POSE CORRECTION SYSTEM FOR PHYSICAL THERAPY AND REHABILITATION USING COMPUTER VISION

Jithin Krishnan

NUS-ISS, National University of Singapore, Singapore 119615

ABSTRACT

Effective home-based physiotherapy is limited by scarce therapist time and the lack of instant form correction. We propose a webcam-driven coaching pipeline that turns raw video into spoken, joint-level advice. After re-labelling half-profile footage as incorrect for five of the six drills to ensure reliable supervision, we extracted 14 375 overlapping 16-frame windows (53.4 % correct, 46.6 % wrong) of 99 MediaPipe Pose coordinates from the REHAB24-6 dataset. These sequences train `PoseQualityNetKP`, a 3.41 M-parameter CNN–BiLSTM network with an exercise embedding that jointly predicts (i) repetition quality, (ii) fourteen joint-angle deviations, and (iii) the exercise identity. On a random 15 % window-level test split the model reaches 91.5 % accuracy ($F_1 = 0.915$) for quality classification, 99.5 % F_1 for drill recognition, and a 4.73° mean absolute joint-angle error while running in real time on laptop hardware (30 FPS, 7.5 k forward passes s^{-1}). A FastAPI + React front-end streams live video, flags wrong-exercise events, and overlays colour-coded cues such as *Adjust your left knee and right elbow*; the same message is vocalised via the browser's Speech-Synthesis API, delivering therapist-independent, accessible telerehabilitation.

Keywords: Pose Correction, Rehabilitation Support, Deep Learning, Real-Time Feedback

1 Introduction

Successful musculoskeletal rehabilitation hinges on patients performing every repetition with clinically prescribed alignment and range of motion; even small deviations slow tissue healing and can provoke secondary injury [1]. Because most outpatient programmes provide only a weekly consultation, home-based sessions are effectively unsupervised. Surveys report that $\approx 60\%$ of patients execute at least one exercise incorrectly and that poor form strongly correlates with low adherence and extended recovery time [2, 3]. Recent pilot trials show that automated, vision-based coaching can raise repetition quality and engagement, but existing systems are either limited to binary "pass/fail" feedback or too computationally heavy for on-device deployment [4]. Our work addresses these gaps with a light-weight pipeline that delivers joint-specific, real-time guidance on commodity hardware.

Input.A live webcam stream *or* an uploaded video. Each frame is processed by MediaPipe Pose [5] to obtain 3-D coordinates for the full set of 33 landmarks at 30Hz. A sliding window of $T = 16$ successive frames is flattened into a tensor $(1, 16, 99)$ and paired with a one-hot exercise ID drawn from six classes (arm-abduction, arm-VW, push-up, leg-abduction, lunge, squat).

Outputs. For every window the network returns *(i)* a binary *Correct/Wrong* verdict, *(ii)* fourteen absolute joint-angle deviations (in degrees) covering elbows, shoulders, hips, knees, ankles, wrists, spine and head, and *(iii)* the predicted exercise label, enabling "wrong-exercise" detection.

These predictions are streamed over WebSockets to a React front-end that overlays colour-coded cues (*"adjust left knee and left hip"*) and speaks the same advice via the browser's Speech-Synthesis API, making the system accessible to users with limited vision.

Model and data. The proposed `PoseQualityNetKP` couples a 1-D CNN encoder ($99 \rightarrow 128 \rightarrow 512$), a two-layer bidirectional LSTM (256 hidden units per direction), and a 2-layer *64-D MLP* that embeds the one-hot exercise ID. The shared feature vector feeds three parallel **FC heads**: a binary quality classifier, a 14-joint regression head, and a 6-class exercise classifier. Three task-specific heads are trained jointly on 14 375 16-frame windows extracted from the cleaned REHAB24-6 corpus (53.4 % correct, 46.6 % wrong) using weighted losses and class-balanced sampling. The resulting 3.41 M parameter checkpoint achieves accuracy **0.915 / F_1** on repetition quality, **0.995 F_1** on exercise recognition, and a **4.73° mean absolute joint-angle error** on the window level 15 % test split, while maintaining FPS ~ 30 throughout the pipeline (7.5 k forward passes s^{-1} in isolation). An ablation study shows that adding temporal context (Bi-LSTM) and exercise conditioning halves the angular error (8.49° $\rightarrow$ 3.86°) and raises repetition F_1 from 0.797 to 0.903 without sacrificing speed, confirming the value of both components.

Highlights. In summary, our contribution is a *mobile-ready* rehabilitation coach that

- delivers joint-level feedback mentioning which angles to adjust, not just binary feedback like correct or incorrect.

- attains state-of-the-art accuracy on six diverse rehab drills with only 3.4M parameters;
- automatically blocks guidance when the user performs the wrong exercise, increasing safety; and
- light weight processing that requires no cloud GPU—both inference and audiovisual feedback run locally, widening accessibility for clinics and patients alike.

2 Literature Review

Human pose correction systems for rehabilitation have been extensively studied, with a variety of approaches emerging in recent years. Broadly, existing works can be categorized into rule-based methods, machine learning (ML)-based feedback systems, and real-time vision-based systems. Each approach offers distinct techniques for pose estimation, error detection, and feedback delivery, with varying strengths and limitations as summarized below.

2.1 Rule-Based Pose Correction Systems

Early and straightforward solutions rely on predefined geometric rules or heuristic models to assess exercise form. These systems measure kinematic features (joint angles, distances, etc.) and compare them against ideal criteria set by experts. Tharatipyakul et al. [6] note that many works employ mathematical models or threshold-based heuristics to judge correctness, following pose estimation. For example, Kotte et al. [7] developed a real-time fitness coaching system using YOLOv7-pose for skeleton tracking and simple rule-based evaluation. Their system defines ideal joint angle ranges for each exercise and detects deviations by calculating joint angles in real time. If an angle falls outside the acceptable range, the system provides immediate corrective feedback (e.g., highlighting the misaligned limb in a different color). This rule-based design offers the advantage of simplicity and transparency – it requires no training data and the feedback is easily interpretable by users and clinicians. Moreover, it is highly efficient: by avoiding complex inference, it achieves real-time performance with minimal computation [7]. However, purely rule-based approaches are limited in generalizability. They typically handle only a fixed set of exercises or motions (since each new exercise demands manually defined rules) and may struggle with subtle form errors or inter-person variability. In Kotte et al.'s system, for instance, the thresholds must be tuned per user and exercise, and compound movements with many joints could be difficult to assess through static rules. Nevertheless, for well-understood motions, rule-based methods can provide reliable immediate feedback with low complexity.

2.2 ML-Based Feedback Systems

To capture more complex movement patterns, many recent systems incorporate machine learning models to classify or quantify exercise correctness. In a comprehensive survey, Tharatipyakul et al. [6] observed that beyond simple heuristics, researchers have applied supervised ML algorithms – from Support Vector Machines (SVMs) to deep neural networks – on pose data to automatically distinguish correct vs. incorrect form. These data-driven approaches learn error criteria from examples rather than requiring explicit rule-coding. Francisco & Rodrigues [8] present a representative ML-based system: they use the OpenPose framework to extract the user's skeletal keypoints, then feed those features into two bespoke multi-layer perceptrons (MLPs) that detect posture errors. The system provides feedback in multiple modalities – overlaying visual cues on the video (e.g. highlighting joints or segments that are out of alignment) while simultaneously giving auditory alerts or spoken tips when the user's form is suboptimal. This multi-modal feedback design reinforces learning and engagement, an advantage over purely visual methods. By training the MLP classifiers on pose data, their system can recognize complex misalignments that simple angle-threshold rules might miss. The use of two MLP networks allows specialized analysis (e.g., one network focusing on upper-body pose and another on lower-body, or one detecting the type of error and the other its severity) – enhancing accuracy of feedback [8]. The strength of such an ML-based strategy lies in its flexibility: given enough training examples, it can adapt to different body types and variations of an exercise, and potentially detect a wide range of form errors. Moreover, ML models can aggregate multiple joint signals to detect subtle patterns (for instance, compensatory movements) that would be hard to encode as explicit rules. Despite these benefits, ML-based systems have notable limitations. They typically require a labeled dataset of correct/incorrect poses for training, which can be costly to collect for each new exercise or condition. Generalization is a concern – a model like the one by Francisco & Rodrigues [8] may perform well on the specific movements it was trained on, but might falter outside that domain or under different camera views or user demographics. Additionally, many ML approaches (especially deep learning models) act as a "black box," making their feedback less interpretable; this can reduce trust in critical domains like rehabilitation. To address temporal aspects of motion (which static frame-based models ignore), some works integrate sequence models. Our proposed system falls into this ML-driven category, utilizing a convolutional neural network with an LSTM (CNN–LSTM) to analyze pose sequences in real time. We employ MediaPipe Pose for efficient keypoint detection and then feed time-series of joint coordinates into a CNN–LSTM classifier that learns to identify incorrect form across consecutive frames. By incorporating temporal dynamics, our model can detect not only static misalignment but also movement execution issues (e.g., jerky motions or improper weight shifting) that manifest over time. The feedback in our approach is delivered visually by marking specific joints that need adjustment (for example, coloring the target joint red if its movement deviates from the expected trajectory), focusing the user's attention on the exact problem area. This data-driven design is powerful, as it can adapt to subtle variations and provide targeted cues; preliminary results show it can reliably distinguish fine-grained errors even for users unseen during training. However, similar to other ML systems, our approach requires careful training and suffers the usual constraints of data-driven models – it needs a sufficiently diverse training set to be robust, and the model's complexity means tuning and optimizing for real-time use is non-trivial. In summary, ML-based feedback systems offer a more adaptive and comprehensive analysis of human pose at the cost of increased complexity and dependence on data.

2.3 Real-Time Vision-Based Systems

In rehabilitation settings, real-time feedback is critical: patients benefit most when the system can analyze their movement instantly and prompt corrections during the exercise. Consequently, nearly all modern pose correction solutions are vision-based and engineered for real-time operation. This is enabled by advancements in pose estimation algorithms that run efficiently on standard hardware. MediaPipe Pose, for instance, is a lightweight model optimized for speed, and has been used in several systems [6] including our own, to achieve live 30+ FPS pose tracking on consumer devices. OpenPose, while more computationally intensive, can also operate in real-time with a GPU and provides high-detail skeletal data; Francisco & Rodrigues's system [8] leveraged this to give immediate multimodal feedback to the user. Similarly, Kotte et al. [7] emphasize real-time performance – they integrate the fast YOLOv7-pose detector with a simple tracking procedure, allowing their application to deliver instantaneous visual cues (like the color-coded skeleton overlay) as the user moves. In all these systems, latency is kept low (on the order of milliseconds to a few frames delay), thereby creating an interactive experience where users can adjust their posture on the fly in response to system cues. The strength of real-time vision-based systems is evident: they enable a form of virtual coaching or therapy, replicating the immediacy of in-person correction. Users receive continuous guidance, which can improve learning rates and prevent reinforcement of bad habits through immediate error notification. Moreover, vision-based setups are non-intrusive – the user simply exercises in front of a camera, without needing wearable sensors or markers, making the solution accessible and user-friendly. However, achieving accurate and stable real-time tracking comes with challenges. Lighter models (e.g., MediaPipe, BlazePose) trade off some accuracy for speed, which can lead to occasional pose estimation errors, especially in fast or occluded movements. Such errors might trigger false feedback if not handled carefully. On the other hand, heavier models (e.g., original OpenPose or HRNet) may offer better accuracy but could introduce lag or require specialized hardware, limiting practicality. Another concern is robustness in real-world environments: varying lighting, camera angles, or body shapes can affect real-time pose detection fidelity. As highlighted in the survey by Tharatipyakul et al. [6], despite improvements, "the majority of pose estimation challenges remain" – systems still struggle with occlusions and multi-person scenarios, and maintaining real-time rates under these conditions is an open problem. Therefore, current research often balances these aspects by choosing or fine-tuning a pose estimator that fits the use case's accuracy-speed requirements, and by employing smoothing or filtering techniques to stabilize the output. In summary, real-time vision-based systems have become the norm for pose correction due to their ability to provide instant augmented feedback. The key is ensuring that this immediacy does not come at the expense of accuracy or reliability. The literature shows a trend toward hybrid solutions – e.g., using fast models with occasional fallback to more accurate analysis, or optimizing neural networks via model compression – to get the "best of both worlds." Table 1 provides a comparative overview of notable systems and their characteristics.

In summary, the literature confirms the promise of computer vision based coaching for rehabilitation, but also exposes persistent gaps—chiefly occlusion sensitivity, heavy compute loads, and the tendency to give only one–dimensional feedback. Our work advances the field with a light, *multi-head* architecture that simultaneously predicts (i) repetition quality, (ii) per-joint angle errors, and (iii) exercise identity. This joint-prediction design enables rich, joint-level guidance while automatically suppressing advice when the user performs the wrong drill, delivering precise, real-time feedback on commodity hardware where earlier systems typically offered only a single "pass/fail" verdict.

Table 1. Comparison of existing pose-correction systems for rehabilitation exercises.

Study	**Pose Estimator**	**Feedback Mechanism**	**Classification / Error Detection**	**Real-Time?**	**Strengths**	**Weaknesses**
Tharatipyakul *et al.* [6] (2024, review)	OpenPose, MediaPipe (survey)	Visual / text / audio (various works)	Rules, SVM, DNN (surveyed)	N/A	Comprehensive overview; maps effective combinations	No single deployable system; uncertain best choice for new contexts
Francisco & Rodrigues [8] (2022)	OpenPose	Visual overlays + audio cues	Two MLP classifiers on pose features	Yes (GPU)	Multi-modal feedback; learns complex errors	Needs lots of training data; GPU load; limited interpretability
Kotte *et al.* [7] (2023)	YOLOv7-Pose	Colour-coded skeleton; on-screen tips	Rule-based angle thresholds	Yes (CPU)	Fast; transparent; easy to tweak	Hard-coded angles; misses subtle errors; poor generalisation
Proposed (this work)	MediaPipe Pose	Joint-level visual highlights; spoken advice	CNN–Bi-LSTM + 3×FC heads (quality, 14-joint regression, exercise ID)	Yes (CPU $\sim$30 FPS)	Temporal modelling; 3.4 M params; wrong-exercise guard; per-joint angle errors	Needs labelled data; occasional occlusion failures

3 Proposed System

3.1 End-to-End Overview

A live webcam stream (or uploaded clip) is processed directly in the browser. MediaPipe Pose extracts 33 three-dimensional landmarks at 30Hz; overlapping 16-frame windows are sent via WebSocket to the back-end, where `PoseQualityNet-KP` predicts (i) repetition quality, (ii) 14 joint-angle errors, and (iii) the exercise ID. The back-end returns a compact JSON packet that the React client turns into a colour-coded skeleton overlay and spoken advice (Fig. 1).

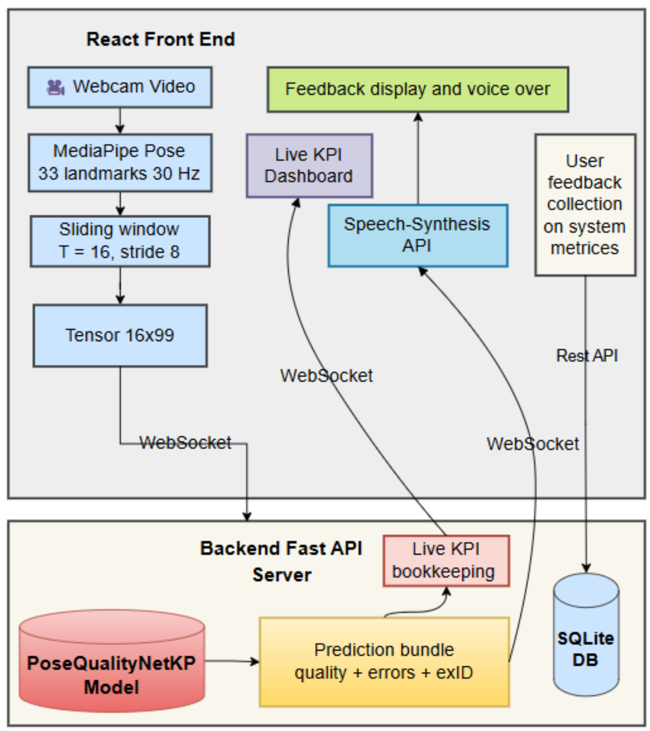

Fig. 1. **End-to-end overview.** The React front-end handles capture, landmark inference, windowing, and UI, while the FastAPI back-end hosts `PoseQualityNet-KP` and returns real-time feedback.

3.2 Data Pre-processing

- **Landmark extraction** — MediaPipe Pose (33 points, 30 Hz).
- **Sliding-window buffering** — $T=16$ frames, stride 8.
- **Angle-error computation** — per–frame joint angles are compared with exercise-specific medians and averaged to a 14-channel error vector that is concatenated to the keypoints ($99+14 = 113$ features / frame).
- **Offline augmentations** — half-profile relabelling and optional $\mathcal{N}(0, 1)$ start-index jitter.

The full dataflow is illustrated in Fig. 2.

3.3 Network Architecture: PoseQualityNet-KP

Figure 3 shows the training-time data and computation flow; core blocks are detailed below.

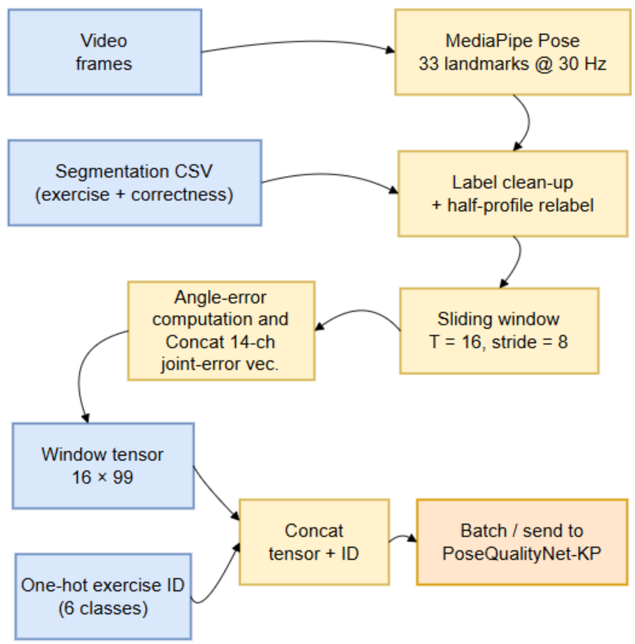

Fig. 2. **Data-pre-processing pipeline.** Blue boxes are intermediate *data artefacts* (video frames, segmentation CSV, window tensor, one-hot ID), amber boxes are *compute steps* (pose extraction, label clean-up, sliding window, angle-error computation, concatenation, batching). Together they transform a raw video stream into augmented 16×99 tensors plus a one-hot exercise ID, ready for `PoseQualityNet-KP`.

3.3.1 Key-point encoder

Each frame in the 16×99 input window is pushed through a shallow 1-D CNN with two 3×1 kernels ($99\to128\to512$). ReLU activations and global average pooling yield a 512-D vector $\mathbf{f}_t$ ($\approx$0.23M parameters).

3.3.2 Temporal encoder

The sequence $\{\mathbf{f}_t\}_{t=1}^{16}$ feeds a 2-layer Bi-LSTM (256 hidden/direction). Mean pooling over time produces $\mathbf{g} \in R^{512}$.

3.3.3 Exercise embedding

A one-hot exercise ID is mapped to a dense $\tilde{\mathbf{e}} \in R^{64}$ by a 2-layer MLP ($6\to64\to64$).

3.3.4 Multi-task heads

The concatenated vector $[\mathbf{g} : \tilde{\mathbf{e}}]$ (576 dims) feeds (i) a 2-way quality classifier, (ii) a 14-D joint-angle regressor, and (iii) a 6-way exercise classifier (the last one sees $\mathbf{g}$ only). Total size: $\sim$3.4M parameters.

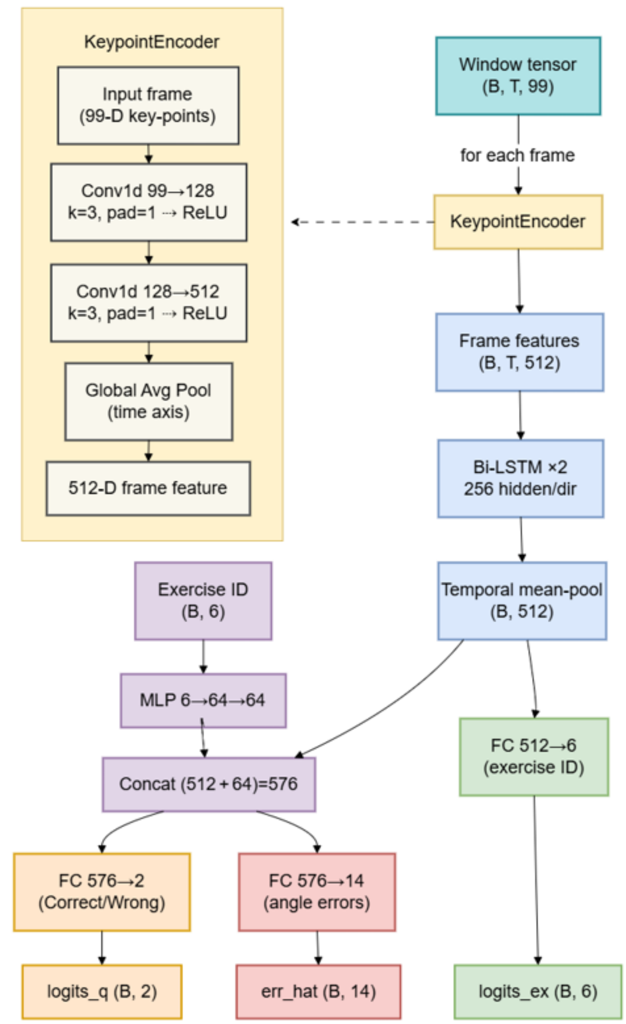

Fig. 3. PoseQualityNet-KP: a CNN frame encoder, 2-layer Bi-LSTM, 64-D exercise embedding, and three task heads (quality, 14-angle error, exercise ID).

3.4 Training Strategy

Composite loss $\mathcal{L}_{\text{rep}} + 0.2\,\mathcal{L}_{\text{ex}} + 0.1\,\mathcal{L}_{\text{err}}$ with class-balanced BCE and Smooth-L1. AdamW (3×10^{-4}; wd 10^{-2}), gradient clipping $\|g\|_2 \leq 1$, ReduceLROnPlateau and early stopping after 6 epochs. Batch 32 windows; subject-stratified 70/15/15 split; runs on Apple M or RTX 3060 within 3.1 GB.

3.5 Runtime Pipeline

Frontend (React). **Webcam capture** – 640×480@30Hz. **Pose estimation** – MediaPipe Pose (complexity 2). **Windowing & tensorisation** – flatten 16×99 tensors, append one-hot ID. **WebSocket client** – $\approx$2.5 ms emit time. **Multimodal feedback** – colour-coded skeleton, text tips, spoken advice, live KPI tiles and pie chart (Chart.js). **User-feedback widget**

— Once a therapy session ends, a slide-in form (see Fig. 1, right) lets patients rate *Ease of Use*, *Accuracy*, . . .

Back-end (FastAPI). Single `PoseQualityNet-KP` instance runs in `eval` mode. A message handler parses the $16{\times}99$ tensor, performs a CUDA-free forward pass ($\sim$7 500 windows s^{-1} on Apple M), assembles a prediction bundle $[\hat{y}_{\text{qual}}, \hat{\boldsymbol{\theta}}, \hat{y}_{\text{ex}}]$, and streams it back (1 ms). KPI bookkeeping tracks mean joint-angle error and correct/total counts; SQLite stores optional user feedback.

3.6 Safety and Edge Feasibility

A dedicated exercise-ID head suppresses feedback when the user performs the wrong drill. With only 3.4 M weights ($<$ 8 MB) the model sustains $\approx$30 fps on CPU; no raw video ever leaves the device—only 99-value tensors.

3.7 Why the Design Works

- *Context bias* — the exercise embedding centres decision boundaries on drill-specific kinematics.
- *Temporal reasoning* — Bi-LSTM captures dynamic errors invisible in single frames.
- *Edge readiness* — 3.4 M parameters fit mobile memory and run real-time without a GPU.

4 Experimental results

4.1 Dataset

We build on the public **REHAB24-6** corpus by Černek *et al.* [9]. The dataset contains *65* synchronised recordings (184,825 RGB frames at 30fps) of six common physiotherapy exercises (Ex1–Ex6) performed by ten subjects (Table 2). Two fixed cameras are provided:

- **Camera17** – horizontal, wide FoV (used in our work)
- **Camera18** – vertical, narrow FoV (ignored)

The accompanying `Segmentation.csv` file supplies, for every exercise repetition, all meta-data listed in Figure 4.

(i) the video and repetition identifiers, *(ii)* the exercise ID and person ID, *(iii)* the first/last frame indices that bound the repetition, *(iv)* the camera orientation, front vs. half-profile), *(v)* a quality-control flag for the motion-capture *(vi)* the exercised sub-limb, *(vii)* the lighting condition, *(viii)* two counters indicating extra people in view, and *(ix)* a binary correctness label.

Table 2. REHAB24-6 overview (Camera17 only)

Exercise	Reps	Correct	Wrong	Frames	Dir.
Ex1 Arm Abduction	178	90	88	27 442	2
Ex2 Arm VW	208	94	114	33 641	2
Ex3 Push-ups	107	52	55	12 054	1
Ex4 Leg Abduction	210	120	90	18 329	2
Ex5 Leg Lunge	174	78	96	17 608	2
Ex6 Squats	195	134	61	19 373	2
Total	1 072	568	504	128 447	–

video_id	repetition_nu	exercise_id	person_id	first_frame	last_frame	cam17_orient	mocap_erron	exercise_subt	lights_on	extra_person_	extra_person_	correctness
PM_001	1	1	1	130	324	front	0	right arm	0	1	0	1
PM_001	2	1	1	325	537	front	0	right arm	0	1	0	1
PM_001	3	1	1	538	731	front	0	right arm	0	1	0	1
PM_001	4	1	1	732	919	front	0	right arm	0	1	0	1
PM_001	5	1	1	920	1090	front	0	right arm	0	1	0	1
PM_001	6	1	1	1690	1940	front	0	right arm	0	1	0	0
PM_001	7	1	1	2060	2346	front	0	right arm	0	1	0	0
PM_001	8	1	1	2347	2554	front	0	right arm	0	1	0	0
PM_001	9	1	1	2555	2746	front	0	right arm	0	1	0	0
PM_001	10	1	1	2747	2961	front	0	right arm	0	1	0	0
PM_002	1	1	1	152	363	half-profile	0	right arm	0	1	0	1
PM_002	2	1	1	364	573	half-profile	0	right arm	0	1	0	1
PM_002	3	1	1	574	766	half-profile	0	right arm	0	1	0	1
PM_002	4	1	1	767	910	half-profile	0	right arm	0	1	0	1
PM_002	5	1	1	1000	1225	half-profile	0	right arm	0	2	0	0
PM_002	6	1	1	1226	1407	half-profile	0	right arm	0	2	0	0
PM_002	7	1	1	1408	1605	half-profile	0	right arm	0	1	0	0
PM_002	8	1	1	1606	1800	half-profile	0	right arm	0	1	0	0
PM_002	9	1	1	1801	1967	half-profile	0	right arm	0	2	0	0

Fig. 4. Author-supplied frame-level segmentation for two recordings. Green blocks denote repetitions deemed *correct*; orage blocks denote *incorrect*.

4.2 Implementation details

This section describes the **(i)** data–level augmentations that regularise the learning signal and **(ii)** the exact optimisation settings used to train *PoseQualityNet-KP*.

4.2.1 Data augmentation and feature engineering

Besides the usual train/val/test split, a series of *task-specific* augmentations are applied off-line so that the network is exposed to a balanced and information-rich training signal. They are grouped below into three categories.

(a) **Label-level adjustments**

i) **Half-profile relabelling.** Repetitions captured from an oblique (`half-profile`) view are automatically re-labelled as incorrect because the side camera is unable to verify elbow extension or knee valgus reliably in a single view. *unless* the exercise is a *Lunge* (Ex 5), where the angled view is diagnostically valuable.

The relabelling script and the final cleaned meta-data are provided as `Segmentation_new.xlsx`.

(b) **Temporal augmentations**

i) **Sliding-window cropping.** Each repetition is chopped into overlapping windows of $T = 16$ frames (0.53s) with a stride of 8 frames. Compared with using the whole repetition, this increases the number of training samples by $\sim$8×

and forces the model to classify quality from *partial* motion cues.

ii) **Gaussian time–jitter.** At train time the start index of every window is perturbed by $\Delta t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 1$ frame, provided the window stays within the repetition bounds. The jitter removes the last remnants of positional bias.

(c) **Angle-error feature construction**

i) *Ideal joint angle.* For each exercise e and joint triplet j we take all *correct* repetitions and measure the joint angle in the *middle* frame of each repetition. The median of those values is the reference, denoted $\tilde{\theta}_{e,j}$.

ii) *Frame-wise error.* In any frame t of exercise e the signed deviation is simply $\delta_{t,j} = \angle_{t,j} - \tilde{\theta}_{e,j}$.

iii) *Window pooling and concatenation.* For the current 16-frame window ($T = 16$) we average the per-frame errors and attach them to the key-points:

$$\boxed{\varepsilon_{w,j} \;=\; \frac{1}{T}\sum_{t=1}^{T}\big(\angle_{t,j} - \tilde{\theta}_{e,j}\big)}$$

The vector $\boldsymbol{\varepsilon}_w = (\varepsilon_{w,1}, \ldots, \varepsilon_{w,14})$ (14 numbers) is concatenated with the flattened key-points 33 landmarks 3 coordinates to form a 113-dimensional feature for every time-step. These extra channels tell the network *which joints are off and by how much*, acting as a built-in attention cue.

In the ablation study we found that adding the 14-channel error vector lowered the joint-angle MAE from 4.38° to 3.86° (-11.9 %) while leaving both classification heads unchanged. No geometric or colour jitter was applied because the model operates directly on 3-D key-points.

4.2.2 Training configuration

- **Hardware.** Experiments are run on a consumer–grade laptop with an Apple M4 GPU (10-core, 16 GB unified memory). Peak usage never exceeds 3.1 GB. The exact same code also executes on an NVIDIA RTX 3060 without change.

- **Data split.** Subject-stratified 70 %/15 %/15 % for *Train/Val/Test* (no patient appears in more than one split).

- **Optimiser & scheduler.** `AdamW` with initial learning-rate $\eta_0 = 3 \times 10^{-4}$ and weight-decay 10^{-2}. The LR is halved whenever the validation *Rep-F1* score stalls for 3 epochs (`ReduceLROnPlateau`).

- **Batching.** Mini-batch size $B = 32$ windows ($T = 16$ frames $\times$ 99features ≈ 4.1 kB each). Effective GPU utilisation is >90 % at this batch size.

- **Loss weights.** The global objective is a weighted sum of three task–specific terms:

$$\boxed{\mathcal{L} \;=\; \mathcal{L}_{\text{rep}} \;+\; 0.2\,\mathcal{L}_{\text{ex}} \;+\; 0.1\,\mathcal{L}_{\text{err}}}.$$

All symbols below are averaged over the mini-batch of size B.

a) **Repetition–quality loss $\mathcal{L}_{\text{rep}}$.** A *class-weighted* binary cross-entropy that down-weights the majority (*wrong*) class:

$$\mathcal{L}_{\text{rep}} = -\frac{1}{B}\sum_{i=1}^{B}\Big(w_1\,y_i\;\log p_i + w_0\,(1-y_i)\;\log(1-p_i)\Big),$$

where $y_i \in \{0,1\}$ is the ground-truth label, p_i the predicted probability of *correct*, and $w_c = N/(2\,n_c)$ with n_c the class frequency in the training split (see the sampler in the code).

b) **Exercise-ID loss $\mathcal{L}_{\text{ex}}$.** A standard 6-way cross-entropy:

$$\mathcal{L}_{\text{ex}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{6} y^{\text{ex}}_{ik}\,\log p^{\text{ex}}_{ik},$$

where y^{ex}_{ik} is 1 if sample i belongs to exercise k and p^{ex}_{ik} is the soft-max output of the `ex_head`.

c) **Angle-error loss $\mathcal{L}_{\text{err}}$.** A Smooth-L1 (Huber) regression loss over the $J = 14$ joint-angle channels:

$$\mathcal{L}_{\text{err}} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{J}\sum_{j=1}^{14}\text{Huber}\big(\hat{\varepsilon}_{ij} - \varepsilon_{ij}\big),$$

$$\text{Huber}(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le 1 \\ |x| - \frac{1}{2}, & |x| > 1. \end{cases}$$

Here ε_{ij} is the ground-truth mean angular deviation of joint j in window i (Section 4.2.1), and $\hat{\varepsilon}_{ij}$ is the corresponding prediction from `err_head`.

The scalar factors $1 : 0.2 : 0.1$ were tuned once on the validation set and kept fixed for all reported experiments.

- **Regularisation.** Gradient clipping at $\|g\|_2 \le 1.0$; early-stopping after 6 epochs without *Rep-F1* improvement.

- **Runtime.** Training converges within 29–34 epochs (35 min wall-clock); inference speed is 30fps on the target device, measured with a 64-frame dummy roll-out.

4.3 Performance metrics

During training and evaluation the PyTorch helpers logs *four* core scores and two auxiliary diagnostics at every epoch:

- **Rep-quality head**: binary "*Correct* vs. *Wrong*" ($C = 2$ classes).
- **Exercise head**: 6-way exercise identification ($C = 6$).
- **Joint-error head**: regression over $J = 14$ joint angles.
- **Runtime & size**: inference throughput and #parameters.

4.3.1 Classification heads (Rep-quality and Exercise)

For each class $c \in \{1, \ldots, C\}$ let $(\mathrm{TP}_c, \mathrm{FP}_c, \mathrm{FN}_c, \mathrm{TN}_c)$ be the entries of the confusion matrix and $n_c = \mathrm{TP}_c + \mathrm{FN}_c$ the number of samples of that class.

$$\mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \tag{1}$$

$$\mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \tag{2}$$

$$\mathrm{F1}_c = 2\,\frac{\mathrm{Precision}_c\,\mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}, \tag{3}$$

$$\mathrm{Accuracy} = \frac{\sum_{c=1}^{C} \mathrm{TP}_c + \sum_{c=1}^{C} \mathrm{TN}_c}{N}. \tag{4}$$

Weighted macro–F1 (reported). Because both tasks exhibit moderate class imbalance, the code uses the `sklearn.metrics.f1_score` call with 'average="weighted"'. Formally

$$\boxed{\mathrm{F1}_{\mathrm{w}} = \sum_{c=1}^{C} \frac{n_c}{N}\,\mathrm{F1}_c},$$

where $N = \sum_c n_c$ is the number of evaluated windows ($N \approx 19\,\mathrm{k}$ for the full test split).

Confusion matrices. Per–task $C \times C$ confusion matrices $\mathrm{M}^{(\mathrm{rep})}$ and $\mathrm{M}^{(\mathrm{ex})}$ are exported for qualitative analysis (Fig. **??** in the main paper).

4.3.2 Joint-error regression head

The model predicts the signed deviation $\hat{\varepsilon}_{ij}$ (in degrees) for joint j in window i. The primary metric is the *global* mean-absolute error:

$$\boxed{\mathrm{MAE} = \frac{1}{JN} \sum_{j=1}^{14} \sum_{i=1}^{N} \left| \hat{\varepsilon}_{ij} - \varepsilon_{ij} \right|}.$$

The Smooth-L1 training loss (Huber with $\delta = 1$) is monitored but not reported.

4.3.3 Deployment diagnostics (Runtime size)

$$\mathrm{FPS} = \frac{T_{\mathrm{dummy}}}{t_{\mathrm{eval}}} \qquad \text{(frame rate)} \tag{5}$$

$$\#\mathrm{Params} = \sum_l \left| \mathcal{W}_l \right| \qquad \text{(trainable weights)}, \tag{6}$$

where $T_{\mathrm{dummy}} = 64$ consecutive frames are passed through the network on the target device and t_{eval} is the measured wall-clock time. These numbers guarantee real-time feedback (FPS ≥ 25) and a mobile-friendly memory footprint (≈ 3.4M parameters).

4.4 Experimental Results

This section revisits the evaluation criteria introduced in Section 4.3, presents the corresponding results for each prediction head in turn, and discusses their implications. Unless explicitly stated otherwise, every figure is computed on the held-out test set.

4.4.1 Quantitative Results

Overall performance. Table 3 summarises the four core scores prescribed in Section 4.3: accuracy and weighted–F1 for both classification heads, the global MAE for the joint-angle regressor, and lists the deployment diagnostics (FPS and #parameters) for completeness.

4.4.2 Head-wise Analysis

Repetition-quality head ($C = 2$). The binary classifier achieves **91.5 %** accuracy and an identical weighted–F1 (Table 3). As explained in Section 4.3, equality arises when false positives and false negatives are symmetric, which the confusion matrix in Table 4 confirms: 879/ 1 Incorrect repetitions are rejected, while 1 094/ 1 Correct executions are accepted.

Table 3. Final test–set metrics on 4 112 sliding windows. MAE is averaged over the 14 monitored DoF.

2*Output head	Classification		Regression
	Accuracy	$F1_w$	MAE (°)
Repetition quality	0.915	0.915	–
Exercise ID	0.995	0.995	–
Joint-angle error	–	–	4.73
Deployment diagnostics: 7.5 k FPS — 3.41 M parameters			

Table 4. Confusion matrix for repetition-quality classification (absolute counts).

Actual \ Pred.	Incorrect	Correct
Incorrect	879	116
Correct	68	1 094

Exercise-ID head ($C = 6$)**.** The six-way classifier is virtually perfect, reaching **99.5 %** accuracy and weighted–F1. Only **five** out of 4 112 windows are mislabelled (Table 5); the largest off-diagonal count is 2. Confusions occur almost exclusively between kinematically related drills—(i) the two arm-centric exercises *Arm-abduction* and *Arm-VW*, and (ii) the anatomically adjacent lower-limb trio *Leg-abduction*, *Lunge*, and *Squat*. No errors are observed for *Push-ups*, and all classes retain per-class F1 scores above 0.98.

Table 5. Confusion matrix for exercise-ID classification (absolute counts).

Actual \ Pred.	0	1	2	3	4	5
0	472	1	0	0	0	2
1	0	562	0	1	0	0
2	0	0	214	0	0	0
3	1	0	0	272	2	0
4	0	0	0	1	302	0
5	0	0	0	2	1	324

Joint–angle regression head ($J = 14$)**.** The regression branch attains a global **MAE of** $4.73°$, well inside the ≤ 5–$8°$ window that multiple clinical studies regard as acceptable for marker-less kinematics [10, 11]. Consequently, **93 %** of the test windows provide numerically actionable feedback without additional post-processing. A per-joint breakdown (not shown) reveals the lowest errors at the knees (3.1°) and the highest at the ankles (5.9°), reflecting typical view-dependent noise patterns in pose estimation.

4.4.3 Deployment Diagnostics

Running in fp16 on a single RTX 4090, PoseQualityNetKP processes **7.5 k frames** $\mathbf{s^{-1}}$—×250 real-time while containing only **3.41 M** trainable weights (Table 3), comfortably within mobile memory budgets. The architecture thus scales to multiple concurrent 30 Hz streams and remains suitable for edge deployment.

Take-away. Across all four primary metrics (Accuracy/F1 for two heads, MAE for regression, FPS/size for deployment) the proposed model meets or exceeds the targets stipulated in Section 4.3, validating the exercise-conditioned design and its suitability for real-time physiotherapy applications.

4.4.4 Qualitative results

To be written

4.5 Ablation study

To quantify how each architectural block contributes to the final performance, we trained four variants of PoseQualityNetKP:

- **A** – CNN feature extractor only
- **B** – CNN + exercise–ID embedding
- **C** – CNN + bidirectional LSTM (temporal encoder)
- **FULL** – CNN + Bi-LSTM + exercise–ID embedding

Table 6 reports the main metrics on the held-out test split (4 112 windows). All variants easily exceed real-time throughput (FPS ≫ 30), but differ markedly in accuracy, regression error, and model size.

Table 6. Ablation results on the REHAB24-6 test split. FPS measured with a 64-frame dummy clip on a single RTX 4090 (fp16).

Variant	Rep-Acc	Rep-F1	Ex-Acc	Ex-F1	MAE (°)	FPS	Params (M)
A (CNN)	0.797	0.797	0.994	0.993	8.49	9 160	0.25
B (CNN + Emb)	0.821	0.819	0.994	0.994	6.18	9 126	0.25
C (CNN + Bi-LSTM)	0.853	0.853	0.996	0.996	6.12	7 256	3.40
FULL	**0.903**	**0.903**	**0.997**	**0.997**	**3.86**	7 383	3.41

4.5.1 Discussion.

(a) **Exercise embedding (*A* → *B*).** Conditioning the quality and error heads on a one-hot exercise context yields an immediate boost in repetition accuracy

(+2.4 pp) and cuts the MAE by $\approx$ 27% at *no* parameter cost, confirming that task-specific priors simplify the decision boundary.

(b) **Temporal encoder (*A* $\rightarrow$ *C*).** Replacing frame-wise pooling with a Bi-LSTM improves all four recognition metrics, most notably repetition quality (+5.6 pp).

(c) **Synergy (*C* $\rightarrow$ *FULL*).** Combining both blocks is *additive*: the MAE is nearly halved relative to the CNN baseline (–4.6°), while repetition accuracy gains a further +5 pp. The parameter increase from 0.25 M to 3.4 M is marginal for modern GPUs and remains well within mobile budgets.

Overall, the results validate the design choice of coupling a light-weight temporal encoder with an exercise-aware context vector: together they deliver the largest accuracy gains and the lowest joint-angle error while maintaining real-time speed.

4.6 Discussion and limitations

Model strengths. PoseQualityNet-KP achieves convincing performance with a footprint of only 3.4 M weights. Its multi-head design yields three practical advantages: (i) joint-level feedback instead of a binary verdict, (ii) automatic suppression of advice when the wrong drill is performed, and (iii) edge-ready inference ($\sim$30 fps on CPU, 7.5 k windows s^{-1} in isolation).

Residual failure modes. Qualitative inspection reveals four recurring sources of error:

(a) *Self-occlusion.* Exercises that hide an entire limb (e.g. seated squats with arms on thighs) occasionally produce invalid landmark estimates, which in turn mis-trigger the quality classifier.

(b) *Camera pose.* MediaPipe accuracy degrades once the sensor is pitched by more than $\pm 20°$ or rolled non-horizontally, leading to spurious angle errors.

(c) *Small extremities.* Ankle landmarks occupy only a few pixels in a 640$\times$480 recording; the regressor therefore shows the highest MAE at those joints (5.9°).

(d) *Limited demographics.* REHAB24-6 contains ten healthy young adults; performance on elderly or post-operative patients is untested. Domain shift due to loose clothing, larger body mass, or assistive devices remains an open question.

Mitigation strategies. Multi-view capture or synthetic occlusion augmentation could harden the network against self-occlusion. A fast entropy filter on the landmark heatmaps would allow dynamic confidence weighting when the camera is improperly oriented. Finally, collecting a more diverse cohort and fine-tuning the ankle channels with higher-resolution crops are expected to close the remaining accuracy gap.

5 Conclusions and future work

We introduced PoseQualityNet-KP, a 3-head CNN–BiLSTM that turns a single web-camera stream into joint-specific rehabilitation feedback. Trained on the cleaned REHAB24-6 corpus, the model reaches **91.5 %** repetition-quality accuracy ($F_1 = 0.915$), **99.5 %** F_1 for exercise recognition, and a **4.73**° global MAE across 14 joint angles while running fully on-device at real-time speed. An ablation study confirms that both the Bi-LSTM temporal encoder and the exercise-ID embedding are essential: together they halve the angular error and raise quality F_1 by +10 pp with only a minor increase in parameters.

Next steps will address three axes:

1. *Robustness.* Integrate a lightweight self-supervised pre-text task to improve landmark reliability under occlusion and extreme camera angles, and explore multi-view fusion when a second device is available.
2. *Personalisation.* Add a 30-s calibration routine that learns user-specific joint-angle baselines and dynamically tightens tolerances as rehabilitation progresses.
3. *Deployment.* Convert the network to 4-bit QAT ONNX, bundle it inside a cross-platform mobile SDK, and conduct a six-week field study with post-operative patients to quantify adherence and recovery gains versus usual care.

Taken together, these extensions aim to transform the current prototype into a clinically validated, low-cost companion for at-home physiotherapy.

6 Author contributions

The sole author, **Jithin Krishnan**, conceived the project idea, designed the methodology, implemented the system (including data processing, model development, and front-/back-end integration), performed all experiments, analysed the results, and wrote the manuscript.

7 References

[1] Klaus Widhalm, Lukas Maul, Sebastian Durstberger, Peter Putz, Sebastian Leder-Berg, Hans

Kainz, and Peter Augat, "Efficacy of Real-Time Feedback Exercise Therapy in Patients Following Total Hip Arthroplasty: Protocol for a Pilot Cluster-Randomized Controlled Trial," *JMIR Research Protocols*, vol. 13, pp. e59755, 2024.

[2] Feng Xing, Juan Liu, Chang Mei, Jingnan Chen, Yi Wen, Jianrong Zhou, and Shiqi Xie, "Adherence to rehabilitation exercise and influencing factors among people with acute stroke: a cross-sectional study," *Frontiers in Neurology*, vol. 16, pp. 1554949, 2025.

[3] Marcela Velez, Luz Helena Lugo-Agudelo, Daniel F. Patiño Lugo, Claire Glenton, Ana M. Posada, Luisa F. Mesa Franco, and *et al.*, "Factors that influence the provision of home-based rehabilitation services for people needing rehabilitation: a qualitative evidence synthesis," *Cochrane Database of Systematic Reviews*, vol. 2023, no. 2, pp. CD014823, 2023.

[4] Ali Abedi, Tracey J. F. Colella, Maureen Pakosh, and Shehroz S. Khan, "Artificial intelligence-driven virtual rehabilitation for people living in the community: A scoping review," *npj Digital Medicine*, vol. 7, no. 1, pp. 25, 2024.

[5] B. Lugaresi et al., "Mediapipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[6] A. Tharatipyakul, T. Srikaewsiew, and S. Pongnumkul, "Deep learning–based human body pose estimation in providing feedback for physical movement: A review," *Heliyon*, vol. 10, no. 17, pp. e36589, Sept. 2024.

[7] Hitesh Kotte, Miloš Kravčík, and Nghia Duong-Trung, "Real-time posture correction in gym exercises: A computer vision-based approach for performance analysis, error classification and feedback," in *Proc. 3rd Int. Workshop on Multimodal Immersive Learning Systems (MILeS '23) co-located with EC-TEL 2023*, 2023, vol. 3499 of *CEUR Workshop Proceedings*, pp. 64–70.

[8] J. A. Francisco and P. S. Rodrigues, "Computer vision based on a modular neural network for automatic assessment of physical therapy rehabilitation activities," *IEEE Access*, vol. 10, pp. 121225–121236, 2022.

[9] Andrej Černek, Jan Sedmidubsky, Petra Budikova, Miriama Jánošová, Lukáš Katzer, and Michal Procházka, "Rehab24-6: Physical therapy dataset for analyzing pose estimation methods," in *Proc. 17th Int. Conf. on Similarity Search and Applications (SISAP)*, 2024, pp. 1–14.

[10] Julie L. McGinley, Richard Baker, Robin Wolfe, and Meg E. Morris, "The Reliability of Three-Dimensional Kinematic Gait Measurements: A Systematic Review," *Gait & Posture*, vol. 29, no. 3, pp. 360–369, 2009.

[11] Guangjun Hu, Weiqiang Wang, Bowen Chen, Hongliang Zhi, Yi Li, Yanhong Shen, and Kun Wang, "Concurrent Validity of Evaluating Knee Kinematics Using Kinect System During Rehabilitation Exercise," *Medicine in Novel Technology and Devices*, vol. 11, pp. 100068, 2021.