, Keras 2.3.1, Scikit-learn, NumPy 1.18.5, and Pandas 1.0.5. The dataset was manipulated using NumPy libraries for matrix operations, Pandas libraries for CSV file manipulation, and Scikit-learn for class-wise sampling across the training, testing, and validation datasets.

3.2. **Experimental results.** Table 3 demonstrates the recognition interpretation outcomes of DL models operating accelerometer data for thigh position. This experiment indicates that the proposed CNN-LSTM model achieved the most satisfactory effectiveness with the highest F1-score of 97.20%.

Table 3. Recognition interpretation outcomes of DL models using accelerometer data from thigh position

Model	**Recognition performance**		
	Accuracy	**Loss**	**F1-score**
CNN	94.79% (±1.205%)	0.25 (±0.039)	94.72% (±1.262%)
LSTM	60.92% (±3.862%)	0.98 (±0.161)	59.75% (±3.688%)
CNN-LSTM	97.23% (±0.524%)	0.13 (±0.025)	97.20% (±0.534%)

As shown in Figure 3, the obtained findings demonstrate that the proposed CNN-LSTM network can execute efficient classification exercise activities with significant F1-scores of more than 90%.

Table 4 shows the outcomes of applying accelerometer data for wrist position in DL models designed for recognition. The outcomes of this study demonstrate that the proposed CNN-LSTM model performed at a superior level, with an F1-score of 91.69%.

Table 4. Recognition interpretation outcomes of DL models using accelerometer data from wrist position

Model	**Recognition performance**		
	Accuracy	**Loss**	**F1-score**
CNN	80.69% (±3.099%)	1.000 (±0.238)	80.49% (±3.134%)
LSTM	43.42% (±10.982%)	1.230 (±0.235)	41.34% (±11.640%)
CNN-LSTM	91.72% (±1.515%)	0.324 (±0.040)	91.69% (±1.524%)

FIGURE 3. A confusion matrix of the CNN-LSTM trained by accelerometer data from thigh position

FIGURE 4. A confusion matrix of the CNN-LSTM trained by accelerometer data from wrist position

The CNN-LSTM trained with accelerometer data from the wrist position can categorize exercise activities with a lower F1-score than the CNN-LSTM trained with accelerometer data from the thigh position, as considering a confusion matrix in Figure 4.

4. **Conclusion and Future Works.** This work investigated multi-modal EAR utilizing accelerometer data, a vital element for automating digital interventions that give appropriate guidance and assistance. To automate EAR, we proposed CNN-LSTM, a novel hybrid CNN and LSTM neural network that conducts EAR using sensor input from two body sites (thigh and wrist). The suggested network considerably surpasses multiple benchmarks and efficiently learns modality combinations suited to identify various activities with the most outstanding accuracy of 97.23% and F1-score of 97.20%.

For the activity recognition future studies of multi-modal exercise, we aim to propose applying different DL networks, such as ResNeXt, InceptionTime, and Temporal Transformer. In addition, a data augmentation is an engaging approach for enhancing models using unbalanced datasets. The issue can be resolved using this strategy.

Acknowledgment. This research is supported by Thailand Science Research and Innovation Fund and University of Phayao (Grant No. FF66-UoE001).

REFERENCES

[1] Y. Wang, S. Cang and H. Yu, A survey on wearable sensor modality centred human activity recognition in health care, *Expert Systems with Applications*, vol.137, pp.167-190, 2019.

[2] A. Das Antar, M. Ahmed and M. A. R. Ahad, Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: A review, *Proc. of the 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Spokane, WA, USA, pp.134-139, 2019.

[3] A. Garg, S. Nigam and R. Singh, Vision based human activity recognition using hybrid deep learning, *Proc. of the 2022 International Conference on Connected Systems & Intelligence (CSI)*, Kerala, India, pp.1-6, 2022.

[4] S. Gonwirat and O. Surinta, DeblurGAN-CNN: Effective image denoising and recognition for noisy handwritten characters, *IEEE Access*, vol.10, pp.90133-90148, 2022.

[5] K. Fithriasari and U. S. Nuraini, Face identification using multi-layer perceptron and convolutional neural network, *ICIC Express Letters*, vol.15, no.2, pp.157-164, 2021.

[6] A. G. D'Sa and B. G. Prasad, A survey on vision based activity recognition, its applications and challenges, *Proc. of the 2019 2nd International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, Gangtok, India, pp.1-8, 2019.

[7] U. A. Usmani, J. Watada, J. Jaafar, I. A. Aziz and A. Roy, Particle swarm optimization with deep learning for human action recognition, *International Journal of Innovative Computing, Information and Control*, vol.17, no.6, pp.1843-1870, 2021.

[8] S. Mekruksavanich, A. Jitpattanakul, K. Sitthithakerngkiet, P. Youplao and P. Yupapin, ResNet-SE: Channel attention-based deep residual network for complex activity recognition using wrist-worn wearable sensors, *IEEE Access*, vol.10, pp.51142-51154, 2022.

[9] S. Mekruksavanich, N. Hnoohom and A. Jitpattanakul, A hybrid deep residual network for efficient transitional activity recognition based on wearable sensors, *Applied Sciences (Switzerland)*, vol.12, no.10, pp.1-21, 2022.

[10] S. Mekruksavanich and A. Jitpattanakul, Deep residual network for smartwatch-based user identification through complex hand movements, *Sensors*, vol.22, no.8, 2022.

[11] W. Zhang, C. Su and C. He, Rehabilitation exercise recognition and evaluation based on smart sensors with deep learning framework, *IEEE Access*, vol.8, pp.77561-77571, 2020.

[12] S. Mekruksavanich and A. Jitpattanakul, Sport-related activity recognition from wearable sensors using bidirectional GRU network, *Intelligent Automation & Soft Computing*, vol.34, no.3, pp.1907-1925, 2022.

[13] D. Burns, E. Leung, M. Hardisty, C. Whyne, P. Henry and S. McLachlin, Shoulder physiotherapy exercise recognition: Machine learning the inertial signals from a smartwatch, *Physiological Measurement*, vol.39, 2018.

[14] M. Guo, Z. Wang and N. Yang, Aerobic exercise recognition through sparse representation over learned dictionary by using wearable inertial sensors, *Journal of Medical and Biological Engineering*, vol.38, pp.544-555, 2018.

[15] L. N. N. Nguyen, D. Rodríguez-Martín, A. Català, C. Pérez-López, A. Samà and A. Cavallaro, Basketball activity recognition using wearable inertial measurement units, *Proc. of the XVI International Conference on Human Computer Interaction*, New York, USA, pp.1-6, 2015.

[16] A. Hussain, K. Zafar, A. R. Baig, R. Almakki, L. AlSuwaidan and S. Khan, Sensor-based gym physical exercise recognition: Data acquisition and experiments, *Sensors*, vol.22, no.7, 2022.

[17] I. P. E. S. Putra and R. Vesilo, Window-size impact on detection rate of wearable-sensor-based fall detection using supervised machine learning, *Proc. of the 2017 IEEE Life Sciences Conference (LSC)*, Sydney, NSW, Australia, pp.21-26, 2017.

[18] S. Dara and P. Tumma, Feature extraction by using deep learning: A survey, *Proc. of the 2018 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, pp.1795-1801, 2018.

[19] S. Noppitak and O. Surinta, dropCyclic: Snapshot ensemble convolutional neural network based on a new learning rate schedule for land use classification, *IEEE Access*, vol.10, pp.60725-60737, 2022.

[20] Y. Chen, H. Jiang, C. Li, X. Jia and P. Ghamisi, Deep feature extraction and classification of hyperspectral images based on convolutional neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, vol.54, no.10, pp.6232-6251, 2016.

[21] S. Sani, S. Massie, N. Wiratunga and K. Cooper, Learning deep and shallow features for human activity recognition, *Proc. of 2017 International Conference on Knowledge Science, Engineering and Management (KSEM 2017)*, Melbourne, Australia, pp.469-482, 2017.

[22] F. J. Ordóñez and D. Roggen, Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition, *Sensors*, vol.16, no.1, 2016.

[23] A. Wijekoon, N. Wiratunga and K. Cooper, *MEx: Multi-Modal Exercises Dataset*, IEEE DataPort, https://dx.doi.org/10.21227/h7g2-a333, 2019.

[24] A. Wibawa, A. B. Putra Utama, H. Elmunsyah, U. Pujianto, F. Dwiyanto and L. Hernandez, Time-series analysis with smoothed convolutional neural network, *Journal of Big Data*, vol.9, 44, 2022.