Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development of a Lightweight Detachable Balloon Robot with Inertial Movement Recognition for Older-Adult Assistance

Weihao Wang

wang-weihao1@ed.tmu.ac.jp

Tokyo Metropolitan University

Chyan Zheng Siow

Naoyuki Kubota

Tokyo Metropolitan University

Research Article

Keywords:

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11127138/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Development of a Lightweight Detachable Balloon Robot with Inertial Movement Recognition for Older-Adult Assistance

Weihao Wang[1*], Chyan Zheng Siow[1], Naoyuki Kubota[1], Azhar Aulia Saputra[1], Qingwei Song[1]

[1*]Tokyo Metropolitan University, Hachioji, 1910065, Tokyo, Japan.

*Corresponding author(s). E-mail(s): wang-weihao1@ed.tmu.ac.jp;

Abstract

Service robots for older adults must balance physical manageability with sensing and mobility. We present a detachable balloon robot that separates a lightweight inflatable companion unit from a Mecanum-wheel mobile base. The approximately 2 kg upper unit includes an embedded inertial sensor for movement recognition when the head-mounted camera is occluded. Seven participants performed five prescribed movements, producing 35 recordings. Six raw inertial and orientation channels sampled at approximately 10 Hz were transformed into 14 sequence features and segmented into 5 s windows with 50% overlap. Four pipelines—radial basis function support vector machine (RBF-SVM), convolutional neural network (CNN), long short-term memory network (LSTM), and CNN-LSTM—were evaluated using nested subject-independent leave-one-subject-out (LOSO) validation. RBF-SVM achieved the highest participant-wise window-level accuracy ($\mathbf{0.842 \pm 0.128}$) and macro-averaged F1 score ($\mathbf{0.840 \pm 0.125}$). CNN-LSTM was the strongest deep model ($\mathbf{0.742 \pm 0.140}$ accuracy; $\mathbf{0.717 \pm 0.135}$ macro-averaged F1 score). These results favor compact statistical features for the present short, pre-segmented task but do not establish superiority for longer or continuous activities. Because CNN-LSTM preserves temporal order, it remains a motivated candidate for future evaluation on longer movement sequences and action transitions. The study contributes a modular service-robot architecture and a rigorous cross-subject comparison of conventional and deep temporal recognition pipelines.

Keywords: service robotics, companion robot, inertial sensing, human–robot interaction, support vector machine, CNN–LSTM

1 Introduction

Service robots for older adults support mobility, coaching, monitoring, and social interaction. Studies in intelligent service robotics show that usability depends not only on control performance but also on clear multimodal interfaces, appropriate coaching, and preservation of user agency [1–3]. Companion robots such as PARO and LOVOT further illustrate the value of embodied interaction in care and home settings [4–6]. However, robots designed to be held or moved by users face a practical trade-off: locomotion hardware increases mass, while close physical contact can occlude camera-based sensing. We therefore developed a detachable balloon robot in which a lightweight inflatable companion unit separates from a Mecanum-wheel mobile base. The upper unit weighs approximately 2 kg and combines a camera with embedded inertial sensing, allowing movement recognition when camera visibility is reduced.

This study evaluates the movement-recognition module under subject-independent conditions. Seven participants performed five prescribed movements, and four pipelines—RBF-SVM, CNN, LSTM, and CNN-LSTM—were compared using nested LOSO validation. The study makes three contributions:

1. A modular service-robot architecture that separates the handheld companion unit from the heavier mobile base.
2. An object-embedded inertial sensing and preprocessing pipeline for camera-occluded physical interaction.
3. A cross-subject comparison of conventional and deep temporal recognition pipelines, with explicit limits on what the short-action benchmark can establish about future temporal modeling.

Accordingly, movement recognition is treated as one component of the service-robot system rather than as an isolated machine-learning benchmark.

2 Related Work and Design Motivation

2.1 Service and companion robots for older adults

Robot-Era showed that multimodal interfaces can improve the usability of multi-robot services for older adults [1], while Dr.VCoach integrated deep learning with robot-based exercise coaching [2]. These studies indicate that assistive robotics depends on the joint design of sensing, interaction, and physical accessibility. Companion robots provide a complementary form of support: PARO has been studied in elder-care settings [4, 5], and LOVOT illustrates how continuous motion and character behavior can support long-term home interaction [6]. The present system differs by physically separating the handheld companion unit from the locomotion hardware.

2.2 Human-centered interaction and user agency

Human-centered service robotics should preserve users' ability to choose and modify an interaction. Shared-control research similarly emphasizes maintaining user freedom unless intervention is necessary [3]. HRI research also cautions against reductive representations of older adults and supports more participatory approaches to robot design

[7, 8]. The concept of mattering—the sense that one's actions are significant to others—provides an additional design motivation, although it is not evaluated in the present experiment [9]. In the present design, users can hold and move the upper unit independently while the mobile base provides locomotion and charging. Programmed prompts should remain transparent and optional because autonomy, privacy, and reciprocity are established concerns in human–robot interaction (HRI) [10, 11].

2.3 IMU-based movement recognition

Inertial measurement units (IMUs) provide compact, low-power motion sensing for activity and rehabilitation analysis. Wearable-IMU studies have explored lightweight CNNs for gait recognition [12] and hybrid CNN-LSTM models for estimating energy expenditure across walking conditions [13]. Long short-term memory networks were introduced to model dependencies in sequential data [14]. These studies motivate temporal deep models for inertial time series, while the present work retains RBF-SVM as a conventional comparison and evaluates all pipelines under the same subject-independent outer folds.

2.4 Relation to our prior balloon robot study

Our earlier balloon-robot study compared seven recognition methods on a different dataset and found the highest score with LSTM [15]. The present study uses a five-movement benchmark, a different feature representation, and a stricter subject-independent nested LOSO protocol. RBF-SVM, CNN, LSTM, and CNN-LSTM are evaluated on the same held-out participants. Because the datasets and validation procedures differ, the earlier results provide design context rather than a direct baseline.

3 Balloon Robot System Design

3.1 System architecture and design requirements

The system was designed around four requirements: low user-handled mass, separation of handheld and locomotion hardware, sensing during close physical contact, and modular deployment. As shown in Fig. 1, the inflatable upper companion unit detaches from the mobile base. The upper unit weighs approximately 2 kg and has a prototype-scale cost of approximately USD 640. The base carries the heavier locomotion and charging hardware and is intended to support multiple upper units. This arrangement reduces the mass handled during direct interaction; total system cost has not yet been evaluated.

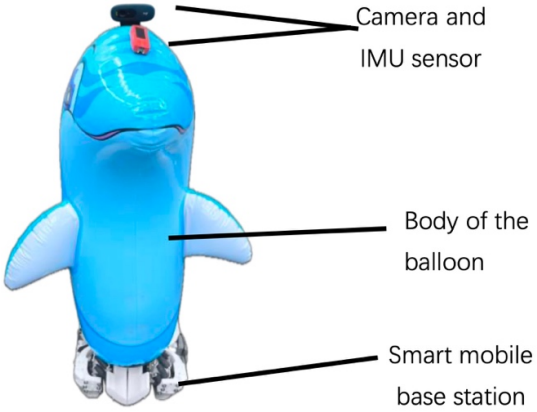


Fig. 1 Prototype of the detachable balloon robot system

3.2 Upper companion unit

The upper companion unit has an inflatable, self-righting body with a compact camera and M5StickC Plus sensing module mounted near the top (Fig. 2). Ballast in the lower body compensates for the elevated sensor mass and preserves self-righting behavior. The camera supports face or gesture sensing when the user remains visible. During close holding, however, the camera may be occluded or directed away from the user; the embedded inertial sensor is therefore the primary source for the present movement-recognition experiment.



Fig. 2 Upper inflatable companion unit

3.3 Smart mobile base

The smart mobile base (Fig. 3) provides locomotion, posture-related actuation, and charging while keeping these heavier functions separate from the handheld unit. Mecanum wheels enable holonomic planar motion in confined indoor spaces [16]. Wheel motion can also act on the lower shell of the upper unit to generate simple expressive movements such as turning or swaying. The base contains a high-capacity battery and a charging interface. The proposed shared-base deployment could reduce duplicated hardware, but this operational model has not yet been evaluated.

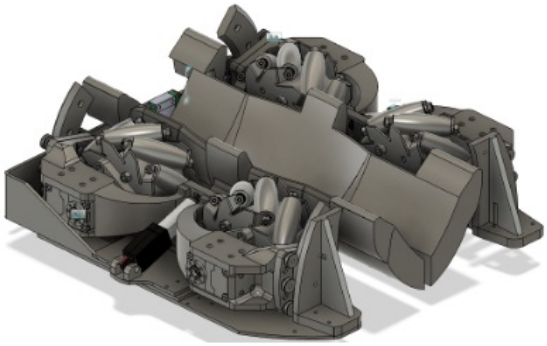


Fig. 3 Smart mobile base with Mecanum wheels

3.4 Interaction and sensing workflow

The robot combines visual and inertial sensing. When the user is visible, the camera can support face or gesture interaction; when the upper unit is held, the embedded inertial sensor continues to measure robot motion. The host computer receives User Datagram Protocol (UDP) packets, preprocesses each recording, constructs sliding windows, and applies the selected recognition pipeline. The present study evaluates this sensing and classification path offline. Closed-loop response latency and end-to-end control performance remain future work.

4 Experimental Methods

4.1 Participants and movement protocol

The subject-independent benchmark included seven participants (S01–S07). Each participant completed one recording of each of five prescribed movements while holding the upper companion unit, yielding 35 recordings. The movement set was developed in consultation with geriatric rehabilitation therapists and was treated as five recognition classes (Table 1). The experiment assesses technical recognition feasibility under controlled conditions and is not a clinical rehabilitation trial.

Table 1 Operational definitions of the five movement classes

Class	Movement description
A	Hold the balloon robot with both hands, extend the arms, and raise it overhead.
B	Hold the robot with extended arms and rotate the upper body to the left.
C	Hold the robot with extended arms and rotate the upper body to the right.
D	Hold the robot with extended arms and bend forward until the lower part of the robot approaches or contacts the floor, according to the study protocol.
E	Raise the robot overhead and sway the body and robot from side to side.

4.2 Sensor acquisition and preprocessing

The M5StickC Plus provided six raw channels—`accX`, `accY`, `accZ`, pitch, roll, and yaw—at approximately 10 Hz. Preprocessing was performed independently for each recording. Pitch, roll, and yaw were unwrapped to remove $\pm 180^\circ$ discontinuities and smoothed with a Savitzky–Golay filter (5-sample window, second-order polynomial). Angular rates were computed using the nominal 0.1 s sampling interval. Additional channels comprised acceleration magnitude, sine/cosine representations of smoothed pitch and roll, and pitch/roll/yaw displacement relative to the first sample of each window, yielding 14 sequence features in total. Each recording was segmented into 50-sample windows (approximately 5 s) with a 25-sample step (50% overlap). Scaling parameters were fitted only on training participants within each inner or outer partition and then applied to the corresponding validation or test data.

4.3 Classification models and training

Four recognition pipelines were evaluated using the same outer LOSO test participants:

1. **RBF-SVM:** each 50×14 window was summarized by eight statistics (mean, standard deviation, minimum, maximum, median, root-mean-square, range, and interquartile range) for each channel, yielding 112 handcrafted features. Inner LOSO selected C from $\{0.1, 1, 10, 100\}$ and γ from $\{\text{scale}, 0.001, 0.01, 0.1, 1\}$.
2. **CNN:** the full 50×14 sequence was processed by two one-dimensional convolutional blocks (32 and 64 filters) with batch normalization, rectified linear unit (ReLU) activation, dropout, adaptive average pooling, and a fully connected classifier.
3. **LSTM:** a unidirectional LSTM with 64 hidden units was followed by layer normalization and fully connected classification.
4. **CNN-LSTM:** the 32/64-channel convolutional feature extractor was followed by a unidirectional LSTM with 32 hidden units.

The comparison is therefore pipeline-level rather than classifier-only: RBF-SVM uses explicit statistical summaries, whereas the neural models receive ordered sequence features. Fig. 4 summarizes the preprocessing and nested evaluation workflow.

The deep models were implemented in PyTorch and trained with class-weighted cross-entropy loss, label smoothing of 0.05, AdamW optimization (learning rate 3×10^{-4}; weight decay 10^{-3}), batch size 32, and gradient clipping at 1.0. Class weights were computed only from the current training labels. Each inner LOSO fold used

early stopping on validation macro-averaged F1 score (macro-F1; maximum 80 epochs; patience 12). The median best epoch across the six inner validation participants determined the fixed duration for final outer-fold retraining. Each final model was trained from scratch on all six non-test participants. A base random seed of 42 was used; optimization variability across seeds was not evaluated.

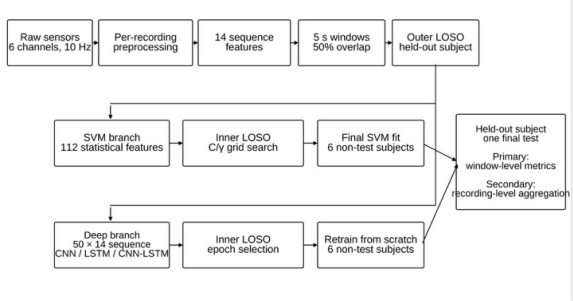


Fig. 4 Preprocessing and nested subject-independent LOSO evaluation pipeline

4.4 Evaluation protocol

The primary evaluation used nested subject-independent leave-one-subject-out (LOSO) validation:

1. **Outer split:** one complete participant was held out for final testing; the remaining six formed the candidate training set.
2. **Inner selection:** inner LOSO within the six training participants selected C and γ for RBF-SVM or the final training duration for each deep model.
3. **Final evaluation:** after selection, the model was refitted on all six non-test participants, with scaling fitted only on those participants, and the held-out participant was evaluated once.

Primary results are the mean ± sample standard deviation of participant-wise window-level metrics across the seven outer folds. Because overlapping windows from the same recording are correlated, windows were not treated as independent replicates for inference. Recording-level performance was computed separately by averaging class probabilities over all windows from each pre-segmented movement recording. Exploratory model comparison used participant-level outer-fold macro-F1 values in a Friedman test followed by exact paired Wilcoxon signed-rank tests with Holm correction.

Table 2 Nested LOSO performance across the seven held-out participants (mean ± standard deviation)

Model	Window Acc.	Window Macro-F1	Recording Acc.	Recording Macro-F1
RBF-SVM	**0.842 ± 0.128**	**0.840 ± 0.125**	**0.943 ± 0.098**	**0.924 ± 0.130**
CNN	0.684 ± 0.146	0.615 ± 0.155	0.714 ± 0.157	0.633 ± 0.177
LSTM	0.644 ± 0.087	0.605 ± 0.097	0.743 ± 0.151	0.671 ± 0.164
CNN-LSTM	**0.742 ± 0.140**	**0.717 ± 0.135**	**0.800 ± 0.200**	**0.748 ± 0.234**

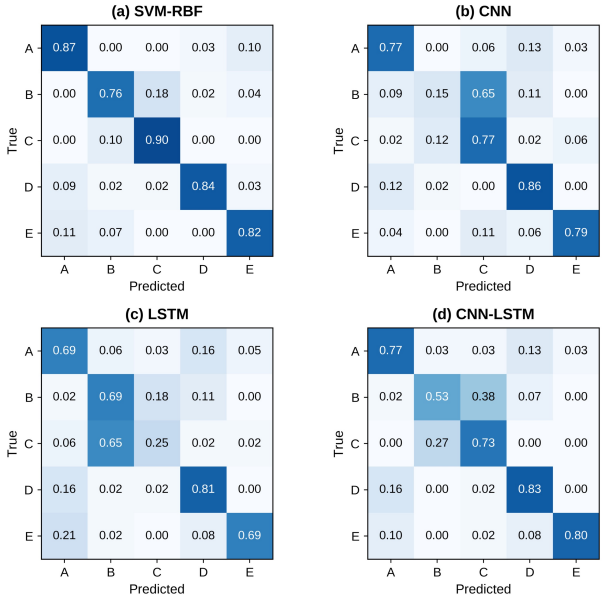


Fig. 5 Pooled row-normalized window-level confusion matrices across the seven outer LOSO folds: (a) RBF-SVM, (b) CNN, (c) LSTM, and (d) CNN-LSTM

5 Results

Table 2 summarizes the nested LOSO results. RBF-SVM ranked first at the window level (accuracy 0.842 ± 0.128; macro-F1 0.840 ± 0.125). CNN-LSTM was the strongest deep model (0.742 ± 0.140; 0.717 ± 0.135), exceeding CNN and LSTM. Relative to CNN-LSTM, RBF-SVM improved accuracy by 10.0 percentage points and macro-F1 by 12.3 percentage points. Recording-level aggregation preserved the same ordering (RBF-SVM: 0.943 ± 0.098 accuracy, 0.924 ± 0.130 macro-F1; CNN-LSTM: 0.800 ± 0.200, 0.748 ± 0.234). Because each participant contributed only five recordings, recording-level results are treated as secondary.

Fig. 5 shows the pooled row-normalized confusion matrices. RBF-SVM achieved class recalls between 0.76 and 0.90, indicating comparatively balanced recognition. CNN's main error was movement B: recall fell to 0.15 and 65% of movement B were

predicted as movement C. LSTM shifted the movement B/C bias in the opposite direction (movement B recall 0.69; movement C recall 0.25), whereas CNN-LSTM produced a more balanced deep-model result (movement B 0.53; movement C 0.73). A Friedman test on participant-level macro-F1 detected an overall model effect ($\chi^2(3) = 17.91$, $p < 0.001$). Exact paired Wilcoxon tests comparing RBF-SVM with each deep model, followed by Holm correction, yielded adjusted $p = 0.0469$ for each comparison. Given the sample of seven participants, these inferential results are exploratory.

6 Discussion

6.1 Implications for intelligent service-robot design

The system separates functions that require close physical interaction from those that add mass. The upper companion unit contains the interaction surface and inertial sensing, while locomotion and charging remain in the mobile base. This reduces the mass handled by the user and avoids requiring an additional body-worn sensor. A shared mobile base could also reduce duplicated locomotion hardware, although this deployment concept requires operational validation.

6.2 Interpretation of the movement-recognition results

For the present short, pre-segmented movements, RBF-SVM with handcrafted statistical features outperformed all three deep models by 12.3 percentage points in macro-F1 relative to CNN-LSTM. This advantage should be interpreted at the pipeline level: RBF-SVM receives 112 explicit statistical descriptors, whereas the neural models receive the ordered sequence. The experiment therefore does not isolate whether the observed advantage arises from the classifier, the statistical representation, or both. Given the small benchmark (seven participants, 35 recordings), compact descriptors of movement amplitude and variability appear well matched to the current task.

CNN-LSTM nevertheless remains relevant for a different reason: it preserves temporal order. LSTMs update an internal state across ordered samples [14], and hybrid CNN-LSTM models have been applied to IMU time series in related movement-analysis tasks [13]. These properties provide a methodological basis for evaluating CNN-LSTM in future longer-horizon or continuous Balloon Robot interactions.

The current experiment does not test that hypothesis. Each recording contains one labeled movement and is segmented into 5 s windows; long-duration movements, continuous multi-movement streams, transitions, and dependencies between successive movements are absent. The present evidence therefore supports RBF-SVM for the bounded recognition task, while CNN-LSTM should be regarded as a motivated sequence-modeling candidate for future tasks in which temporal order or wider context determines class identity. Direct evaluation will require longer recordings, transition-aware labels, and online testing.

6.3 Human-centered interaction and deployment

The balloon robot is intended as a companion service platform rather than a diagnostic device. User participation should remain optional and transparent, consistent

with human-centered service-robot design [1, 3, 7, 8]. Camera sensing should be used only when needed; for movement tasks that do not require images, embedded inertial sensing may reduce reliance on continuous visual monitoring. Institutional deployment also requires evaluation of charging availability, cleaning, maintenance, pairing, staff workload, and failure recovery. These operational factors are outside the scope of the present recognition benchmark.

6.4 Limitations and required validation

The study has five main limitations:

1. **Sample size:** only seven participants and 35 recordings were available, limiting statistical power and precision across participants.
2. **Task scope:** the five movements were prescribed and pre-segmented; transitions, partial movements, spontaneous handling, and long-duration activities were not represented.
3. **Correlated windows:** 50% overlapping windows from the same recording are not independent, so participant-level outer folds remain the unit of inference.
4. **Representation mismatch:** RBF-SVM used 112 statistical features, whereas the deep models used ordered sequences. The study therefore does not isolate whether the SVM advantage arises from the classifier, the representation, or both. Representation-controlled baselines should be added in future work.
5. **System validation:** neural models were evaluated with one base seed, and the mobile base, charging workflow, closed-loop latency, and end-to-end interaction were not quantitatively tested.

Accordingly, the results should not be generalized to long-term deployment, continuous activity understanding, rehabilitation effectiveness, or populations beyond the tested participants.

7 Conclusion

This study presents a lightweight detachable balloon robot and a subject-independent inertial movement-recognition benchmark. The approximately 2 kg companion unit separates direct interaction and sensing from the heavier Mecanum-wheel base. Under nested LOSO evaluation, RBF-SVM with handcrafted statistical features achieved the highest window-level performance (accuracy 0.842 ± 0.128; macro-F1 0.840 ± 0.125), while CNN-LSTM was the strongest deep model (0.742 ± 0.140; 0.717 ± 0.135). Thus, RBF-SVM is the stronger pipeline for the current short, pre-segmented task. CNN-LSTM remains a justified candidate for future sequence-oriented studies because it preserves temporal order, but any advantage for longer movements or action transitions must be demonstrated experimentally.

Acknowledgements. This work was partially supported by the Japan Science and Technology Agency (JST), Moonshot R&D, grant JPMJMS2034, and TMU local 5G research support.

References

[1] Di Nuovo, A., Broz, F., Wang, N., *et al.*: The multi-modal interface of Robot-Era multi-robot services tailored for the elderly. Intelligent Service Robotics **11**(1), 109–126 (2018) https://doi.org/10.1007/s11370-017-0237-6

[2] Cauli, N., Reforgiato Recupero, D.: Dr.VCoach: Employment of advanced deep learning and human–robot interaction for virtual coaching. Intelligent Service Robotics **18**(3), 613–629 (2025) https://doi.org/10.1007/s11370-025-00603-3

[3] Falqueto, P., Antonucci, A., Palopoli, L., Fontanelli, D.: Humanising robot-assisted navigation. Intelligent Service Robotics **17**(2), 155–165 (2024) https://doi.org/10.1007/s11370-023-00495-1

[4] Shibata, T., Coughlin, J.F.: Trends of robot therapy with neurological therapeutic seal robot, PARO. Journal of Robotics and Mechatronics **26**(4), 418–425 (2014) https://doi.org/10.20965/jrm.2014.p0418

[5] Hori, Y., Kato, K., Kobayashi, M., *et al.*: Use of robotic pet in a distributed layout elderly housing with services: A case study on elderly people with cognitive impairment. Journal of Robotics and Mechatronics **33**(4), 784–803 (2021) https://doi.org/10.20965/jrm.2021.p0784

[6] Yoshida, N., Yonemura, S., Emoto, M., *et al.*: Production of character animation in a home robot: A case study of LOVOT. International Journal of Social Robotics **14**(1), 39–54 (2022) https://doi.org/10.1007/s12369-021-00746-0

[7] Burema, D.: A critical analysis of the representations of older adults in the field of human–robot interaction. AI & SOCIETY **37**(2), 455–465 (2022) https://doi.org/10.1007/s00146-021-01205-0

[8] Weiss, A., Spiel, K.: Robots beyond science fiction: Mutual learning in human–robot interaction on the way to participatory approaches. AI & SOCIETY **37**(2), 501–515 (2022) https://doi.org/10.1007/s00146-021-01209-w

[9] Flett, G.L., Heisel, M.J.: Aging and feeling valued versus expendable during the COVID-19 pandemic and beyond: A review and commentary of why mattering is fundamental to the health and well-being of older adults. International Journal of Mental Health and Addiction **19**(6), 2443–2469 (2021) https://doi.org/10.1007/s11469-020-00339-4

[10] Fronemann, N., Pollmann, K., Loh, W.: Should my robot know what's best for me? human–robot interaction between user experience and ethical design. AI & SOCIETY **37**(2), 517–533 (2022) https://doi.org/10.1007/s00146-021-01210-3

[11] Van Wynsberghe, A.: Social robots and the risks to reciprocity. AI & SOCIETY **37**(2), 479–485 (2022) https://doi.org/10.1007/s00146-021-01207-y

[12] Huang, H., Zhou, P., Li, Y., Sun, F.: A lightweight attention-based CNN model for efficient gait recognition with wearable IMU sensors. Sensors **21**(8), 2866 (2021) https://doi.org/10.3390/s21082866

[13] Lee, C.J., Lee, J.K.: IMU-based energy expenditure estimation for various walking conditions using a hybrid CNN–LSTM model. Sensors **24**(2), 414 (2024) https://doi.org/10.3390/s24020414

[14] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997) https://doi.org/10.1162/neco.1997.9.8.1735

[15] Wang, W., Siow, C.Z., Kubota, N., Saputra, A.A., Song, Q., Obo, T.: Balloon robot: Movement recognition, motion of robot and design of robot. In: Intelligent Robotics and Applications. Lecture Notes in Computer Science, vol. 15210, pp. 42–55. Springer, Singapore (2025). https://doi.org/10.1007/978-981-96-0786-0_4

[16] Wang, Y., Chang, D.: Motion performance analysis and layout selection for motion system with four mecanum wheels. Journal of Mechanical Engineering **45**(5), 307–310 (2009) https://doi.org/10.3901/JME.2009.05.307