Personalized Monitoring in Home Healthcare: An Assistive System for Post Hip Replacement Rehabilitation

Alaa Kryeem
University of Haifa
akryeem@gmail.com

Shmuel Raz
University of Haifa
razshmu@gmail.com

Dana Eluz
Galilee Medical Center
DanaE@gmc.gov.il

Dorit Itah
Galilee Medical Center
DoritI@gmc.gov.il

Hagit Hel-Or
University of Haifa
hagit@cs.haifa.ac.il

Ilan Shimshoni
University of Haifa
ishimshoni@is.haifa.ac.il

Abstract

The rehabilitation process for hip replacement surgery relies on supervised exercises recommended by medical authorities. However, limitations in therapist availability, budget constraints, and evaluation inconsistencies have prompted the need for a more accessible and user-friendly solution. In this paper, we propose a scalable, user-friendly, and cost-effective vision-based human action recognition system utilizing machine learning (ML) and 2D cameras. By providing personalized monitoring, our solution aims to address the limitations of traditional rehabilitation methods and support productive home-based healthcare. A key component of our work involves the use of deep learning (DL) method to align time-series exercise data, which ensures accurate analysis and assessment. Additionally, we introduce the concept of a Golden Feature, which plays a critical role in the framework by providing valuable insights into exercise execution and contributing to overall system accuracy. Furthermore, our framework goes beyond predicting exercise scores and focuses on predicting comments for partially successful cases using a multi-label ML model. This allows for a deeper understanding of the clinical reasons behind partial success, such as the patient's physical condition and their execution of the exercise. By identifying and analyzing these factors, our framework provides meaningful feedback and guidance to support effective rehabilitation. When evaluated on multiple exercises, the system achieved an accuracy level of 80% or higher on predicting execution score, and 72% on predicting the execution feedback.

1. Introduction

Human action recognition (HAR) in the healthcare sector, and particularly in diagnostics and rehabilitation of patients has gained significant attention in recent years due to the potential to facilitate more effective and efficient rehabilitation interventions [7, 12, 17, 29, 31]. The ability to accurately identify and assess performance of a patient's physical motion can provide valuable insights into their recovery progress, which in turn can guide personalized treatment plans. Furthermore automation of HAR can promote remote and home based medical treatment. The use of HAR in rehabilitation in remote settings, can improve patient engagement, as well as provide clinicians with objective measures of progress.

In this study, we focus on rehabilitation following hip replacement surgery, but the approach easily generalizes to any rehabilitation program. Currently, rehabilitation following hip replacement surgery involves a set of exercises of increasing difficulty that are supervised and tuned to the patient's abilities by a medical professional (occupational therapist, physiotherapist etc). The rehabilitation program is often initiated at the hospital following surgery and is continued at a rehabilitation facility, hospital day visits, or the patient's home. Rehabilitation in the hospital is costly, leaning heavily on the limited medical professional's resources, and thus patients are typically discharged soon after surgery. Unfortunately, rehabilitation in the home setting is mostly unsupervised with intermittent monitoring and guidance by the medical professional. This results in imprecise assessment of the patient's' performance and results in a less efficient and less adaptive treatment. These limitations underscore the need for more effective and efficient methods of rehabilitation in a home setting, which automatically assess performance, provide feedback, and adapt the program to the patient's performance. Such a system has the potential to improve patient recovery and reduce costs while minimizing the need for professional resources.

In this paper we introduce a novel framework that utilizes 2D cameras and machine learning (ML), to automatically assess the performance of rehabilitation exercises following hip replacement surgery, enabling accessible home-based rehabilitation. Our framework incorporates an innovative deep learning (DL) method for aligning time-series data, ensuring accurate synchronization of exercise signals. Additionally, we introduce the concept of a Golden Feature, which aids in signal alignment, iteration splitting, and similarity comparison between test and aligned training data. By leveraging these techniques, our framework identifies key intervals within the exercise routine, allowing for a more precise and effective assessment of exercise performance. Furthermore, in addition to performance scores, our framework provides feedback regarding specific aspects of the execution that may require improvement, offering valuable insights to enhance the rehabilitation process.

2. Related work

Automated rehabilitation systems have gained significant attention in recent years due to their potential to improve the effectiveness and accessibility of rehabilitation programs. Tracking the activity of a patient can be accomplished through the utilization of cameras or wearable body sensors [1]. Several reviews and surveys have covered the diverse methodologies employed in HAR [16]. These approaches can be categorized into founr primary categories: radio frequency-based [34], sensor-based [33], wearable device-based [5], and vision-based [14] methods.

In the domain of vision-based HAR, a notable work [20] proposes an integrated framework for fine-grained human action quality assessment, combining category classification and regression-based evaluation running on RGB videos. Local motion patterns of body joint-based feature representation are extracted for action classification, and a class-specific learning algorithm is used for evaluation. However, this approach suffers from several challenges: firstly, the segmentation of long videos is considered without extensive research on segment synchronization. The method is more suitable for assessing well-segmented action instances. Secondly, when complex activities are present in videos with long durations, the quality score is significantly affected. Future studies are needed to address semantic segmentation and alignment methods to enhance the practical application of the proposed framework. An exercise recognition method based on RGB-D human skeleton models using an Asus Xtion camera was proposed in [18]. Several works in the field of HAR have utilized the Microsoft Kinect RGB-D camera, which combines an RGB camera with an infrared (IR) camera to develop HAR systems. For instance, [13] proposed A-MAL, an automatic motion evaluation learning algorithm, which leveraged Kinect's capabilities to learn from correctly performed motion 3D videos. Similarly, [24] utilized the Kinect camera to automate the process of predicting fall probability in older adults using the BBS fall assessment. Researchers in [10] combined Kinect with three machine learning algorithms to classify actions. They employed Kinect to capture body joint data, which was then processed using techniques such as K-means clustering to identify relevant joints involved in each activity. Subsequently, a multi-class support vector machine was employed to validate the obtained postures, and a discrete hidden Markov model was used to model each activity as a sequence of known postures. Hybrid DL model was also developed in [19] by incorporating CNN and Long Short-Term Memory (LSTM) for activity recognition where CNN is used for spatial features extraction and LSTM network is utilized for learning temporal information, also in this work Kinect was used to extract human body joints. In the domain of smart sensor-based systems for HAR, the authors of [36] developed the Smart Sensor-based Rehabilitation Exercise Recognition (SSRER) system, which utilizes a hybrid CNN of Sensor-CNN (S-CNN) and dynamic platform (D-CNN). Furthermore, a sensory-based DL framework was proposed in [21] for assessing physical rehabilitation exercises. However, this research encountered certain limitations that can be categorized into two main sets: dataset limitations and equipment limitations. In the field of sports activities, [37] demonstrated an application of human motion recognition utilizing DL and smart wearable devices. Additionally, other studies, such as [2], addressed the same problem using smartwatches. Other methods such as in [27] used virtual reality to enhance the rehabilitation experience for individuals with Tetraplegia in home settings, while allowing supervision by clinical therapists.

The current literature on vision-based HAR systems faces a common obstacle, which is the limited availability and usability of the technology. In previous works, joint detection has typically been accomplished using 3D cameras, which offer the advantage of depth tracking and increased accuracy, but are less accessible and more complicated to set up than 2D cameras. By contrast, 2D cameras, such as those found on smartphones or webcams, are widely available and easy to use, making them a more user-friendly and cost-effective option for tracking. Additionally, 2D-based systems have the advantage of being able to analyze any video recorded with a camera, while 3D-based systems are often device-dependent (e.g. Intel RealSense or Microsoft Kinect), limiting their versatility. In a comparative analysis made in [23], evaluating the differences between 2D and 3D cameras for HAR, it was demonstrated that extracting joint positions from standard resolution 2D videos can provide equally informative features as those obtained from depth cameras like Kinect. The study showed that 2D joint positions can be as informative as their 3D counterparts, challenging the notion that depth information is essential for accurate joint extraction and activity recognition. Our proposed solution will leverage the simplicity and accessibility of 2D camera technology to provide a highly available, low-cost, and user-friendly solution for tracking patient activity.

3. Dataset

For this study, a dataset was collected at the occupational therapy unit at a major medical center consisting of 35 patients recovering from hip replacement surgery and 7 healthy individuals. The ages of the participants varied from 50 to 85 years, with the average age being 68.6 years. Out of these, there were 30 females and 12 males. Each participant performed a total of 24 exercises, with each exercise consisting of three or more repetitions of a well-defined task. Participants' performance on the exercises was assessed by occupational therapists who are co-authors of this paper. Performance scores were on a three-level scoring system: 0 - an unsuccessful attempt to perform the exercise, 1 - partial success, and 2 - successful execution. When patients receive a partial success score, additional comments were also recorded (e.g. compensatory movement, pain, fatigue etc). All the executions were recorded using two cameras: a frontal view camera and a side view camera. The scores assigned by the occupational therapists served as labels for training the ML models. The dataset focused on two daily functional tasks: (1) moving to the bed and lying in it, and (2) wearing pants. Each task contains several sub-tasks that mimic real-life activities. For instance, task (1) involves raising the legs onto the bed to assume a lying position and rolling from lying on the side to lying on the back. For task (2) the sub-tasks included, bending towards the floor, raising the leg in the air, and threading a ring onto the leg. By incorporating these functional tasks and their corresponding sub-tasks, the dataset captures essential movements and actions that individuals encounter in their daily lives. The inclusion of such tasks ensures the relevance and practicality of the exercises, enabling the proposed framework to provide targeted guidance and assessment for activities that directly impact daily functioning. [1]

4. An Automated System for monitoring Post Hip Replacement Rehabilitation Exercises

We propose an automated system that utilizes ML techniques and leverages the extraction of body joints from 2D videos. The system employs features derived from the extracted joints to assess the performance of rehabilitation exercises, predict performance scores as well as providing feedback when performance receives a score of 1. The major components of the system include (see Figure 1):

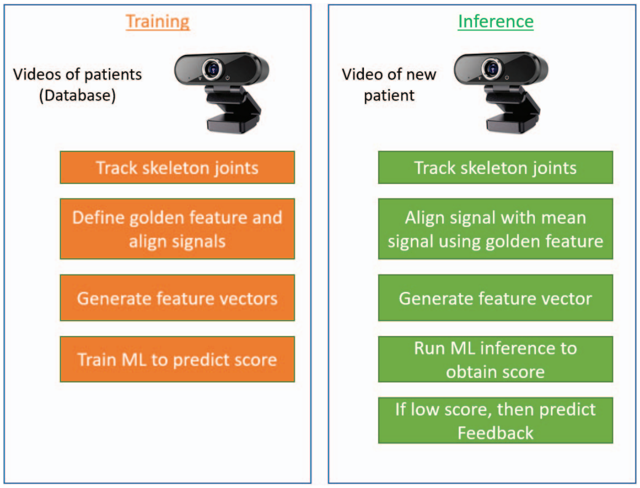

Training
Videos of patients (Database)
Track skeleton joints
Define golden feature and align signals
Generate feature vectors
Train ML to predict score
Inference
Video of new patient
Track skeleton joints
Align signal with mean signal using golden feature
Generate feature vector
Run ML inference to obtain score
If low score, then predict Feedback

Figure 1: Overview of the proposed framework.

- Alignment of the exercise signals
- Feature extraction
- Training a ML model to predict performance scores of exercises
- Training a ML model to predict the feedback when exercise performance is poor (score of 1).

4.1. Signal Alignment

An important aspect of exercise assessments is the variability in performance (even when correctly performed) between individuals. We address this challenge by aligning the exercise signals across subjects. Several approaches have been proposed for alignment of time-series signals including various Time-warping methods such as DTW [26], DBA [28], and SoftDTW [6]. While these methods have been widely used, they exhibit certain limitations such as difficulties in generalization, handling multi-classes, and computational complexity. To overcome these limitations, we use Diffeomorphic Temporal Alignment Nets (DTAN) [30] that can effectively handle the non-linearities and variability present in signals extracted from exercises performed by subjects. Additionally, DTAN overcomes the challenges of generalization, making it suitable for aligning signals from diverse exercise tasks and individuals. To validate the superiority of DTAN, we conducted a comparison with DTW [11] on our data. Results show that DTAN outperforms DTW in terms of alignment accuracy and robustness. Specifically, the mean standard deviation $\mathrm{std}(\mathrm{V}_{\mathrm{mean}})$ using DTW was 8.31 whereas using DTAN it reduced to 6.12 (see Figure 2).

[1] This study complies with the Declaration of Helsinki under the hospital's ethics committee approval (0148-20-NHR).

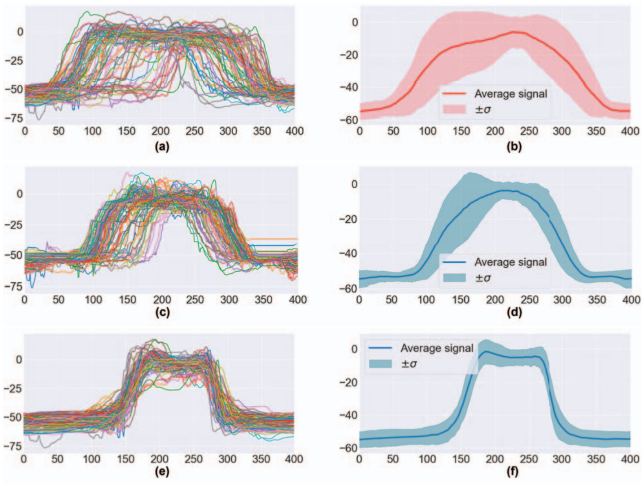


Figure 2: Example of signal alignment: (a) 81 misaligned signals. (b) Average misaligned signal. (c) Signals following DTW alignment. (d) Average signal following DTW alignment. (e) Signals following DTAN alignment. (f) Average signal following DTAN alignment. Shaded areas in b,d,f indicate to ± 1 STD at each time point.

DTAN is a DL based system that aligns time-series ensembles in an input-dependent manner (see Figure 2f). A nonlinear misaligned ensemble can be described as:

$$U_i = (V_i \circ W_i), \tag{1}$$

where $U_i, i = 1 \ldots N$ represents a set of N time-series signals . V_i represents the i^{th} aligned signal and W_i represents the i^{th} warp. Thus, W_i is the warping that when applied to V_i yields the misaligned signal U_i. Given an ensemble of input signals U_i, DTAN computes the mean aligned signal V_{mean} defined as the signal that minimizes:

$$\sum_{i=1}^{N} \|V_i - V_{mean}\|_2^2. \tag{2}$$

DTAN also provides the std of V_{mean} at each sample point:

$$\mathrm{std(V_{mean})} = \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{N}} (\mathrm{V_i} - \mathrm{V_{mean}})^2}. \tag{3}$$

Additionally, DTAN also outputs the warpings W_i which are Continuous Piecewise-Affine based (CPAB) diffeomorphisms as proposed in [8, 9], and are thus invertable continuous piecewise-affine transformations that are expressive and efficient. During inference, given a new signal U and the mean signal V_{mean}, DTAN aligns U by minimizing the following over W:

$$\|V - V_{mean}\|_2^2 \ \text{where} \ U = (V \circ W). \tag{4}$$

DTAN outputs the aligned signal V, as well as the transformation W associated with U.

In our study, we use DTAN to align exercise signals. As described below (Section 4.2), a specific time varying feature termed the *Golden Feature*, is chosen as the exercise signal. This feature will exhibit significant changes over time while maintaining a relatively smooth and accurate trend, free from noise. During training, correctly performed exercise signals are aligned forming a single representative signal V_{mean}. At test time the system attempts to align a new exercise signal U with the representative signal. The alignment using DTAN is as follows:

Pre-processing of the data for signal alignment:

1. Correctly performed exercise samples are selected from the database (received a score of 2 by the occupational therapist).

2. Since the exercise includes several repetitions, and the alignment is applied to a single repetition of the golden feature, the signal is split into repetitions. Smoothing of the signal is applied followed by peak detection, to identify the repetitions. Figure 3 shows the steps involved in the splitting process.

3. To normalize the varying number of frames across different subjects and different repetitions, the signal is interpolated to equate the number of frames.

Alignment iterations:

4. DTAN is trained on the pre-processed golden feature signals (Figure 2a). The "mean" signal representing the exercise is obtained (Figure 2f).

5. Remove samples that, after warping, deviate significantly from the mean aligned signal, and retrain DTAN on the remaining samples. (Figure 2e). This improves the DTAN alignment and refines the accuracy of the mean signal. Compare Figure 2b and 2f.

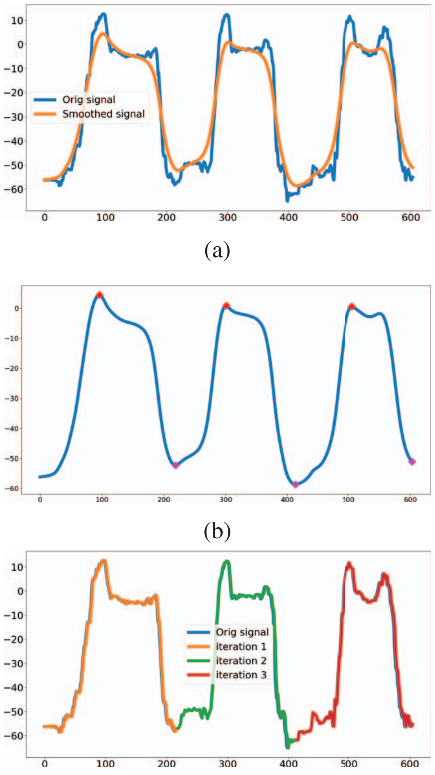
4.2. Feature Extraction

In order to train a ML model to predict the performance score of an exercise, a vector of features must be generated per sample. The features include numerous spatio-temporal characteristics of the motion performed by the subject. Subjects were tracked in the video streams using DL based trackers (e.g. MediaPipe (MP) [22]). and the spatio-temporal features were extracted from the tracked skeletal joints. Examples of relevant features may include average speed of joint motion, maximal height from floor, minimal joint angle etc. (see Figure 4).

Additionally, a single feature termed the *Golden Feature*, is chosen for each exercise, on which the alignment process is implemented. The Golden Feature is a characteristic time-dependent feature of the exercise that defines the time course of the exercise across subjects. Examples of golden features are the distance between the ankle and the bed, or the distance between the hand and the ankle (see example in Figure 4). The time-dependent golden feature serves as the exercise signal U_i, in the alignment process (Section 4.1).

Mapping Time Interval Features

In selecting the feature vector for training and testing the ML models, certain features may only be relevant during specific segments of the movement rather than over the entire exercise duration. For example, knee angle is only relevant when the patient's leg reaches the height of the bed, speed of motion is only relevant in the forward stretch of the arms and not on the return. Determining specific segments in each individual's motion path is very challenging due to the variability between subjects. To address this, we rely on the aligned DTAN mean signal, of the exercise: Important time-points can be marked in the exercise's mean signal V, representing relevant time-intervals. For example the time points when the ankle begins ascending from the floor (interval starting point) and when the ankle reaches the bed (interval end point). (red dots on the black plot in Figure 5b). These time points, marked once in the mean-signal V, are projected onto each aligned exercise signal V_i (green plot in Figure 5b) and then back projected to the original exercise signal U_i using the inverse of the warping transform W_i associated with the signal (blue plot in Figure 5b) (See also Section 4.1). Thus, the required time intervals are determined on the original signals and time dependent features can be extracted in the original signal's time units.



(a)

(b)

(c)

Figure 3: Splitting the exercise signal into separate iterations. (a) Smooth the given signal. (b) Identify local min and local max points. (c) Split the full signal into iterations based on the min/max from (b).

4.3. ML Model for Exercise Score Prediction

We trained ML models to assess the exercise performance and predict the score as assigned by the occupational therapist. We used Random forest classifiers [15] as its use of bootstrapping enables working on small datasets. Additionally, the random forest classifier allows feature ranking [3, 4] in which the predictive power of features can be assessed and used to tune the model. In all models, the number of trees was set to 100, and Gini Impurity measure [25] was used as splitting criterion, with minimum sample number of 2. Due to the unbalanced data (fewer low-scoring samples), balanced class weights were used. Model testing was performed on the patient dataset using a Leave One Out (LOO) approach [25].

4.4. ML Model for Feedback Prediction

When subjects are awarded a score of 1 on an exercise, informative feedback is also provided as one or more comments from a predefined list of possible feedbacks (for example: compensatory movement, pain, fatigue etc).

To automatically provide performance feedback for exercises receiving a score of 1, we employ a multi-label ML classification model [32, 35]. A separate Random Forest classifier is trained to predict each possible feedback response from the subjects spatio-temporal exercise signal

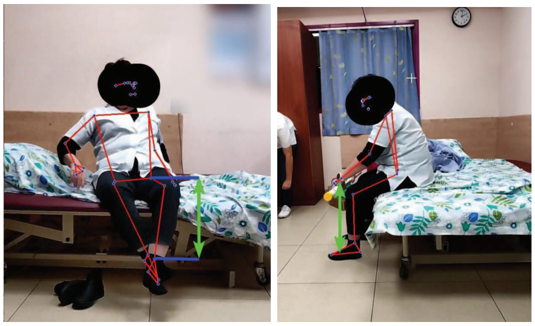


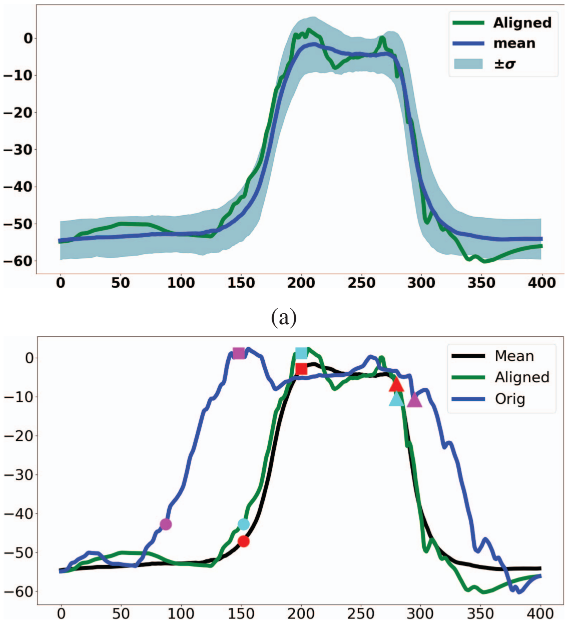
Figure 4: Example of exercise features. Left: Vertical distance of ankle joint from hip joint in the exercise requiring raising leg to bed. Right: Height of the hand from floor (ankle joint) in the exercise requiring bending to floor (illustrated by the therapists involved in this study).

features (see Section 4.2). Model hyper-parameters are set as given above. To combine the independent feedback predictions into a combination of feedbacks associated with the subject's performance, prior probabilities on the distribution of combinations of feedbacks per subject is computed from the database. Let f_i, $i = 1 \ldots k$ be the set of possible feedbacks for a specific exercise. For every combination of possible feedbacks $F_j = \{f_{j1}, f_{j2} \ldots, f_{jk}\}$, the prior probability $P_{prior}(F_j)$ is computed (Table 2). The posterior probability on the combination of feedbacks associated with a subject's performance combines the prior information and the model's predicted probabilities:

Let $X = (x_1, x_2 \ldots, x_n)$ be the feature vector of an incoming sample. The inference process determines the feedback combination $\hat{F_j} = \{\hat{f_{j1}}, \hat{f_{j2}} \ldots, \hat{f_{jk}}\}$ for the sample as:

$$\hat{F_j} = \text{argmax}_{F_j}\{(P_{prior}(F_j) + \epsilon) \times P_{model}(F_j \mid X)\}, \tag{5}$$

where $P_{prior}(F_j)$ is the prior probability, $P_{model}(F_j \mid X)$ is the conditional probability calculated as the product of the probabilities estimated by each classifier, and $\epsilon \in \mathbb{R}$ is the Prior Adjustment Factor. Due to the small sample size, a Prior Adjustment Factor is introduced and it's intended to ensure that all possible feedback combinations are considered, even in low or 0 prior probability.



(a)

(b)

Figure 5: Mapping time-intervals (Raising leg task). (a) Mean exercise signal V computed using DTAN (blue) and an aligned incoming exercise signal V_i (green). Std of the mean signal is shown shaded. (b) Red symbols indicate marked time points on the mean signal. These are mapped onto the aligned incoming signal V_i - cyan symbols. The mapped points are then back-projected to the original incoming signal U_i using the inverse of the warping W_i associated with U_i - pink symbols.

5. Experimental Results

5.1. Predicting Exercise Scores

The proposed framework was tested on multiple exercises using feature vectors derived based on insights from therapists and close monitoring of the exercises, and achieved an accuracy level of 80% or higher on predicting execution score (Figure 6). Table 1 showcases the evaluation metrics of our framework on two specific exercises, including accuracy, precision, recall, and F1 macro scores.

Raising leg onto the bed This task involves various exercises of raising a leg onto the bed: using a canvas belt, with hand assistance, and with assistance from the healthy leg. These three variations share the same goal and evaluation criteria, thus they were combined into a single dataset. This task was performed by 42 participants who completed a total of 115 exercises, of which 20 were unsuccessful (score=0), 49 were partially successful (score = 1), and 46 were performed successfully (score = 2).

The golden feature was defined as the distance between the ankle and the bed as shown in Figure 4 (left).

During each execution of the exercise, which consisted of multiple iterations, several spatio-temporal features were calculated for each iteration:

1. Ankle speed during leg-raising interval
2. Minimum height of ankle
3. Maximum height of ankle
4. Mean knee angle
5. Duration (%) knee angle exceeded 140° (indicating a straight leg) during leg-on-bed interval
6. Mean ankle depth
7. Duration of leg raise
8. Mean distance between left shoulder and left hip along the Y-axis
9. Mean distance between right shoulder and right hip along the Y-axis
10. Distance of the golden feature signal from the mean signal calculated as the mean square error divided by the std at each sample point of the sequences.

These features were taken for the best iteration (that with minimal distance to the mean signal - Section 4.1). Two additional features were computed by taking the mean of features 1 and 10 across all iterations. These two additional

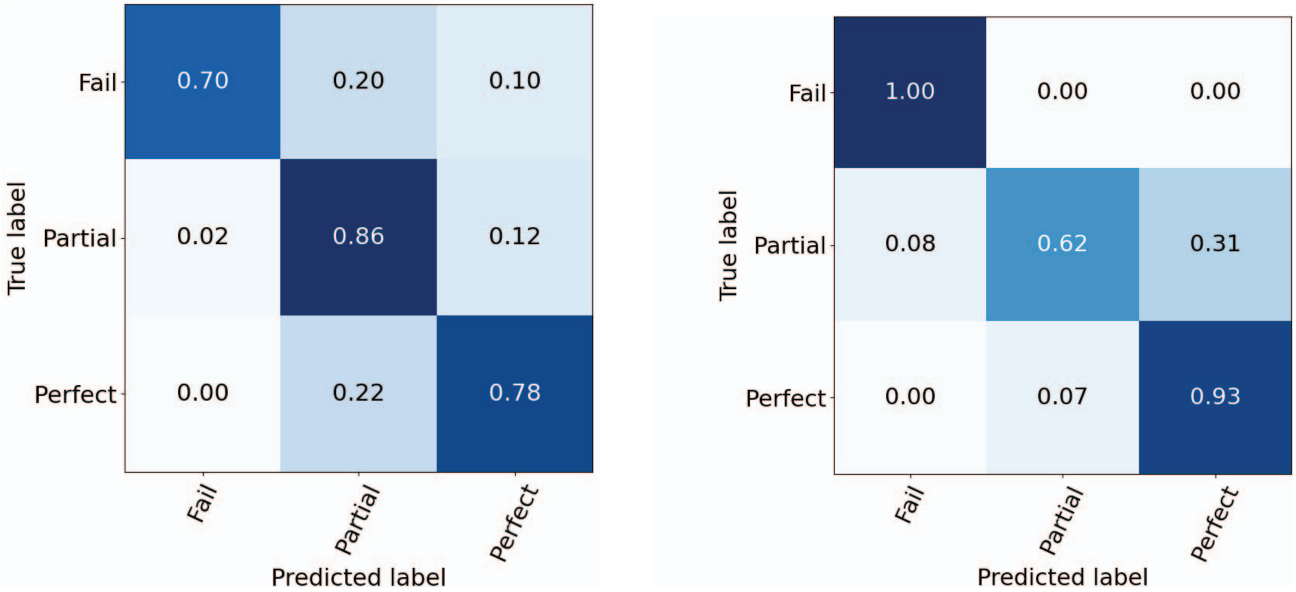


Figure 6: Confusion matrices. Left: Results for raising leg onto bed. Right: Results for bending down to floor exercise.

features (number 11 and 12) aim to capture the overall characteristics of the exercise execution, rather than focusing on a single iteration.

The model was evaluated using LOO. Figure 6(left) shows the resulting confusion matrix. 80% accuracy was obtained with the majority of mis-predictions occurring between scores 1 and 2 (partial success and success). It is worth noting that even therapists often disagree on the score of a specific imperfect execution.

Performing feature ranking revealed that the following five features contributed the most to the classifier's output (in decreasing order): 10, 3, 11, 6, 2. Note that the comparison to the mean signal is the most predictive feature.

Table 1: Evaluation metrics for the predicting exercise scores model. Accuracy calculation incorporates the imbalanced class sizes.

Run	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
Raising leg	0.8	0.83	0.78	0.8
Bending	0.85	0.85	0.85	0.84

Bending down to the floor This task involves bending down to the floor, in two variations: from a seated position on the bed and while holding a heavy stick with both hands. The two variations were combined into a single dataset. This task was performed by 27 participants who completed a total of 48 exercises, of which 8 were unsuccessful (score=0), 13 were partially successful (score = 1), and 27 were performed successfully (score = 2). The number of participants differed from the previous exercises, as not all patients were permitted to perform the bending down exercise due to associated risks.

The golden feature was defined as the distance between the hand and the ankle, captured by a side-facing camera as shown in Figure 7.

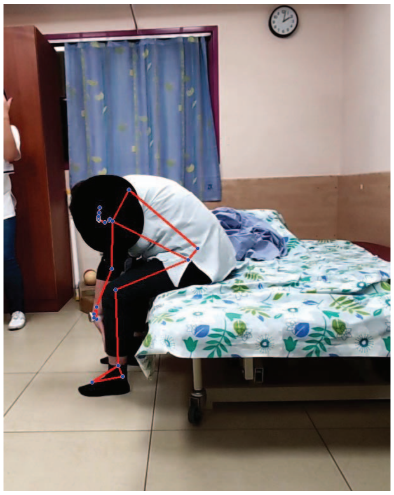


Figure 7: Side-facing camera: Bending down to the floor (illustrated by the therapists involved in this study).

For feature extraction, we defined two time intervals within this exercise: Interval 1 where hand descends to floor (segment marked by circle and square in Figure 8) and Interval 2 between reaching the ground and beginning to ascend (marked by square and triangle in Figure 8). The following

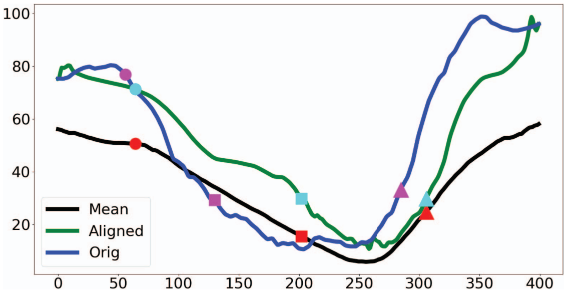


Figure 8: Time intervals for bending down to the floor task.

features were calculated for each iteration of the exercise:

1. Hand speed during Interval 1
2. Minimum height of hand from floor
3. Maximum height of hand from floor
4. Mean hip angle in Interval 2
5. Duration (%) hip angle less than 45° in Interval 2
6. Mean distance between hand and ankle in Interval 1
7. Duration (%) distance between hand and ankle was less than 20cm in Interval 1
8. Distance of the golden feature signal from the mean signal (calculated as the mean square error divided by the std at each sample point of the sequences)

These features were taken for the best iteration (that with minimal distance to the mean signal). Additional features were computed by considering all iterations:

9. Mean of feature 1 across all iterations
10. Minimum of feature 4 across all iterations
11. Maximum of feature 5 across all iterations
12. Minimum of feature 6 across all iterations
13. Maximum of feature 7 across all iterations
14. Mean of feature 8 across all iterations

The model was evaluated using LOO. Figure 6(right) shows the resulting confusion matrix. Accuracy of 85% was obtained.

Performing feature ranking revealed that the following features were most predictive: 4, 5, 12, 8, 14.

5.2. Predicting Feedback for low scoring tasks

We trained a multi-label ML model to predict the feedback comments received when exercise execution was only partially successful (score=1) (see Section 4.4). The leg-raising exercise serves as an example. The feedback in this task includes: partial movement, problem with knee angle, non-smooth movement, and back compensation.

The dataset consists of 35 patients who received a score of 1, excluding patients who verbally reported pain or physical difficulties. The distribution of combinations of feedback comments are presented in Table 2.

Table 2: Labels distribution - Prior Probabilities

Partial Movement	Problem with Knee Angle	Non-Smooth Movement	Back Compensation	Probability
0	0	0	1	$\frac{12}{35}$
1	0	0	0	$\frac{6}{35}$
0	1	0	0	$\frac{5}{35}$
1	0	0	1	$\frac{4}{35}$
0	0	1	0	$\frac{4}{35}$
1	0	1	0	$\frac{2}{35}$
0	0	1	1	$\frac{1}{35}$
0	1	1	0	$\frac{1}{35}$

The sample features defined in Section 5.1 were used to train the multi-label model. The prediction were generated as described in Section 4.4, with $\epsilon = 0.3$. The model achieved 72% accuracy in predicting the specific combination of comments for partially-successful cases. (Note, chance level is 6.25%). If the prior-probability information is ignored and only the predictions of the independent models are considered, the resulting accuracy is only 68% compared to the improved accuracy of 72% when priors are considered.

6. Conclusion and future work

This paper proposed a framework for automatically evaluating rehabilitation exercises using machine learning and 2D cameras. The framework was tested on various exercises, including daily functional tasks, and achieved an accuracy level of 80% or higher. Features were selected based on occupational therapist feedback. Although the framework demonstrated promising results, improvement could be achieved with improved joint tracking.

The use of 2D cameras offers a cost-effective and user-friendly alternative to specialized equipment for exercise tracking. The proposed framework has the potential to enhance rehabilitation programs by providing consistent and accessible quantitative assessments of patient progress. Moreover, the framework has the potential to support effective home-based healthcare, allowing patients to perform exercises remotely while receiving personalized feedback and guidance.

References

[1] Jake K Aggarwal and Lu Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

[2] David M Burns, Nathan Leung, Michael Hardisty, Cari M Whyne, Patrick Henry, and Stewart McLachlin. Shoulder physiotherapy exercise recognition: machine learning the inertial signals from a smartwatch. *Physiological measurement*, 39(7):075007, 2018.

[3] Barak Chizi and Oded Maimon. Dimension reduction and feature selection. *Data mining and knowledge discovery handbook*, pages 83–100, 2010.

[4] Barak Chizi, Lior Rokach, and Oded Maimon. A survey of feature selection techniques. In *Encyclopedia of Data Warehousing and Mining, Second Edition*, pages 1888–1895. IGI Global, 2009.

[5] Maria Cornacchia, Koray Ozcan, Yu Zheng, and Senem Velipasalar. A survey on activity detection and classification using wearable sensors. *IEEE Sensors Journal*, 17(2):386–403, 2016.

[6] Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *International conference on machine learning*, pages 894–903. PMLR, 2017.

[7] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108:107561, 2020.

[8] Oren Freifeld, Soren Hauberg, Kayhan Batmanghelich, and John W Fisher. Highly-expressive spaces of well-behaved transformations: Keeping it simple. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2911–2919, 2015.

[9] Oren Freifeld, Søren Hauberg, Kayhan Batmanghelich, and Jonn W Fisher. Transformations based on continuous piecewise-affine velocity fields. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2496–2509, 2017.

[10] Salvatore Gaglio, Giuseppe Lo Re, and Marco Morana. Human activity recognition process using 3-d posture data. *IEEE Transactions on Human-Machine Systems*, 45(5):586–597, 2014.

[11] Toni Giorgino. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31:1–24, 2009.

[12] Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. A survey on deep learning for human activity recognition. *ACM Computing Surveys (CSUR)*, 54(8):1–34, 2021.

[13] Tal Hakim and Ilan Shimshoni. A-mal: Automatic motion assessment learning from properly performed motions in 3d skeleton videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[14] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

[15] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[16] Zawar Hussain, Michael Sheng, and Wei Emma Zhang. Different approaches for human activity recognition: A survey. *arXiv preprint arXiv:1906.05074*, 2019.

[17] Md Milon Islam, Sheikh Nooruddin, Fakhri Karray, and Ghulam Muhammad. Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects. *Computers in Biology and Medicine*, page 106060, 2022.

[18] Csaba Kertész. Physiotherapy exercises recognition based on rgb-d human skeleton models. In *2013 European Modelling Symposium*, pages 21–29. IEEE, 2013.

[19] Imran Ullah Khan, Sitara Afzal, and Jong Weon Lee. Human activity recognition via hybrid deep learning based model. *Sensors*, 22(1):323, 2022.

[20] Qing Lei, Hong-Bo Zhang, Ji-Xiang Du, Tsung-Chih Hsiao, and Chih-Cheng Chen. Learning effective skeletal representations on rgb video for fine-grained human action quality assessment. *Electronics*, 9(4):568, 2020.

[21] Yalin Liao, Aleksandar Vakanski, and Min Xian. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):468–477, 2020.

[22] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[23] Fiona Marshall, Shuai Zhang, and Bryan Scotney. Comparison of activity recognition using 2d and 3d skeletal joint data. In *Irish Machine Vision & Image Processing IMVIP 2019*, page 13. Irish Pattern Recognition and Classification Society, 2019.

[24] Alaa Masalha, Nadav Eichler, Shmuel Raz, Adi Toledano-Shubi, Daphna Niv, Ilan Shimshoni, and Hagit Hel-Or. Predicting fall probability based on a validated balance scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 302–303, 2020.

[25] Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

[26] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[27] Shanmugam Muruga Palaniappan, Shruthi Suresh, Jeffrey M Haddad, and Bradley S Duerstock. Adaptive virtual reality exergame for individualized rehabilitation for persons with spinal cord injury. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 518–535. Springer, 2020.

[28] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, 44(3):678–693, 2011.

[29] Fatemeh Serpush, Mohammad Bagher Menhaj, Behrooz Masoumi, and Babak Karasfi. Wearable sensor-based human

activity recognition in the smart healthcare system. *Computational intelligence and neuroscience*, 2022, 2022.

[30] Ron A Shapira Weber, Matan Eyal, Nicki Skafte, Oren Shriki, and Oren Freifeld. Diffeomorphic temporal alignment nets. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Abdulhamit Subasi, Kholoud Khateeb, Tayeb Brahimi, and Akila Sarirete. Human activity recognition using machine learning methods in a smart healthcare environment. In *Innovation in health informatics*, pages 123–144. Elsevier, 2020.

[32] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[33] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern recognition letters*, 119:3–11, 2019.

[34] Shuangquan Wang and Gang Zhou. A review on radio based activity recognition. *Digital Communications and Networks*, 1(1):20–29, 2015.

[35] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

[36] Wentong Zhang, Caixia Su, and Chuan He. Rehabilitation exercise recognition and evaluation based on smart sensors with deep learning framework. *IEEE Access*, 8:77561–77571, 2020.

[37] Xiaojun Zhang. Application of human motion recognition utilizing deep learning and smart wearable device in sports. *International Journal of System Assurance Engineering and Management*, 12(4):835–843, 2021.