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.

Arogya: Interpretable Joint-Angle Templates for Camera-Based Assessment of Home Rehabilitation Exercises

Shilpa Katikar
Vishwakarma Institute of Technology
Kavya Bhand
kavya.bhand24@vit.edu

Vishwakarma Institute of Technology
Kabir Khanuja
Vishwakarma Institute of Technology
Pranav Rana
Vishwakarma Institute of Technology

Research Article

Keywords: human pose estimation, rehabilitation, exercise assessment, MediaPipe, joint angles, interpretable models, telerehabilitation

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11822923/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Home exercise programmes are central to rehabilitation, but patients mostly perform them unsupervised, and clinicians learn little about how the exercises were actually done. We present Arogya, a camera-only home rehabilitation system in which each exercise is described by a template of named joint-angle ranges that a physiotherapist can read and adjust for an individual patient. A repetition is scored by how far its measured angles fall outside those ranges, so every warning refers to a specific joint and a specific size of deviation. We evaluated the measurement and decision stages on the public REHAB24-6 dataset (10 participants, six exercises, 1,057 labelled repetitions), using MediaPipe Pose on ordinary RGB video and synchronised optical motion capture as the reference. Repetition-level trunk, hip and knee angles agreed with motion capture with mean absolute errors of 5.1°, 8.9° and 10.8° (r = 0.88–0.94). Elbow angles were less reliable (20.9°), and a camera at half-profile gave the best agreement for most angles. In leave-one-subject-out evaluation, the templates detected incorrect repetitions with a mean balanced accuracy of 0.58 (AUC 0.62). The same templates reached 0.60 on motion-capture angles, and a random forest trained on the same features reached 0.63. Pose estimation and template evaluation together ran at 25 frames per second on four CPU cores. For this task a single camera comes close to motion capture; the main limitation is what joint-angle extremes can express. We discuss what this implies for per-patient, clinician-set criteria.

1. Introduction

Rehabilitation after injury or surgery depends on exercises that the patient repeats many times between clinic visits, and most of that repetition happens at home with nobody watching. Two things tend to go wrong there. Patients drift away from the prescribed movement, for instance by bending a knee further than a recovering ligament should allow or by twisting the trunk to make up for a stiff shoulder, and many stop exercising altogether once the novelty wears off [1], [2]. The clinician usually finds out at the next appointment, from the patient's own account. The need is large: an estimated 2.4 billion people worldwide live with a condition that would benefit from rehabilitation [3].

Real-time telerehabilitation over video calls gives outcomes comparable to in-person care for several musculoskeletal conditions [4], but it ties up the therapist for the length of every session. Wearable inertial sensors measure joint motion well, yet they must be bought, charged and strapped on correctly. A camera that the patient already owns is a cheaper sensor. Modern pose estimators such as OpenPose [5] and BlazePose [6], available on phones through MediaPipe [7], locate body landmarks in ordinary RGB video at interactive frame rates, and joint angles computed from those landmarks have been reported to agree reasonably with marker-based motion capture [8], [9].

Most published work on automatic exercise assessment trains a model, often a deep network, to map a skeleton sequence to a quality score or a correct/incorrect label [10]–[12]. These models can be accurate, but their output is hard for a physiotherapist to inspect or adjust. If a patient is allowed only 90° of knee flexion for the first six weeks after surgery, there is no obvious way to write that restriction into a trained network, and a low score does not say which joint was at fault.

Arogya, the system described here, takes the opposite position. Each exercise is described by a small template of named joint-angle ranges and safety limits that a clinician can read and edit for an individual patient. A repetition is judged by how far its measured angles fall outside those ranges, so every warning points to a specific joint and a specific number of degrees. The question we set out to answer is how much this transparency costs in accuracy, and how reliable the camera-derived angles are in the first place.

The contributions of this paper are:

1. the design of a camera-only home rehabilitation system in which exercise criteria are explicit joint-angle templates that a clinician can adjust per patient, with measurements returned to the clinician after each session;
2. a measurement of joint-angle agreement between MediaPipe Pose, applied to ordinary RGB video, and optical motion capture on the public REHAB24-6 dataset [13], broken down by camera view and lighting;
3. a leave-one-subject-out evaluation of template-based correctness detection on six exercises, compared with a random-forest classifier trained on the same features and with the same templates applied to motion-capture angles;
4. latency measurements showing that the full per-frame pipeline runs in real time on a commodity CPU.

All code is public so that the numbers can be reproduced (see Data Availability).

2. Related Work

Pose estimation. OpenPose introduced part affinity fields for real-time multi-person 2D keypoint detection [5]. BlazePose takes a lighter detector–tracker approach aimed at mobile devices and predicts 33 landmarks, including approximate 3D “world” coordinates in metres relative to the hips [6]; it is distributed through the MediaPipe framework [7]. Lafayette et al. compared MediaPipe and several RGB-D sensors with a marker-based system for shoulder, elbow, hip and knee movements and found the RGB method’s angular error to be within the range typically attributed to manual goniometry [8]. Stenum et al. showed that 2D video pose estimation can recover useful gait kinematics [9]. Multi-camera 3D pipelines such as RePose push accuracy further and add biomechanical modelling, at the cost of more hardware [14].

Exercise assessment. Public skeleton datasets with correctness labels, including UI-PRMD [15] and KIMORE [16], have driven most of the work in this area. Liao et al. proposed a deep framework that scores rehabilitation movements from skeleton sequences [10]. Mennella et al. built a home monitoring system for six resistance exercises around pose estimation and deep classifiers [11]. Sardari et al. review skeleton-based assessment and note that feedback generation and interpretability receive far less attention than classification accuracy [12]. REHAB24-6 adds synchronised RGB video and motion capture to correctness labels, which makes it possible to evaluate a camera pipeline against a physical reference rather than against another estimator [13].

Position of this work. Arogya does not propose a new pose estimator or a new classifier. Its focus is the interface between the measurement and the clinician: criteria that a physiotherapist can read and edit, evaluated on camera-derived angles, with an honest account of how well that works on real video.

3. System Design

3.1 Overview

Figure 1 shows the architecture. Pose estimation, angle computation, template evaluation and feedback all run on the patient's device, so video does not need to leave it. The server stores exercise templates and per-patient overrides, receives a summary of each session, and presents it to the treating clinician, who can edit the patient's templates from the dashboard. Communication is encrypted and access is role-based: a clinician sees only the patients assigned to them. Figure 2 shows two screens of the current mobile prototype.

3.2 Landmarks and joint angles

Each frame is passed to the MediaPipe Pose Landmarker [6], [7], which returns 33 landmarks with metric 3D world coordinates. A joint angle is the angle at the middle landmark of a triple. For landmarks $\varvec{a}$, $\varvec{b}$, $\varvec{c}$,

$$\theta\left(\varvec{a}, \varvec{b}, \varvec{c}\right) = \arccos \frac{\left(\varvec{a} - \varvec{b}\right) \cdot \left(\varvec{c} - \varvec{b}\right)}{\| \varvec{a} - \varvec{b} \| \, \| \varvec{c} - \varvec{b} \|}$$

We use the elbow (shoulder–elbow–wrist), shoulder (elbow–shoulder–hip), hip (shoulder–hip–knee) and knee (hip–knee–ankle) angles on both sides, the angle between the two thigh vectors (leg separation), and trunk inclination, the angle between the mid-hip-to-mid-shoulder vector and the vertical. Because templates should not depend on which side the patient happens to use, bilateral angles are also reduced to their per-frame minimum and maximum across the two sides.

3.3 Exercise templates

A template for exercise e is a set of features $f_1, \ldots, f_K$, each with a permitted interval $[l_k, u_k]$ in degrees. In Arogya, a feature summarises one angle over a repetition; we use the 10th and 90th percentiles of the angle within the repetition as robust estimates of its lowest and highest values. A clinician can narrow an interval or add a hard safety limit for one patient without changing the template for anyone else. For a repetition with feature values x_k, the exceedance of feature k is

$$\delta_k = \max(l_k - x_k,\, 0) + \max(x_k - u_k,\, 0),$$

and the repetition's deviation score is $D = \max_k \delta_k$. The repetition is flagged when $D > m$ for a tolerance margin m. The feature with the largest δ_k names the joint that the feedback should talk about, and δ_k itself says by how much it missed. Breaching a safety limit is handled separately: it triggers an immediate warning regardless of m.

3.4 Feedback and the clinician loop

During a repetition the prototype highlights the segment with the largest current exceedance on the skeleton overlay and shows a short cue such as “keep your back straighter”. An optional spoken cue helps when the patient is on the floor and cannot see the screen. At the end of a session the device uploads repetition counts, per-feature exceedances and scores. The clinician reviews them on the dashboard and can widen a range as the patient recovers or tighten a limit if a joint keeps overshooting. The next session is evaluated against the edited template.

4. Experimental Methodology

4.1 Dataset

We used REHAB24-6 [13], which is publicly available under a CC BY-NC 4.0 licence. Ten adults (six men, four women, aged 25–50) performed six exercises: arm abduction, an arm V-to-W movement, push-ups against a table, leg abduction, lunges and squats. A physiotherapist instructed each participant to perform some repetitions correctly and others with a typical mistake. The recordings come from two RGB cameras at 30 frames per second placed in adjacent corners of the room, so that a participant seen from the front by one camera is seen from the side by the other, and from 16 optical motion-capture cameras that provide a 26-joint skeleton used here as ground truth. Each of the 1,072 annotated repetitions carries a binary correctness label, the participant’s orientation to each camera, and a flag for lighting conditions. Fifteen repetitions with flagged motion-capture errors were excluded, leaving 1,057 for analysis (Table 1). Relative to the first camera, participants faced it from the front in 499 repetitions, at half-profile in 466 and from the side in 107; 308 repetitions were recorded with the room lights switched off.

4.2 Pose extraction

We ran the MediaPipe Pose Landmarker (full model, float16, video mode, one person) on every second frame within the annotated repetitions, that is, at 15 frames per second, which is the lower end of what a phone would deliver. This gave 64,326 frames per camera at 1920 × 1080 pixels. Both cameras were processed independently. No dataset-specific tuning or fine-tuning was performed.

4.3 Angle agreement

Templates operate on repetition-level features, so agreement was assessed at that level too. For every repetition, camera and angle, the 10th- and 90th-percentile values obtained from the MediaPipe world landmarks were compared with the same quantities computed from the motion-capture joints over the same frames. Left and right sides were pooled. We report the mean absolute error (MAE), median absolute error, mean signed error (bias, MediaPipe minus motion capture), the Pearson correlation r and the share of values within 10°, overall and broken down by the participant’s orientation to the camera (front, half-profile, side) and by lighting. The two skeletons define joints slightly differently; for example, MediaPipe’s hip landmarks lie closer to the skin surface than a motion-capture hip-joint centre. The error therefore combines estimator error with this definitional offset, which matters most for the hip and shoulder.

4.4 Correctness detection

For each exercise we selected the features a physiotherapist would plausibly check (Table 1) and evaluated detection of incorrect repetitions with leave-one-subject-out cross-validation. In each fold the template intervals were set to the 5th–95th percentile range of the correct repetitions of the nine training participants, and the margin m was chosen from 0° to 40° in steps of 2.5° to maximise balanced accuracy on those same training participants. The held-out participant's repetitions were then scored. This stands in for a clinician setting ranges, which we could not do for a public dataset, and it keeps the test participant entirely unseen.

Table 1
Exercises, repetitions used, and template features. Every feature enters the template twice, as its 10th and 90th percentile within the repetition. "Min" and "max" are the smaller and larger of the left and right angles in each frame.

Exercise	Repetitions (incorrect)	Template features
Arm abduction	178 (88)	max shoulder, min elbow, trunk inclination
Arm V-to-W	208 (114)	min and max shoulder, min and max elbow, trunk inclination
Push-up (table)	96 (50)	min and max elbow, min and max hip, min knee
Leg abduction	210 (90)	leg separation, min knee, min hip, trunk inclination
Lunge	174 (96)	min and max knee, min and max hip, trunk inclination
Squat	191 (61)	min and max knee, min and max hip, trunk inclination

We compared three settings: (i) templates applied to MediaPipe angles from the camera that saw the participant from the front or half-profile (the realistic home setup, where the patient faces the device), with each camera on its own also reported; (ii) a random forest (400 trees, balanced class weights) trained on the same features with the same folds, as a learned reference; and (iii) templates applied to motion-capture angles, which shows what the method achieves with a near-perfect pose. Repetitions with flagged motion-capture errors were excluded from all settings so that every method sees the same repetitions. We report balanced accuracy, sensitivity (incorrect repetitions caught), specificity, F1 score for the incorrect class, and the area under the ROC curve (AUC) of the deviation score.

4.5 Latency

Inference time was measured for the lite, full and heavy pose models on 300 consecutive frames of one recording after a 30-frame warm-up, in a single process. The input was 1920 × 1080 video. The angle and template stage was timed separately over 20,000 single-frame calls. All experiments ran on Kaggle cloud instances with four virtual CPUs (Intel Xeon, 2.20 GHz) and no GPU.

5. Results

5.1 Pose detection

MediaPipe returned a pose in 99.95% of processed frames from the first camera and 99.99% from the second. Only two repetitions had a pose in fewer than 90% of their frames on either camera. The dataset was recorded indoors with one participant at a time in clear view, so this is a best case. The rate tells us that missing detections were not a factor in the results below. It does not show how the estimator would cope with a cluttered living room.

5.2 Agreement with motion capture

Table 2 summarises agreement between repetition-level angle features from MediaPipe and from motion capture. Trunk inclination agreed best, with an MAE of 5.1°, no systematic bias, and 87% of values within 10°. The hip and knee followed, with MAEs of 8.9° and 10.8° and correlations of 0.94 and 0.92. The upper limb was clearly worse: the elbow MAE was 20.9° and only a third of elbow values fell within 10°. The shoulder correlated well (r = 0.91), but MediaPipe read it about 10° smaller than motion capture on average.

Table 2
Agreement of repetition-level angle features (10th and 90th percentiles) between MediaPipe and motion capture, both cameras and all views pooled. Bias is MediaPipe minus motion capture.

Angle	Pairs	MAE (°)	Median AE (°)	Bias (°)	r	Within 10°
Elbow	8,456	20.9	16.6	+ 6.4	0.81	32%
Shoulder	8,456	15.7	12.5	−10.5	0.91	40%
Hip	8,456	8.9	7.0	+ 6.7	0.94	69%
Knee	8,456	10.8	8.8	+ 5.8	0.92	56%
Leg separation	4,228	9.2	7.2	−5.0	0.86	67%
Trunk inclination	4,228	5.1	3.6	0.0	0.88	87%

Figure 3 plots MediaPipe against motion capture for the knee, elbow and hip with the participant facing the camera. Most of the knee and hip error is a consistent offset. When the joint is strongly flexed, MediaPipe reads it as straighter than motion capture does, so points lie above the diagonal at the low end. For the elbow the spread is far wider: a motion-capture value of 60° could come out of MediaPipe anywhere from about 70° to 140°. This is what one expects when the forearm points towards the camera and the elbow's depth has to be guessed from a single image.

Viewing angle mattered in different ways for different joints (Table 3, Fig. 4). The elbow and shoulder were worst from the side, where the far arm is hidden behind the body. Trunk inclination was worst from the front (MAE 8.2°, against 3.8° at half-profile), because leaning forward then moves the trunk almost directly towards the camera. Hip and knee errors were slightly lower from the side, where those joints flex in the

image plane. Half-profile gave the highest correlation for four of the six angles. Switching the lights off increased the elbow error by about 4.5° and had little effect elsewhere.

Table 3
MAE (°) of repetition-level angle features by the participant's orientation to the camera and by lighting.

Angle	Front	Half-profile	Side	Lights on	Lights off
Elbow	20.5	19.0	24.1	19.6	24.0
Shoulder	15.5	14.3	18.2	15.8	15.6
Hip	9.4	9.0	8.1	9.0	8.4
Knee	11.2	11.2	9.9	10.9	10.6
Leg separation	8.7	8.9	10.1	9.1	9.3
Trunk inclination	8.2	3.8	3.9	5.1	5.0

5.3 Detecting incorrect repetitions

Table 4 gives the leave-one-subject-out results. Templates applied to MediaPipe angles reached a mean balanced accuracy of 0.58 and a mean AUC of 0.62. Detection was best for push-ups (AUC 0.80) and leg abduction (0.70). For the arm V-to-W exercise it was close to chance, and for lunges it was below chance (AUC 0.41): the ranges learned from nine participants did not transfer to the tenth.

Applied to motion-capture angles, the same templates did only slightly better, with a mean balanced accuracy of 0.60 and an AUC of 0.65. The random forest trained on the same MediaPipe features averaged 0.63 and 0.69. It improved most on lunges and arm abduction, but it did worse than the templates on push-ups and the arm V-to-W exercise. On motion-capture angles the forest reached 0.62 balanced accuracy, a smaller margin over the templates. Using either camera on its own changed the template results by less than 0.01 in mean balanced accuracy (0.577 for the first camera, 0.581 for the second). The margins selected on the training folds were small (median 2.5° for most exercises). Most of the decision therefore came from the template intervals themselves rather than from the tolerance.

Table 4

Detection of incorrect repetitions, leave-one-subject-out. BA: balanced accuracy; Sens: sensitivity; Spec: specificity. MP: MediaPipe angles from the camera facing the participant; MoCap: motion-capture angles; RF: random forest on the same features.

Exercise	Templates, MP: BA	Sens	Spec	AUC	RF, MP: BA	AUC	Templates, MoCap: BA	AUC
Arm abduction	0.59	0.50	0.68	0.66	0.68	0.75	0.54	0.53
Arm V-to-W	0.55	0.58	0.51	0.54	0.53	0.59	0.60	0.63
Push-up	0.65	0.84	0.46	0.80	0.61	0.65	0.73	0.84
Leg abduction	0.66	0.54	0.77	0.70	0.76	0.82	0.67	0.74
Lunge	0.45	0.27	0.63	0.41	0.61	0.65	0.52	0.56
Squat	0.57	0.54	0.59	0.63	0.59	0.67	0.54	0.60
Mean	**0.58**	**0.55**	**0.61**	**0.62**	**0.63**	**0.69**	**0.60**	**0.65**

Figure 6 helps explain why the squat was hard. With motion capture, incorrect squats tend to go deeper than correct ones, and part of the incorrect group lies to the left of the template interval. With MediaPipe the knee angles are shifted upwards and spread out, so the two groups overlap more. Here the limiting factor is the measurement, not the template.

Because the score is a maximum over named features, each flagged repetition comes with the feature responsible for it. For correctly flagged repetitions measured with MediaPipe, the most frequent responsible features were:

- the minimum elbow angle for arm abduction (26 of 44), suggesting a bent arm;
- leg separation and trunk inclination for leg abduction (17 and 16 of 49);
- trunk inclination and hip extension for squats;
- hip angles for push-ups, which is consistent with the body sagging or piking.

The dataset does not label which mistake was made in each repetition, so we could not check these attributions directly. They are nonetheless the kind of statement a physiotherapist can act on (“the elbow bent to 120° at the top of the lift”), and a classifier’s probability is not.

5.4 Latency

On four virtual CPUs with 1920 × 1080 input, the full pose model took 39.3 ms per frame on average (25.5 frames/s), the lite model 29.2 ms and the heavy model 99.1 ms (Table 5). Computing the angles and evaluating a template took a median of 0.21 ms per frame, about 0.5% of the pose-estimation time. The template stage therefore adds nothing noticeable to the cost of the pose model, and with the full model the pipeline keeps up with a 15–25 frames/s camera stream.

Table 5
Per-frame processing time on four virtual CPUs (Intel Xeon 2.20 GHz, no GPU), 1920 × 1080 input, 300 frames after warm-up.

Stage	Mean (ms)	Median (ms)	95th percentile (ms)	Frames/s
Pose, lite model	29.2	28.7	30.9	34.3
Pose, full model	39.3	38.0	47.0	25.5
Pose, heavy model	99.1	98.6	102.4	10.1
Angles + template	–	0.21	–	–

6. Discussion

The pose estimator is not the main bottleneck. The most useful finding may be the small gap between camera and motion capture in Table 4. Replacing MediaPipe with a sixteen-camera optical system raised mean balanced accuracy by only two points, from 0.58 to 0.60. Neither came close to reliable detection. A large share of the mistakes in REHAB24-6 are therefore invisible to a description that keeps only the extremes of a few joint angles. Some likely explanations are errors of timing or smoothness, movement outside the plane the angle measures, and compensation at joints that the template does not include. Better pose estimation alone will not fix this. Richer features will be needed, such as angle trajectories compared with a reference by dynamic time warping, left–right asymmetry, or the velocity profile, and they can still be expressed as named, inspectable criteria.

The cost of interpretability was modest. The random forest outperformed the templates by about five points of balanced accuracy on MediaPipe features. Its advantage was not consistent across exercises, though, and on motion-capture features it shrank to under two points. In return, the template states which joint was outside its range and by how much, and a clinician can edit that range for an individual patient. A forest cannot be edited that way. Given how much both approaches are limited by the features, we think this trade is reasonable for a home system whose decisions a clinician has to trust.

Population templates do not transfer well between people. The below-chance result for lunges shows that ranges learned from some people can be wrong for others: body proportions and habitual movement differ, and the “correct” range for one participant may be another participant’s mistake. This supports the design choice in Arogya of having clinicians set and adjust ranges per patient. One practical route is to record a few repetitions that the patient performs correctly under supervision during a clinic visit and build that patient’s template around them.

What the angle accuracy allows. Hip, knee and trunk features agreed with motion capture to within roughly 5–11°, with correlations around 0.9. That is good enough for coarse feedback such as “go a little deeper” or “you are leaning forward”. It is not good enough to enforce a narrow post-operative limit. Errors of this size are consistent with earlier comparisons of RGB pose estimation and marker-based capture [8]. They are also inflated by real differences in where the two systems place joint centres. Much of the hip, knee and shoulder error is a stable bias (Table 2), so a per-joint calibration offset would remove part of it. Elbow

angles from a single camera are unreliable when the forearm points at the lens, and elbow-based limits should be treated with caution. The view analysis also gives simple setup advice. Placing the phone at roughly 45° to the patient, rather than straight in front, gave the best agreement for most angles and was much better for trunk lean.

Limitations. The participants were healthy adults performing staged mistakes, not patients with pain, weakness or restricted range. Templates were derived from data, not set by clinicians, so clinician-authored criteria were not tested. Repetitions were taken from the dataset's annotations, so automatic repetition counting was not evaluated. Correctness was a single binary label without the type of mistake, which prevented a direct check of the feedback attributions. Latency was measured on a cloud CPU rather than a phone. Mobile performance will depend on the device and on the use of GPU delegates. Finally, the evaluation covers one dataset and six exercises, and no study with patients or physiotherapists has yet been carried out with the Arogya prototype.

7. Conclusion

We have described Arogya, a camera-only home rehabilitation system in which exercise criteria are explicit joint-angle templates that a clinician can read and adjust, and we evaluated its measurement and decision stages on the public REHAB24-6 dataset. MediaPipe joint angles from a single RGB camera tracked motion capture closely for the trunk, hip and knee, and less well for the elbow. They supported correctness detection almost as well as motion capture did. The pipeline runs in real time on a modest CPU. Detection itself was only moderately accurate, whether with camera or motion-capture angles, and a random forest did somewhat better. Range-based templates on joint-angle extremes are a transparent starting point, but they are not sufficient on their own. Our next steps are per-patient templates recorded in the clinic, trajectory-based criteria that remain readable by clinicians, and a usability study with physiotherapists and their patients.

Declarations

Conflicts of interest.

The authors declare no conflicts of interest.

Use of AI tools.

Generative AI (Anthropic Claude) was used to help write and edit the evaluation code and to improve the language of the manuscript. The authors reviewed all code, results and text, checked every reference, and take full responsibility for the content.

Funding.

This research received no external funding.

Author contributions.

Conceptualisation and supervision: S.K.; system design and prototype: K.B., K.K., P.R.; evaluation code and analysis: K.B.; writing (original draft): K.B.; writing (review and editing): all authors.

Data and code availability.

REHAB24-6 is available from Zenodo (https://doi.org/10.5281/zenodo.13305826) under a CC BY-NC 4.0 licence. The code is available as two notebooks: pose extraction and feature computation at https://www.kaggle.com/code/kavyabhand/arogya-rehab24-eval, and the agreement, detection and latency analysis, with all result figures, at https://www.kaggle.com/code/kavyabhand/arogya-rehab24-analysis. The extracted repetition-level features are available at https://www.kaggle.com/datasets/kavyabhand/arogya-rehab24-features.

References

1. Argent R, Daly A, Caulfield B (2018) Patient involvement with home-based exercise programs: Can connected health interventions influence adherence? JMIR mHealth uHealth, 6, 3, e47
2. Peek K, Sanson-Fisher R, Mackenzie L, Carey M (2016) Interventions to aid patient adherence to physiotherapist prescribed self-management strategies: A systematic review. Physiotherapy 102(2):127–135. 10.1016/j.physio.2015.10.003
3. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T (2020) Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: A systematic analysis for the Global Burden of Disease Study 2019. Lancet 396(10267):2006–2017. 10.1016/S0140-6736(20)32340-0
4. Cottrell MA, Galea OA, O'Leary SP, Hill AJ, Russell TG (2017) Real-time telerehabilitation for the treatment of musculoskeletal conditions is effective and comparable to standard practice: A systematic review and meta-analysis. Clin Rehabil 31(5):625–638
5. Cao Z, Hidalgo G, Simon T, Wei S-E, Sheikh Y (2021) OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 43(1):172–186
6. Bazarevsky V, Grishchenko I, Raveendran K, Zhu T, Zhang F, Grundmann M BlazePose: On-device real-time body pose tracking, arXiv:2006.10204, 2020
7. Lugaresi C et al (2019) MediaPipe: A framework for building perception pipelines, arXiv:1906.08172
8. de Gusmão TB, Lafayette et al (2023) Validation of angle estimation based on body tracking data from RGB-D and RGB cameras for biomechanical assessment. Sensors 23(1). 10.3390/s23010003
9. Stenum J, Rossi C, Roemmich RT (2021) Two-dimensional video-based analysis of human gait using pose estimation. PLoS Comput Biol, 17, 4, e1008935

10. Liao Y, Vakanski A, Xian M (2020) A deep learning framework for assessing physical rehabilitation exercises. IEEE Trans Neural Syst Rehabil Eng 28(2):468–477
11. Mennella C, Maniscalco U, De Pietro G, Esposito M (2023) A deep learning system to monitor and assess rehabilitation exercises in home-based remote and unsupervised conditions. Comput Biol Med 166. 10.1016/j.compbiomed.2023.107485
12. Sardari S et al (2023) Artificial intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review. Comput Biol Med 158. 10.1016/j.compbiomed.2023.106835
13. Černek A, Sedmidubsky J, Budikova P (2024) REHAB24-6: Physical therapy dataset for analyzing pose estimation methods. in Similarity Search and Applications (SISAP 2024). Lecture Notes in Computer Science. Springer, Cham, Switzerland. doi: 10.1007/978-3-031-75823-2_2.
14. Xue J et al (2026) RePose: A real-time 3D human pose estimation and biomechanical analysis framework for rehabilitation, arXiv:2601.00625
15. Vakanski A, Jun H-P, Paul D, Baker R (2018) A data set of human body movements for physical rehabilitation exercises, *Data*, vol. 3, no. 1, 2
16. Capecci M et al (2019) The KIMORE dataset: KInematic assessment of MOvement and clinical scores for remote monitoring of physical REhabilitation. IEEE Trans Neural Syst Rehabil Eng 27(7):1436–1448

Figures

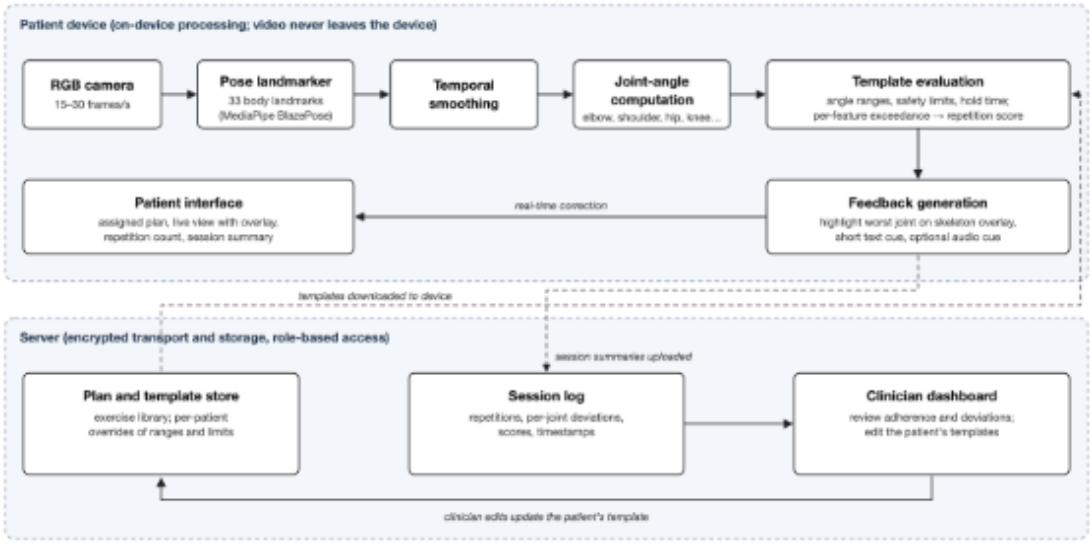
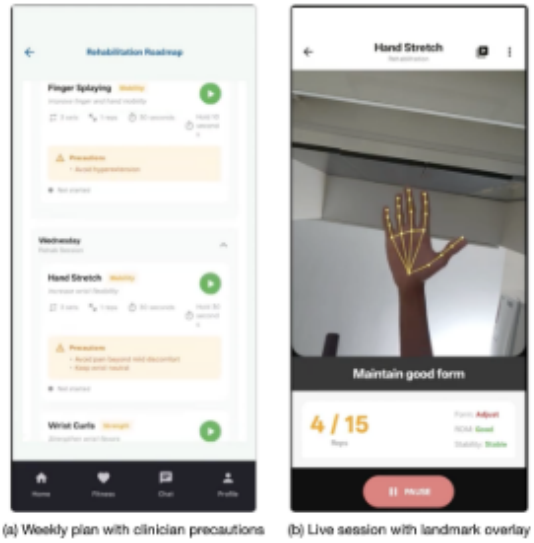


Figure 1

Arogya architecture. Solid arrows show data flow within the device or server; dashed arrows show synchronisation between them.



(a) Weekly plan with clinician precautions (b) Live session with landmark overlay

Figure 2

Mobile prototype: (a) a patient's weekly plan, with precautions entered by the clinician; (b) a live session with landmark overlay and repetition count.

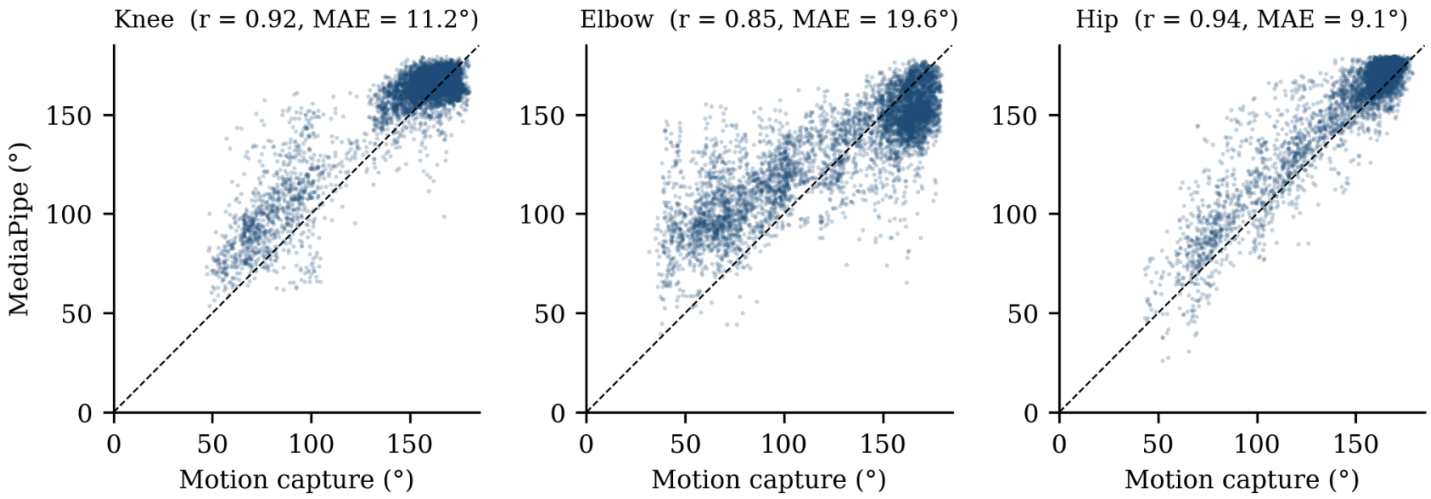


Figure 3

Repetition-level angle features from MediaPipe against motion capture for the knee, elbow and hip, with the participant facing the camera from the front or at half-profile. The dashed line marks perfect agreement.

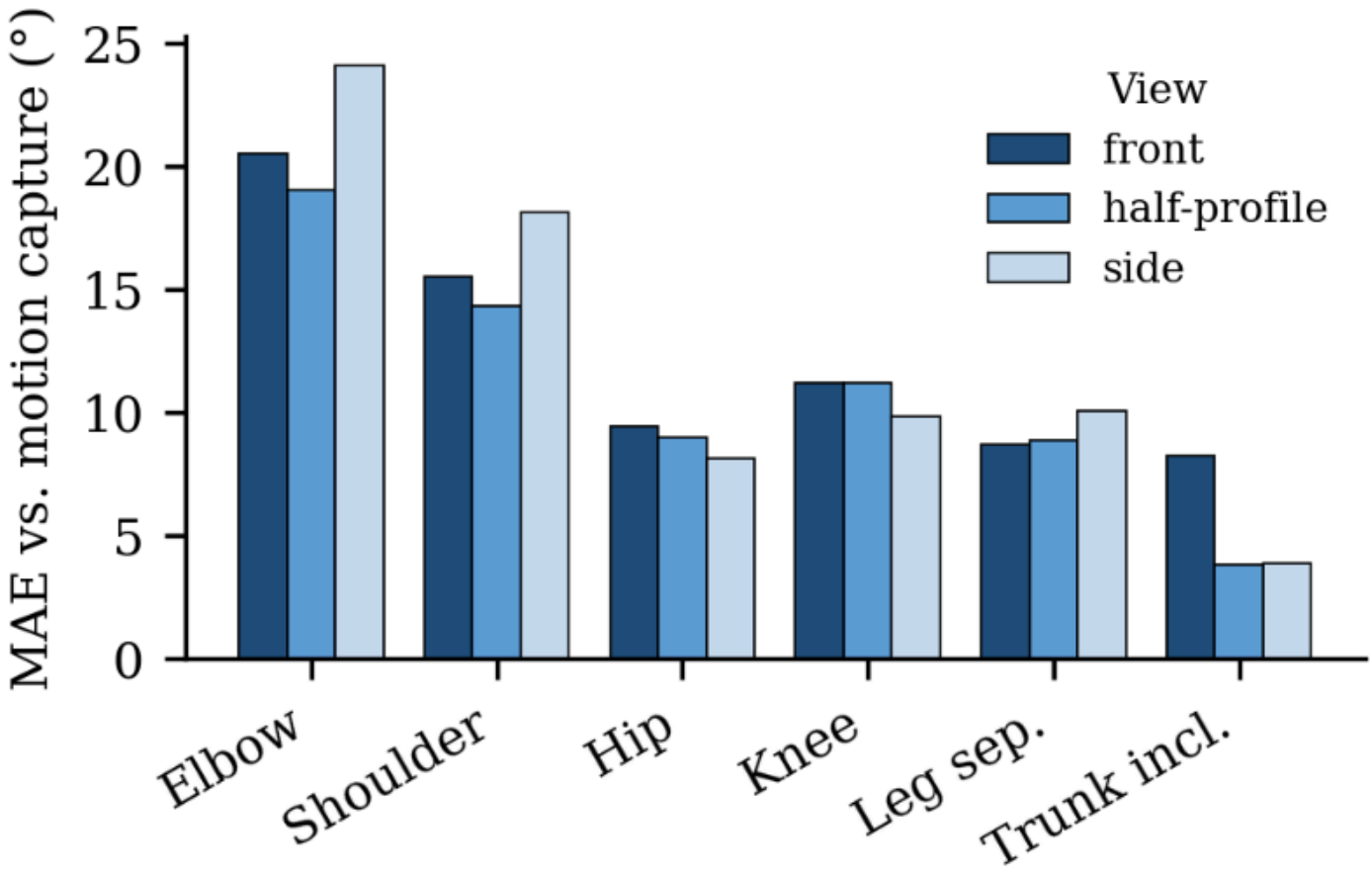


Figure 4

MAE against motion capture for each angle by the participant's orientation to the camera.

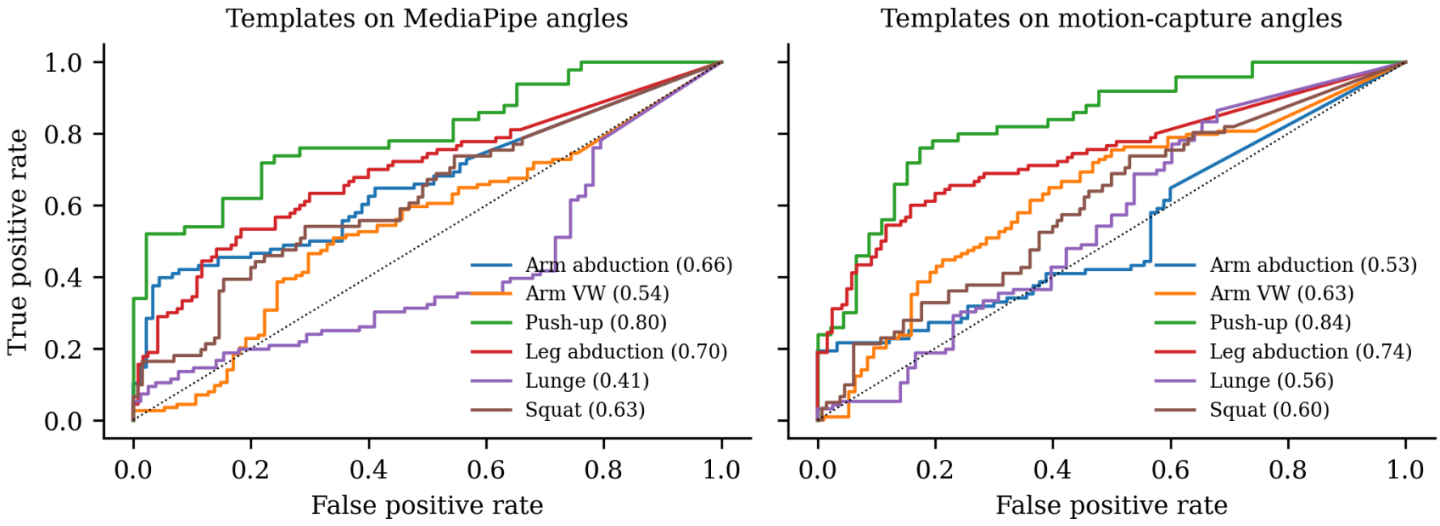


Figure 5

ROC curves of the template deviation score for each exercise, on MediaPipe angles (left) and motion-capture angles (right). AUC in parentheses.

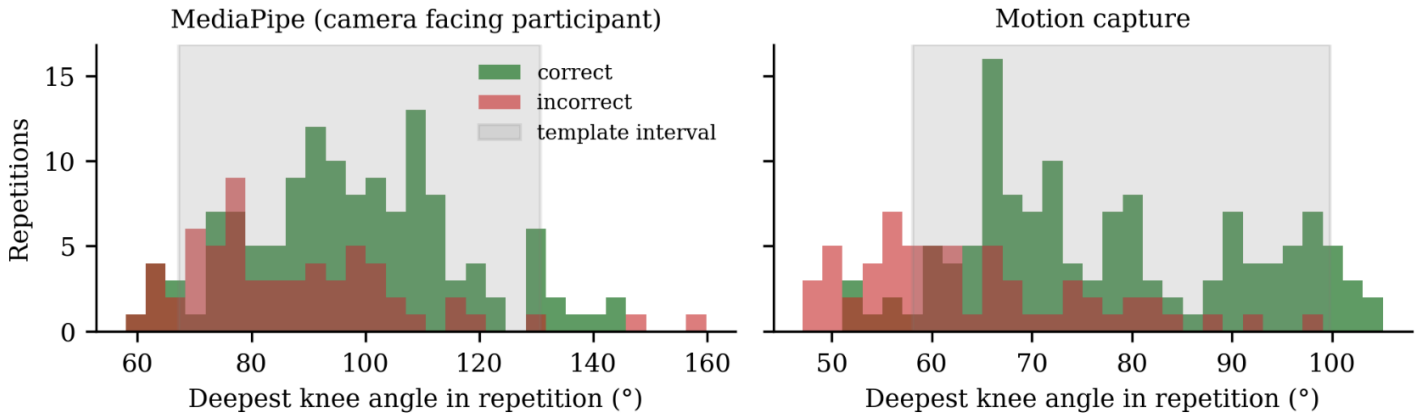


Figure 6

Deepest knee angle per squat repetition (10th percentile) for correct and incorrect repetitions, with the template interval learned from correct repetitions of all participants (shaded).