DOI: 10.15849/IJASCA.250730.20

Int. J. Advance Soft Compu. Appl, Vol. 17, No. 2, July 2025
Print ISSN: 2710-1274, Online ISSN: 2074-8523
Copyright © Al-Zaytoonah University of Jordan (ZUJ)

Deep Learning-Based Monitoring System for Early Childhood Motor Skill Development

Lina Lina[1], Arlends Chris[2], and Ranny Ranny[3]

[1]Faculty of Information Technology, Tarumanagara University, Indonesia
e-mail: lina@untar.ac.id
[2]Faculty of Medicine, Tarumanagara University, Indonesia
e-mail: arlendsc@fk.untar.ac.id
[3]Data Science Department, Bunda Mulia University, Indonesia
e-mail: ranny@bundamulia.ac.id

Abstract

Early childhood education plays a vital role in supporting the cognitive and motor development of children aged zero to six years. However, due to limited verbal communication skills at this stage, young children are often unable to accurately report their school activities, presenting challenges for parental monitoring. To address this issue, this research proposes the development of a Deep Learning-based system capable of automatically analyzing and recording children's daily activities within children education institutions. The system aims to facilitate real-time access to developmental records by both parents and educators. This system received input in the form of images of early childhood activities from video recordings and videos in real time at school. The methodology employs BlazePose for human pose estimation and a modified Convolutional Neural Network (CNN) for classifying activities based on image datasets. The expected output is a software product in the form of an intelligent system to automatically monitor the development of early childhood motor skills. This study focused on activities classified as gross motor skills, which engage the body's large muscle groups and facilitate movements. Examples of such activities include sitting, standing, and sleeping. Experimental results showed that the proposed system demonstrated effective performance, achieving the highest accuracy of 97.77%. Some errors occurred due to dependence on camera angles and the similarity of poses across different viewpoints.

Keywords: *Children Monitoring, BlazePose, Children Activity Detection, Human Pose Estimation, Convolutional Neural Network.*

1 Introduction

Monitoring is an activity that aims to observe an event that occurs within a certain period. Monitoring activities are often carried out with the aim of recording or documenting events that occurred at that place and time. This recording process certainly has constraints and limitations because monitoring generally must be done manually or directly by humans. It is not uncommon for the monitoring process to be carried out with observers going directly to the field. In this increasingly advanced technological world, monitoring can be carried out using technological assistance with an image or video recording hardware. This image or video recording media is capable of recording events that occur in the field automatically at a specified place. Some examples of recording

Received 25 May 2025; Accepted 22 July 2025

media that can be used are Closed Circuit Television (CCTV), web cameras, smartphone cameras, etc. By utilizing the results of this recording, monitoring activities can be carried out to see the results of the recording at the desired time. However, these results must be observed sequentially so that it takes quite a long time if the search is done manually. Thus, this method is inefficient because it takes a lot of time and energy to observe the contents of the entire video for quite a long duration.

Beyond the security considerations of residential and public environments, educational settings—particularly early childhood institutions—require systematic monitoring of human activities. This need arises not only from safety concerns but also from parental interest in the day-to-day experiences of their children. However, direct access to classrooms or regular review of video recordings of classroom activities is often impractical or restricted. According to Papalia and Martorell [1], children in the preschool stage exhibit rapid development in gross motor skills—such as running and jumping—that engage large muscle groups. As their musculoskeletal systems strengthen and pulmonary capacity increases, children become capable of executing these activities with greater speed and endurance. Moreover, young children generally lack the cognitive and linguistic capacity to accurately recount their daily experiences, making external observation essential for understanding their behavior and development [2,3]. To address this limitation, this study proposes the development of an intelligent monitoring system capable of automatically assessing motor skill development in early childhood. Such a system would enable continuous, real-time access for parents, educators, and relevant authorities, thereby enhancing oversight of developmental milestones. The proposed solution employs Artificial Intelligence algorithms to detect and classify motor skill activities autonomously, thereby eliminating the need for manual observation and reporting by teachers or administrative personnel.

Despite the growing number of studies on Human Activity Recognition (HAR), most of the existing research has focused predominantly on adult subjects. This presents a significant challenge in recognizing activities performed by children, primarily due to the reliance on adult-centric datasets and the scarcity of child-specific references in the literature. Several prior studies have primarily focused on the characterization of activities in children with specific medical conditions [4-10], with limited attention given to healthy and typically developing children. Factors such as high articulation variability, small and less discernible joints, occlusions due to clothing, diverse lighting conditions, and motion blur contribute to the increased difficulty of accurately recognizing children's activities. To address these challenges, the proposed system is specifically designed to advance research in child-centered HAR.

Human Activity Recognition (HAR) can be broadly categorized into two main approaches: sensor-based techniques and vision-based techniques [11]. Sensor-based methods typically involve the use of wearable or embedded devices to capture motion or physiological data [12]. However, these techniques are often considered less practical for real-world deployment due to the requirement for additional hardware, which must be worn or carried by the user [13]. This not only introduces logistical constraints but also compromises user comfort and convenience.

In contrast, vision-based HAR has emerged as a viable alternative, leveraging visual data captured by cameras to recognize and classify human activities. Unlike sensor-based approaches, vision-based systems enable non-intrusive activity monitoring, eliminating the need for body-mounted devices and allowing for immediate detection through image or video analysis. Vision-based HAR studies, such as [14–18], typically utilize raw image data captured directly from camera devices. More recent advancements in HAR involve the use of wearable sensors, either individually or in combination, attached directly to the human body. These include 3D triaxial accelerometers, magnetometers, gyroscopes, Radio Frequency Identification (RFID) devices, and Global Positioning System (GPS) sensors [19–23]. These sensors capture detailed biomechanical and spatial data reflecting body movement. The HAR process in such systems involves the acquisition of movement data by the sensors, which is then analyzed by recognition algorithms to classify and interpret the specific activities being performed. Collectively, these sensor modalities enhance the precision and applicability of HAR across diverse real-world scenarios. Moreover, a well-known example of such a system is Microsoft Kinect [24–26], which employs depth-sensing and skeletal tracking capabilities to detect up to 20 key joint positions on the human body, enabling the classification of various physical activities. Numerous studies in HAR also utilize skeletal data for activity identification due to its detailed representation of human posture and movement [27–30]. However, the generation and processing of skeletal data demand substantial computational resources, which can limit the feasibility of real-time implementation and deployment on resource-constrained platforms [31,32]. To address these limitations, Albukhary and Mustafah [33] proposed a computationally efficient human activity recognition method that relies solely on metrics derived from human movement distance and aspect ratio.

2 Research Methodology

The research methodology employed in this study consists of two stages: (1) object detection focusing on early childhood subjects through skeleton-based edge detection, and (2) activity classification employing a Convolutional Neural Network (CNN) architecture. The proposed system accepts image inputs directly from a camera or video source. These images are processed using BlazePose framework to extract skeletal keypoints, which are then transformed into a two-dimensional matrix representation suitable for input to a trained Convolutional Neural Network (CNN) model. Activity prediction within the system is performed by averaging the CNN outputs over the most recent three frames, with frame selection occurring at one-second intervals, starting from the initial frame of each second. The overall system workflow is illustrated in Fig. 1.

The pose estimation method employed in this system is BlazePose [34,35]. BlazePose utilizes a modified stacked-hourglass architecture to predict skeletal keypoints from human subjects. The architecture comprises an encoder-decoder network designed to generate heatmaps for all joint locations, followed by an additional encoder module that refines these heatmaps to accurately estimate the coordinates of each specific joint. BlazePose employs a detector-tracker framework wherein human detection is performed in the initial frame by identifying the region of interest (ROI) corresponding to the

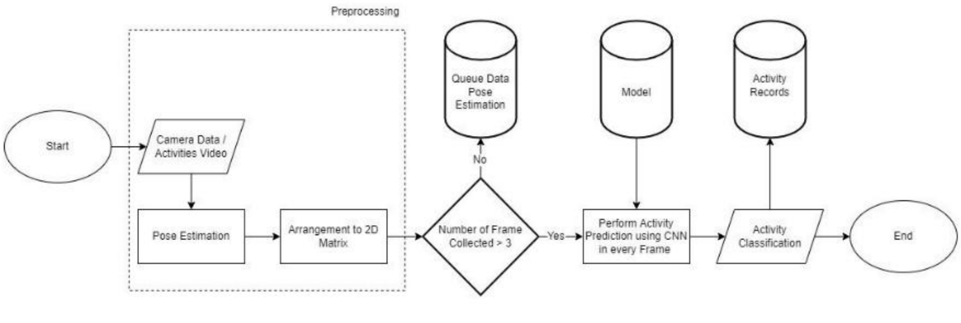

Preprocessing
Start
Camera Data / Activities Video
Pose Estimation
Arrangement to 2D Matrix
Queue Data Pose Estimation
No
Number of Frame Collected > 3
Yes
Model
Perform Activity Prediction using CNN in every Frame
Activity Records
Activity Classification
End

Figure 1. Diagram of the Proposed Human Activity Recognition System.

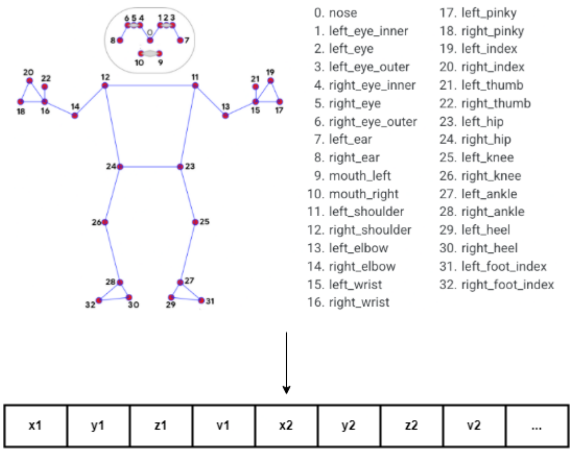


Figure 2. The Location of 33 BlazePose Keypoints.

subject. In subsequent frames, BlazePose utilizes a tracking mechanism to follow the detected human within the previously established ROI. If tracking fails or the human object is not detected, the detection process is reinitiated. The system detects a total of 33 keypoints representing anatomical landmarks, the spatial configuration of which is illustrated in Fig. 2. Each keypoint detected by BlazePose is represented by four values: the x, y, and z coordinates, along with a visibility score. The visibility score, derived from the model's confidence estimation, indicates whether a keypoint is occluded or detected with low accuracy. This mechanism enables BlazePose to robustly estimate keypoints even when certain points are partially outside the frame or obscured.

In the next stage, the Convolutional Neural Networks (CNN) architecture is constructed to identify human activities. Unlike traditional artificial neural networks, CNN requires minimal preprocessing of input data, as they inherently possess the capability to extract complex hierarchical features directly from raw images, provided that sufficient training data is available [28]. The CNN method consists of two main stages, namely the feature training stage consisting of convolutional layers, ReLU (activation function) and pooling layers, while the classification stage consists of the flattening process, fully connected layers, and predictions. Each part of CNN has two main processes, namely feedforward and backpropagation. The first stage in the CNN method is the convolutional layer which is a layer that performs feature extraction that is connected to the local area of the input image. The equation used in the convolutional layer calculation process is as follows [36]:

$$x(i,j) = \sum_m \sum_n w^l_{m,n} * o^{l-1}_{i+m,j+n} + b \qquad (1)$$

where *x(i, j)* is the result of the convolution calculation at position *(i, j)*, *l* is the layer, *o(i,j)* is the input value, *w(m,n)* is the filter, i is the bias, and *i* and *j* represent the row and column of pixels in the image, respectively.

The second stage is the pooling layer which is a layer that functions to reduce the size of the previous layer (downsampling) in the spatial dimension (width, height). There are 2 types of pooling layers, namely max pooling and average pooling. The max pooling process is carried out by searching and applying the maximum value of each part on the feature map, while the average pooling process calculates the average value of each part on the feature map. Max pooling is the most frequently used method including in this paper. Furthermore, ReLU is an activation function that is responsible for being able to normalize the values generated from the convolutional layer. The ReLU stage is the value normalization stage. ReLU will display the value directly if the value is positive while for negative values it will be given a value of zero.

In the classification stage, the first process that occurs is flattening which will change the feature map in the previous layer into a one-dimensional vector. Furthermore, the next stage is the formation of a fully-connected layer for linear classification on CNN. In the fully-connected layer, each neuron has a full connection to all neurons in the previous layer. The output of the fully-connected layer in the form of a y value with weight parameters W and bias b from an input x can be seen in the following formula:

$$y = \sum_i x_i \times W_i + b \tag{2}$$

Softmax is an activation function that will be used in the output layer. The function of softmax is to take a number from the input vector that has gone through the fully-connected layer process and change the number into a range of 0 to 1. The output layer has many similarities with the fully-connected layer, the difference between the two is that the output layer uses the softmax activation function and the fully-connected layer uses the ReLU activation function. The softmax activation function can be seen in the following formula:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{3}$$

where p_i is the probability of the i-th vector, z_i is the i-th input vector, and z is the input vector.

The CNN architecture implemented in this paper is a custom-designed model, with its detailed configuration layers are presented in Table 1.

Table 1: The Proposed CNN Architecture.

Layer Type	Filter/Unit	Size/Padding	Activation Function
Conv2D	32	(3, 3) / same	ReLU
Conv2D	32	(3, 3) / same	ReLU
Conv2D	64	(3, 3) / same	ReLU
Flatten	-	-	-
Dense	32	-	ReLU
Dense	3	-	Softmax

3 Experimental Results and Discussion

In this experiment, a dataset comprising children's activities was compiled from private early childhood education institutions in Indonesia and publicly accessible online sources. Researchers conducted site visits to private schools in Indonesia to recruit parent volunteers who provided informed consent to participate in the study. Children participants were recorded while performing a series of predefined activities under the supervision of the research team. Data were collected in a designated student playroom with camera placement optimized to capture the participant's full body in unobstructed motion. A total of 150 videos were recorded, comprising 30 children aged 6 to 8 years, each performing three distinct activities. Ethical approval for this study was obtained from Universitas Tarumanagara Human Research Ethics Committee (Approval No.001-UTHREC/UNTAR/II/2025). The dataset was subsequently augmented through geometric transformations including translation, rotation, and occlusion to enhance data variability and robustness. The augmentation process was performed using image processing techniques, including affine transformations to simulate translation and rotation effects, as well as motion blur applied through Gaussian filtering. The activity categories consist of standing, sitting, and sleeping. The dataset was partitioned into 75% for training and 25% for testing purposes. Detailed information regarding the quantity and distribution of samples across each activity class is provided in Table 2.

The proposed model was evaluated across a range of hyperparameter settings to assess its performance. Hyperparameter tuning was conducted iteratively, beginning with the learning rate, followed by batch size, and finally the number of epochs. For each stage, the optimal value identified was fixed and utilized in subsequent tests of the remaining hyperparameters. The results of the hyperparameter evaluations are summarized in Table 3.

Table 2: Training and Testing Data Distribution.

Class Type	Amount of Training Data	Amount of Test Data
Sleeping	2861	521
Standing	5827	1005
Sitting	4826	859
Total	13514	2385

Table 3: Hyperparameter Tuning Results.

Hyperparameter	Values	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
Learning rate	0.01	96.60%	0.1047	95.22%	0.1519
	0.001	99.14%	0.0251	97.65%	0.0935
	0.001	97.56%	0.0667	96.60%	0.0946
Batch size	16	99.19%	0.0239	97.61%	0.1218
	32	99.21%	0.0243	97.02%	0.1013
	64	98.61%	0.0243	96.44%	0.0965
	128	99.07%	0.0275	97.40%	0.0729
Epoch	32	99.19%	0.0234	97.78%	0.1162
	75	99.75%	0.0113	97.69%	0.1635
	100	100%	1.67e-6	98.20%	0.1966

Based on the experimental results, the optimal hyperparameter configuration was identified with a learning rate of 0.001, a batch size of 16, and 32 training epochs. Performance metrics, including overall accuracy graph, overall loss graph, accuracy graph for each category, and loss graph for each category are presented in Figures 3 through Figure 6. During the testing phase, the evaluation was conducted using two distinct datasets. The first dataset comprises images of children's activities that were not included in the training set, ensuring the assessment of the model’s generalization capability. The second dataset consists of videos of children's activities, composed of multiple image frames arranged in varying activity sequences, enabling the evaluation of the model's performance in more dynamic and temporally dependent scenarios. Detailed descriptions of Dataset 2 are provided in Table 4. The evaluation procedure involves frame-level prediction alongside a temporal smoothing technique, which computes the average prediction over the three most recent frames.

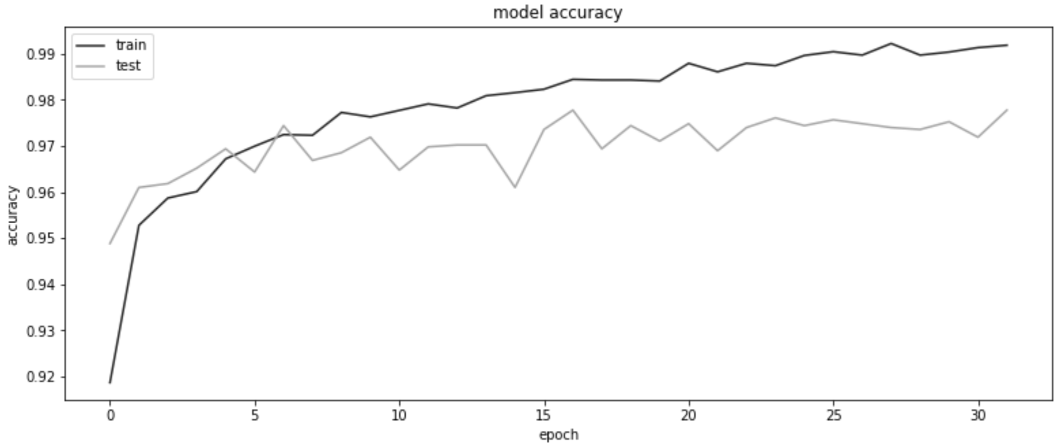


Figure 3. The Overall Accuracy Graph.

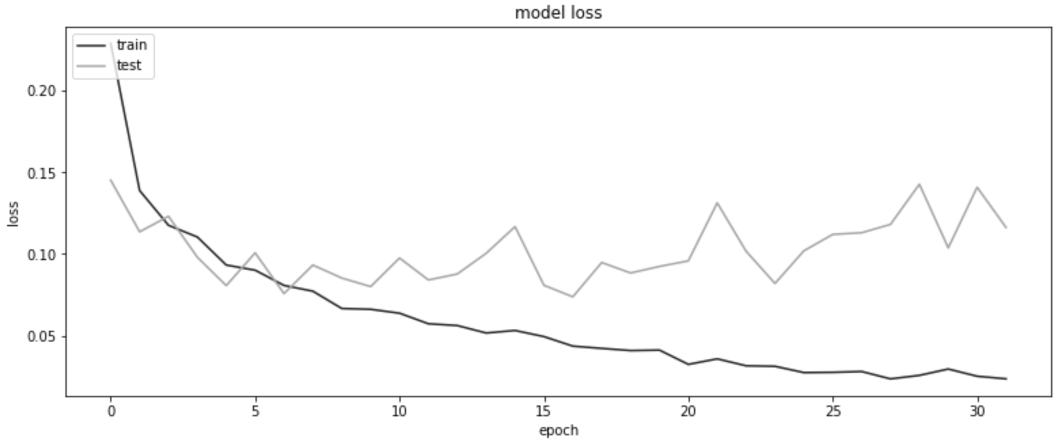


Figure 4. The Overall Loss Graph.

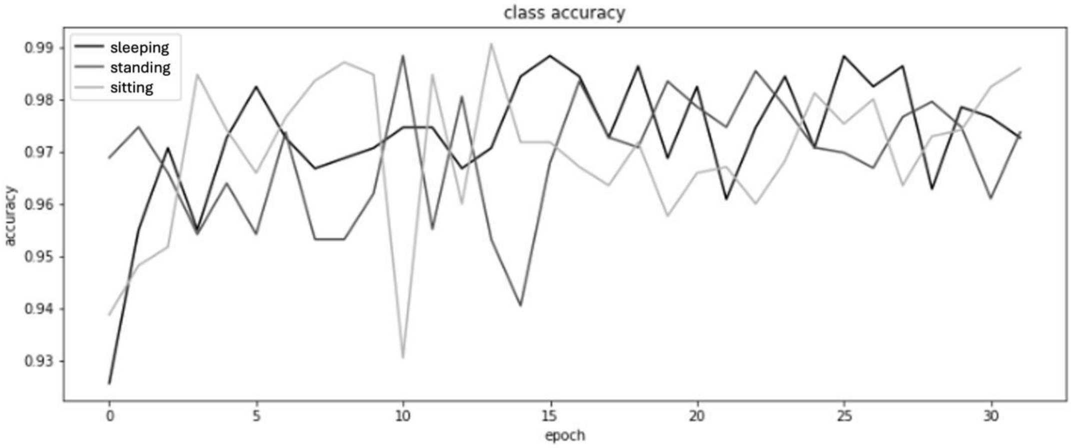


Figure 5. The Accuracy Graph For Each Category.

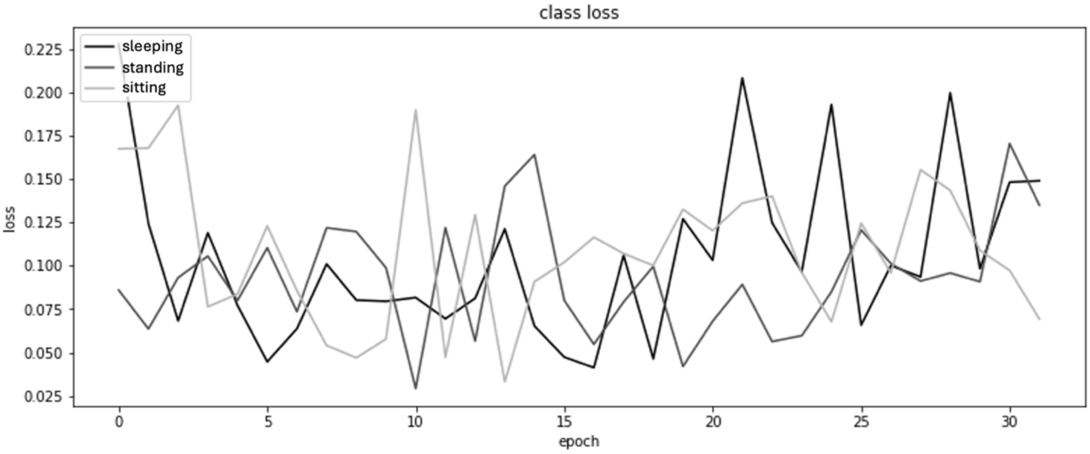


Figure 6. The Loss Graph For Each Category.

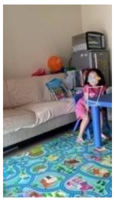
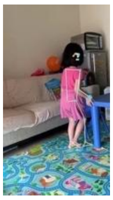
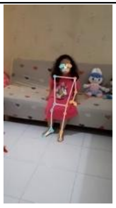
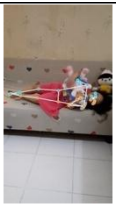
Table 4: Detailed Scenarios of Dataset 2 (Video Input).

Data Label	Time Frame (Seconds)	Activity
Scenario 1	0 – 3	Standing
	4 – 6	Sitting
	7 – 9	Standing
	10 - 13	Sleeping
Scenario 2	0 – 8	Standing
	9 – 11	Sitting
	12 – 19	Sleeping
	20 – 22	Sitting
	23 – 31	Standing
	32 – 39	Sitting
Scenario 3	0 – 2	Standing
	3 – 8	Standing
	9 – 10	Sitting
	11 – 14	Standing
	15 – 18	Sitting
	19 - 21	Sleeping
Scenario 4	0 – 8	Standing
	9 – 10	Sitting
	11 – 12	Sleeping
	13	Sitting
	14	Standing
	15 - 17	Sleeping

Table 5: The Accuracy Results of the Proposed System.

Dataset 1 (Image)	Dataset 2 (Video)							
	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	1 Frame	3 Frame	1 Frame	3 Frame	1 Frame	3 Frame	1 Frame	3 Frame
97.77%	78.57%	83.33%	70.00%	77.14%	71.81%	76.84%	78.88%	70.00%

Table 6: Samples of Testing Images Used in The Experiments.

Test Images	Activity		Test Images	Activity	
	Ground-truth	System's Classification		Ground-truth	System's Classification
	Sitting	Sitting		Standing	Standing
	Sitting	Standing		Sleeping	Standing

The performance evaluation results for the two test datasets are presented in Table 5, while samples of video frames used for testing and their classification results are presented in Table 6. As shown in Table 5, the system demonstrates strong performance in activity recognition of Dataset 1, achieving a high accuracy of 97.77%. Meanwhile, for Dataset 2 which consists of video inputs with varying frame sequences, the system achieves a maximum accuracy of 83.33% in Scenario 1, indicating effective recognition under specific temporal conditions. In Dataset 2, which contains video data, Scenario 4 yielded the lowest accuracy rate. This outcome is attributed to the abrupt positional changes of the subject, transitioning directly from a standing to a sleeping position. In contrast, the other scenarios involved more gradual transitions—from standing to sitting, and subsequently to sleeping. Additionally, activities in Scenario 4 were performed at a higher speed, as indicated by a shorter recording duration compared to the earlier scenarios. The results further demonstrate that the system achieves higher recognition accuracy when temporal consistency is preserved.

4 Conclusion

Based on the experimental results, the proposed system demonstrated effective performance in detecting children's activities, achieving a high accuracy of 97.77%. However, reliance on skeletal keypoints for human activity recognition may lead to misclassification of activities with similar skeletal configurations—such as distinguishing between learning and standing—due to the absence of contextual environmental cues. Additionally, camera positioning was found to significantly influence detection performance, underscoring the importance of incorporating datasets with diverse viewing angles to enhance model generalization. To address the instability associated with individual frame predictions, aggregating outputs over the most recent three frames proved beneficial in improving both detection accuracy and robustness.

References

[1] Martorell, G. (2021) Experience Human Development. 13th ed. McGraw-Hill Education.

[2] Morgenstern, A. (2023) Children's Multimodal Language Development From an Interactional, Usage-Based, and Cognitive Perspective. *Wiley Interdisciplinary Reviews: Cognitive Science*. 14 (2), e1631.

[3] Alduais, A., Al-Qaderi, I., Alfadda, N., and Alfadda, H. (2022) Pragmatics: Mapping Evidence on Enhancing Children's Use of Linguistic and Non-Linguistic Capacities for Interactive Communication. *Children*. 9 (9), 1318.

[4] Torring, M.F., Logacjov, A., Braendvik, S.M., Ustad, A., Roeleveld, K., and Bardal, E.M. (2024) Activity Recognition in Children with CP: Development and Validation of a Human Activity Recognition Model. *Gait & Posture*. 113 (1), 222–233.

[5] Patel, J., Shah, S., and Teraiya, V. (2024) Enhancing Child Physical Exercise with Pose Estimation Using Visual Cartoon Avatar. in: 2024 Int. Conf. on Sus. Comm. Net. and Comms.

[6] Ali, M.M. and Mohamed, S.I. (2024) A Pose Estimation for Motion Tracking of Infants Celebral Palsy. *Multimedia Tools Application.* 84, 8261–8286.

[7] Jasrotia, A., and Rajput A. (2025) Real-Time Pose Estimation: Reconnoitring Freebie Artificial Intelligence Applications in Physical Education and Sports. *Int. Journal of Physical Education, Sports, and Health.* 12(4), 32–35.

[8] Said, H., Mahar, K. Sorour, S.E., Elsheshai, A., Shaaban, R., Hesham, M., Khadr, M., Mehanna, Y.A., Basha, A., and Maghraby, F.A. (2024) IMITASD: Imitation Assessment Model for Children with Autism Based on Human Pose Estimation. *Mathematics.* 12 (21), 3438

[9] Anusha, D., Bhavani, K.P.S., Sai., N.V., Firdose, S.K., Chaitanya, Y.S. (2024) Children ADHD Disease Detection using Pose Estimation Technique. *Int. Journal of Engineering Science and Advanced Technology*. 24 (5), 279–291.

[10] Anderson, J.T., Stenum, J., Roemmich, R.T., and Wilson, R.B. (2025) Validation of Markerless Video-Based Gait Analysis Using Pose Estimation in Toddlers With and Without Neurodevelopmental Disorders. *Front. Digit. Health,* 7:1542012.

[11] Maret, Y., Oberson, D., and Gavrilova, M. (2018) Real-time Embedded System for Gesture Recognition. in: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 30–34.

[12] Benhaili, Z., Balouki, Y., and Moumoun, L. (2021) A Hybrid Deep Neural Network for Human Activity Recognition based on IoT Sensors. *International Journal of Advanced Computer Science and Applications*. 12 (11),.

[13] Janardhanan, J. and Umamaheswari, S. (2022) Vision based Human Activity Recognition using Deep Neural Network Framework. *International Journal of Advanced Computer Science and Applications*. 13 (6),.

[14] Gao, Y., Lu, J., Li, S., Ma, N., Du, S., Li, Y., et al. (2023) Action Recognition and Benchmark Using Event Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45 (12), 14081–14097.

[15] Yang, X., Xiong, B., Huang, Y., and Xu, C. (2024) Cross-Modal Federated Human Activity Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[16] Pandey, M., Mishra, R., and Khare, A. (2024) Vision-based Human Activity Recognition Using Local Phase Quantization. *Journal of Artificial Intelligence and*

Technology. 4 (3), 208–215.

[17] Rehman, S.U., Yasin, A.U., Ul Haq, E., Ali, M., Kim, J., and Mehmood, A. (2024) Enhancing Human Activity Recognition through Integrated Multimodal Analysis: A Focus on RGB Imaging, Skeletal Tracking, and Pose Estimation. *Sensors*. 24 (14), 4646.

[18] Dang, L.M., Min, K., Wang, H., Piran, M.J., Lee, C.H., and Moon, H. (2020) Sensor-based and Vision-based Human Activity Recognition: A Comprehensive Survey. *Pattern Recognition*. 108 107561.

[19] Wu, Y., Ranasinghe, D.C., Sheng, Q.Z., Zeadally, S., and Yu, J. (2011) RFID Enabled Traceability Networks: A Survey. *Distributed and Parallel Databases*. 29 397–443.

[20] Ko, C.-H. (2017) Accessibility of Radio Frequency Identification Technology in Facilities Maintenance. *Journal of Engineering, Project & Production Management*. 7 (1),.

[21] Huang, X., Xue, Y., Ren, S., and Wang, F. (2023) Sensor-Based Wearable Systems for Monitoring Human Motion and Posture: A Review. *Sensors*. 23 (22), 9047.

[22] Khan, D., Alonazi, M., Abdelhaq, M., Al Mudawi, N., Algarni, A., Jalal, A., et al. (2024) Robust human locomotion and localization activity recognition over multisensory. *Frontiers in Physiology*. 15 1344887.

[23] Sazonov, E.S., Fulk, G., Hill, J., Schutz, Y., and Browning, R. (2010) Monitoring of Posture Allocations and Activities by A Shoe-Based Wearable Sensor. *IEEE Transactions on Biomedical Engineering*. 58 (4), 983–990.

[24] Tran, T., Ruppert, T., Eigner, G., and Abonyi, J. (2023) Assessing Human Worker Performance by Pattern Mining of Kinect Skeleton Data. *Journal of Manufacturing Systems*. 70 538–556.

[25] Brambilla, C., Marani, R., Romeo, L., Nicora, M.L., Storm, F.A., Reni, G., et al. (2023) Azure Kinect Performance Evaluation for Human Motion and Upper Limb Biomechanical Analysis. *Heliyon*. 9 (11),.

[26] Bernal, F., Feipel, V., and Plaza, M. (2024) Kinect-Based Gait Analysis System Design and Concurrent Validity in Persons with Anterolateral Shoulder Pain Syndrome, Results from a Pilot Study. *Sensors*. 24 (19), 6351.

[27] Rouali, M.L., Boulahia, S.Y., and Amamra, A. (2023) Structure and Sequencing Preserving Representations for Skeleton-based Action Recognition Relying on Attention Mechanisms. *Journal of Signal Processing Systems*. 95 (8), 1003–1019.

[28] Alshurafa, N., Xu, W., Liu, J.J., Huang, M.-C., Mortazavi, B., Roberts, C.K., et al. (2013) Designing A Robust Activity Recognition Framework for Health and Exergaming Using Wearable Sensors. *IEEE Journal of Biomedical and Health Informatics*. 18 (5), 1636–1646.

[29] Gaglio, S., Re, G. Lo, and Morana, M. (2014) Human Activity Recognition Process Using 3-D Posture Data. *IEEE Transactions on Human-Machine Systems*. 45 (5), 586–597.

[30] Komang, M.G.A., Surya, M.N., and Ratna, A.N. (2019) Human Activity Recognition Using Skeleton Data and Support Vector Machine. in: J Phys Conf Ser, p. 12044.

[31] Atalaa, B.A., Ziedan, I., Alenany, A., and Helmi, A. (2021) Feature Engineering for Human Activity Recognition. *International Journal of Advanced Computer Science and Applications*. 12 (2).

[32] Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. (2022) Deep Learning for Sensor-based Human Activity Recognition. *ACM Computing Surveys*. 54 (4), 1–40.

[33] Albukhary, N. and Mustafah, Y.M. (2017) Real-Time Human Activity Recognition. in: IOP Conf Ser Mater Sci Eng, p. 12017.

[34] Dubois, A. and Charpillet, F. (2017) Measuring Frailty and Detecting Falls for Elderly Home Care Using Depth Camera. *Journal of Ambient Intelligence and Smart Environments*. 9 (4), 469–481.

[35] Kim, J.-W., Choi, J.-Y., Ha, E.-J., and Choi, J.-H. (2023) Human Pose Estimation Using Mediapipe Pose and Optimization Method Based on A Humanoid Model. *Applied Sciences*. 13 (4), 2700.

[36] Lina, L., Marunduh, A.A., Wasino, W., and Ajienegoro, D. (2022) Emotion Identification of Video Conference Users using Convolutional Neural Network. *Journal of Information Technology and Computer Science (in Indonesian)*. 9 (5), 1047–1054.

Lina Lina is a Professor at the Faculty of Information Technology, Tarumanagara University, Indonesia. Her research interests include computer vision, image recognition, and intelligent systems. She is a member of the Institute of Electrical and Electronic Engineers (IEEE) and the Institute of Electronics, Information and Communication Engineers (IEICE).

Arlends Chris is an Assistant Professor at Medical Faculty, Tarumanagara University, Indonesia. His main teaching and research interests include Human Anatomy, Histology and Physiology, Mental Health and Primary Care Medicine. He has published several research articles in international journals of medicine.

Ranny Ranny is a faculty member at the Department of Data Science, Universitas Bunda Mulia, Indonesia. Her academic and research interests are focused on Machine Learning, Artificial Intelligence, and Sound Processing. She has been involved in various research projects and has published several papers in the fields of audio signal processing and machine learning.