Multimedia Tools and Applications
https://doi.org/10.1007/s11042-025-21050-3

Towards an RGB camera-based rehabilitation exercise assessment using an enhanced spatio-temporal transformer framework

Amanpreet Chander[1] · **Deepti R. Bathula**[2] · **Ashish Kumar Sahani**[1]

Received: 12 August 2024 / Revised: 4 February 2025 / Accepted: 11 July 2025

Abstract

Rehabilitation following surgery or trauma is crucial for patient recovery. However, many patients attempt to rehabilitate at home by performing exercises without professional supervision, potentially hindering their progress or causing complications. To address this, deep learning systems have been extensively explored. Traditional motion detection systems such as Vicon or Kinect, while effective, are costly or impractical for at-home use. Therefore, leveraging the ubiquity of mobile devices, we propose an affordable and robust home-based rehabilitation system that uses a simple RGB camera for recording physical exercises. We used Mediapipe, an open-source framework, to extract joint coordinate information from video frames. Our spatio-temporal transformer encoder model is enhanced with dual attention and weighted residual connections for assessing human motion quality using skeleton data. Our model has been evaluated on a custom RGB dataset, as well as two benchmark datasets—UI-PRMD and KIMORE, and achieves state-of-the-art performance in motion quality assessment, demonstrating an average reduction of 12.9% in mean absolute error (MAE) across all datasets. Our approach has the potential to be a cost-effective and reliable solution for home-based rehabilitation, offering advanced motion assessment capabilities without the need for expensive equipment.

Keywords Physical rehabilitation · Action quality assessment · Deep learning · Transformer encoder · Residual connection

✉ Amanpreet Chander
2018bmz0002@iitrpr.ac.in

Deepti R. Bathula
bathula@iitrpr.ac.in

Ashish Kumar Sahani
ashish.sahani@iitrpr.ac.in

[1] Department of Biomedical Engineering, Indian Institute of Technology, Ropar, Rupnagar 140001, Punjab, India

[2] Department of Computer Science and Engineering, Indian Institute of Technology, Ropar, Rupnagar 140001, Punjab, India

Published online: 06 August 2025

1 Introduction

Rehabilitation exercises are essential for individuals recovering from injuries, surgery, or suffering from conditions that affect mobility and strength for considerable duration [1, 2]. These exercises, crafted by healthcare professionals, aim to improve strength, flexibility, and functional abilities to support independence in daily life. Research highlights the importance of consistency, frequency, precision, and intensity of these exercises, as these factors strongly influence recovery outcomes. Although, ideally, rehabilitation must be performed in clinics under professional supervision, the high cost of healthcare and limited therapist availability often reduces the frequency of clinical sessions. Consequently, over 90% of rehabilitation programs, especially for chronic conditions, end up being performed by patients at home, that is, without supervision [3]. Lack of real-time supervision and feedback is a major barrier in home-based rehabilitation, as patients miss the immediate guidance and corrective feedback that therapists provide in clinical settings [4]. This lack of oversight can lead to uncertainty about exercise quality, reduced motivation, and a sense of isolation from the therapeutic process, which in turn can lead to extended recovery times, further complications, or additional healthcare costs [5–7]. Therefore, addressing these challenges has become a focal point in rehabilitation research, with new technologies emerging to bridge the gap between home-based exercise and professional supervision.

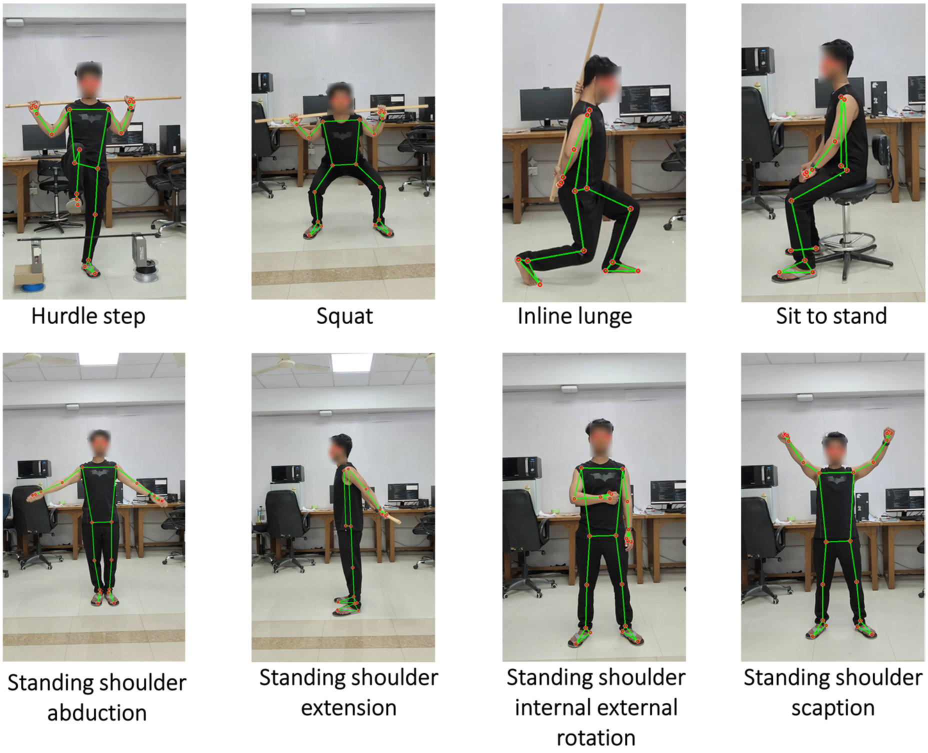
Action Quality Assessment (AQA) provides automatic evaluation of human actions and activities and is thus considered to have great potential in the field of rehabilitation medicine. AQA aims to provide an unbiased and consistent measure of action quality by eliminating subjective variability of human assessors, typically using automated systems. It has significant applications in various fields, including healthcare and rehabilitation [8–13], sports and athletic training [14–20], and education and skill training [21, 22]. Vision-based Action Quality Assessment systems capture body movement information using various motion detection technologies. Generally, these systems utilize an RGBD camera, such as the Kinect, or multiple cameras, like in the Vicon system, to track body motion [23]. The captured movement data is then used to assess motion quality. Several studies have approached quality assessment as a classification problem, determining whether movements are correct or incorrect [24, 25]. However, such systems cannot detect subtle variations in the way actions are performed, which limits their ability to provide useful feedback. Hence, researchers have developed regression-based evaluation systems capable of providing continuous scores to describe the quality of actions [26–28]. With advancements in pose estimation technology, skeleton data can now be obtained even from RGB videos. Several studies have used pose detection software to extract joint information [29, 30]. However, the low signal-to-noise ratio (SNR) in RGB-based camera systems reduces the accuracy of joint detection, making the development of a reliable rehabilitation system challenging.

In this study, we propose a cost-effective home-based rehabilitation system that leverages the ubiquity of a simple RGB camera (smartphones, tabs, webcams) for AQA to capture video. In addition, it utilizes the MediaPipe [31], an open-source framework that detects landmarks of human body to extract skeleton data. Subsequently, the skeleton-based pose estimation from video frames is used to train a deep learning model to assess the quality of actions performed by the patient. The key contributions of this study are as follows:

- We propose an affordable RGB camera-based home-rehabilitation system that uses deep learning to assess the quality of human body actions while performing different rehabilitation exercises.

- We provide a dataset on eight healthy individuals repeating eight different rehabilitation exercises (see Fig. 1) captured using a smartphone. Each exercise was performed both correctly and incorrectly to highlight the distinction between the two.
- We introduce a novel, transformer-based, encoder-only architecture with Dual Attention Block that extends the traditional spatio-temporal transformer architecture to process and analyze sequences of frames (multi-frame) in a sophisticated manner. It allows for more granular capture of short-range spatio-temporal dependencies in addition to long-range temporal dependencies.
- Additionally, we introduce a weighted-residual transformer block that incorporates adaptive skip connection with learnable weight across the sub-layers of the transformer block to enhance the model's ability to predict quality scores.
- We demonstrate the efficacy of the proposed model and compare it with state-of-the-art approaches using our custom RGB camera dataset. Furthermore, we establish the generalizability and robustness of our model using two benchmark datasets: UI-PRMD [32] and KIMORE [33].

The article is organized as follows: Section 2 presents an overview of related literature. Section 3 describes the data collection process, joint coordinate extraction from video frames, motion quality estimation, and the spatio-temporal transformer-based deep learning archi-



Hurdle step

Squat

Inline lunge

Sit to stand

Standing shoulder abduction

Standing shoulder extension

Standing shoulder internal external rotation

Standing shoulder scaption

Fig. 1 Sample video frames of a subject performing all eight exercises

tecture. Sections 4 and 5 provide details on the datasets, evaluation metrics, implementation, and model performance comparisons. Finally, section 6 titled "Discussion" analyzes the model's performance, identifies its limitations, and outlines future work directions.

2 Related works

This section comprehensively reviews previous studies on human motion quality assessment using various deep learning methods. Numerous studies have focused on evaluating human movement by predicting continuous scores. For instance, Liao et al. proposed a deep spatio-temporal neural network that combines CNN and LSTM layers to predict quality scores. In their study, they recorded their own data and used Gaussian Mixture Models (GMM) to generate ground-truth scores for each repetition of different rehabilitation exercises [34]. Deb et al. introduced a deep learning architecture that combines a graph convolution network (GCN) with LSTMs, effectively capturing spatio-temporal features to predict continuous scores [35]. Similarly, Yao et al. proposed a multi-task contrastive learning framework to capture subtle and critical differences in skeleton sequences, introducing a joint attention matrix designed to evaluate the significance of various joints in physical exercise, thus enabling feedback on motion errors [36].

Traditional systems such as Vicon and Kinect capture both positional and angular information, and some studies have leveraged both to assess performance. For example, Mourchid et al. introduced a Multi-Residual Spatio-Temporal Graph Network (MR-STGN) incorporating both angular and positional 3D skeleton data [37]. Sardari et al., aiming to balance computational complexity with scoring accuracy, proposed a lightweight physical rehabilitation assessment system (LightPRA) based on a Temporal Convolutional Network (TCN), utilizing the capabilities of dilated causal convolutional networks [38].

In recent years, the use of transformer architectures has expanded significantly across various fields. For example, Zhang Q. et al. presented a Recurrent Spatio-Temporal Transformer (RSTformer) [39], while Kanade et al. proposed an attention-guided framework combining transformer encoder, dense layers, and embedding layers for quality score prediction [40]. Mourchid et al. introduced a Dense Spatio-Temporal Graph Conv-GRU Network with a Transformer, integrating a modified STGCN with transformer architectures to efficiently handle spatio-temporal data. Their model utilizes convolutional gated recurrent units (Conv-GRU) to enhance computational efficiency [41]. Wang K. et al. developed a transformer-based model augmented with self-attention and channel attention mechanisms, which allows the network to selectively attend to anatomical regions crucial for motion execution [42]. He et al. incorporated expert knowledge into a graph convolutional approach to enhance spatial feature extraction while leveraging a transformer module to capture long-range temporal dependencies [43]. Table 1 summarizes the methods, input systems, pose detection methods, and datasets used for the current state-of-the-art technologies.

Numerous studies demonstrate several applications of convolutional and residual architectures, such as image classification [44–47]. Residual network-based models, such as ResNet, have achieved significant success in terms of performance [48]. Deep convolutional models incorporating residual units have shown compelling accuracy and improved convergence on a range of computer vision tasks [49–52]. However, despite their efficacy, studies have identified certain limitations to these architectures. Shen et al. highlighted two key

Table 1 Summary of methods, input systems, pose detection methods, and datasets

Authors	Method	Input System	Pose Detection Method (only for RGB)	Dataset
Jleli et al., 2024 [27]	YOLO V5–ShuffleNet V2-based image processor with bidirectional LSTM for scoring	Kinect	YOLO V5	KIMORE
Liao et al., 2020 [34]	Deep spatio-temporal neural network using CNN and LSTM layers	Kinect V2 and Vicon	-	UI-PRMD and KIMORE
Deb et al., 2022 [35]	Graph convolution network with LSTM	Kinect and Vicon	-	UI-PRMD and KIMORE
Yao et al., 2023 [36]	Multi-task contrastive learning framework	Kinect, Vicon, and RGB camera	BlazePose and VideoPose3D	KIMORE and UI-PRMD
Mourchid et al., 2023 [37]	Multi-Residual Spatio-Temporal Graph Network (MR-STGN) incorporating 3D skeletons	Kinect and Vicon	-	UI-PRMD
Sardari et al., 2024 [38]	LightPRA (Light Physical Rehabilitation Assessment) system using a Temporal Convolutional Network (TCN)	Kinect and Vicon	-	UI-PRMD and KIMORE
Zhang et al., 2023 [39]	Transformer-based network structure Recurrent Spatio-temporal Transformer (RSTformer)	Kinect V2 and Vicon	-	KIMORE and UI-PRMD
Kanade et al., 2023 [40]	Transformer encoder with embedding layers	Kinect and Vicon	-	UI-PRMD and KIMORE
Mourchid et al., 2023 [41]	Dense Spatio-Temporal Graph Conv-GRU Network with Transformer	Kinect and Vicon	-	UI-PRMD and KIMORE
Karlov et al., 2024 [62]	Spatial-Temporal Graph Convolutional Network (ST-GCN)	Kinect and Vicon	-	UI-PRMD, IRDS, and KIMORE
Pan et al., 2023 [63]	Self-supervised framework	Kinect and Vicon	-	UI-PRMD

issues in the original residual networks: incompatibility between ReLU and element-wise addition, as well as challenges in achieving convergence with models exceeding 1000 layers when using the “msra” initializer. They proposed weighted residual connections, adding learnable weights to the residual branch, which enables very deep models (layers > 1000) to train and converge more efficiently with minimal impact on GPU memory and computational load [53].

Taking inspiration from this, we propose a deep learning model incorporating a weighted residual transformer encoder. The proposed architecture demonstrates promising performance, as evidenced by the results. Additionally, the aforementioned studies often rely on publicly available datasets collected using high-cost systems such as Vicon. Although the Vicon system provides high precision, it is cost-prohibitive and impractical for home or hospital use, as it requires the user to wear a suit with body markers. Similarly, the Kinect system, although it eliminates the need for body markers by utilizing an RGB-D camera to track joint movements, is no longer supported by Microsoft and is incompatible with many modern systems. In contrast, RGB-based approaches, which can utilize a smartphone or laptop camera, offer a more accessible and practical alternative for home-based rehabilitation.

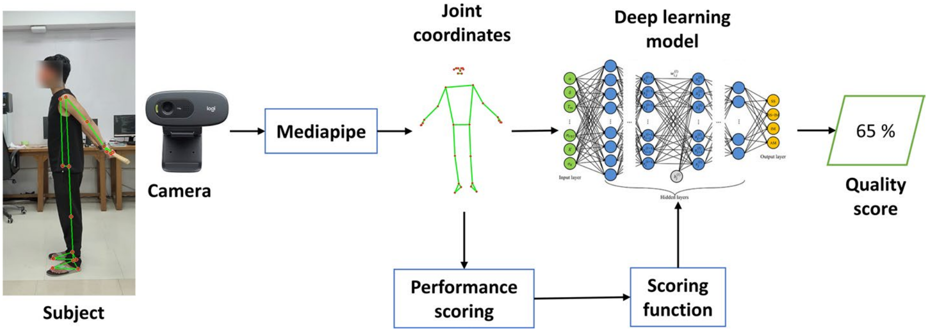


Fig. 2 Overview of the proposed home-based action quality assessment framework using RGB camera

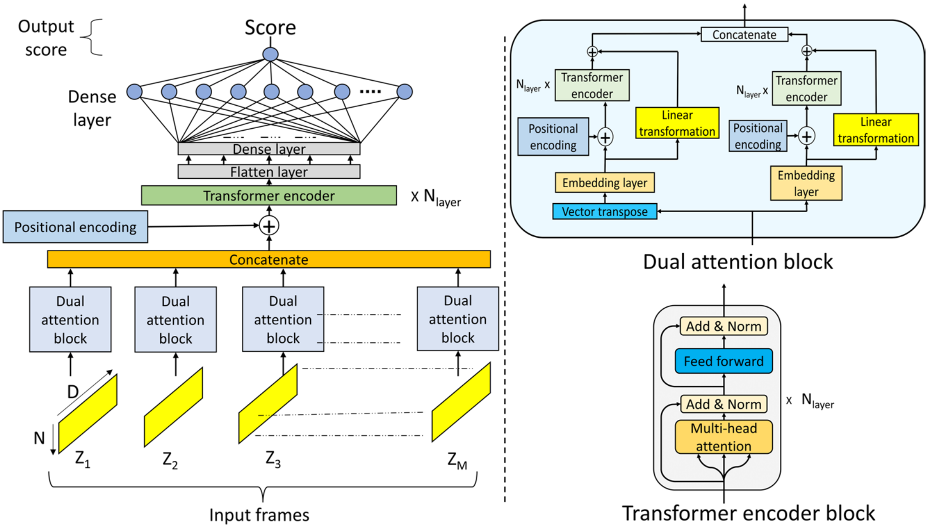


Fig. 3 Overview of the proposed spatio-temporal transformer encoder-based architecture for quality assessment

3 Proposed method

Figure 2 presents an overview of the proposed framework for human movement quality assessment. The system captures human motion in real time using an RGB camera or processes video frames from pre-recorded videos. Each frame, whether from a live camera feed or a video, is analyzed using MediaPipe to extract joint coordinates. For datasets that lack clinical scores, we employ a probabilistic model to quantify how closely a patient's movements match those of healthy individuals. The estimated performance metric values are mapped to movement quality scores to imitate ground truth clinical scores. Finally, the raw joint coordinates along with movement quality scores are used to train a transformer-based deep learning model (see Fig. 3) for performance score estimation.

Table 2 Demographic data of the participants

Subject	Age	Height (cm)	Weight (Kg)	Gender	Total repetitions per exercise
S1	28	163	67	M	160
S2	18	162.5	57	F	160
S3	18	162.5	56	F	160
S4	25	164	50	F	160
S5	24	163	50	F	160
S6	23	172	75	M	160
S7	26	177	74	M	160
S8	32	177	65	M	160
				Total	1280

Table 3 Rehabilitation exercises included in the study protocol

Order	Exercise
E1	Deep squat
E2	Inline lunge
E3	Hurdle step
E4	Sit-to-stand
E5	Standing shoulder scaption
E6	Standing shoulder abduction
E7	Standing shoulder internal-external rotation
E8	Standing shoulder extension

3.1 Data collection

We enrolled eight healthy individuals, four males and four females, aged 18–32 [mean(SD) = 24.25(4.74)]. All the enrolled subjects signed the informed consent. The study was approved by the local Institutional Ethics Committee (Humans) at the Indian Institute of Technology Ropar. Demographic characteristics of the study participants are provided in Table 2.

The subjects were asked to perform the experimental protocol, which included eight different rehabilitation exercises executed both correctly (to mimic healthy movement) and incorrectly (to mimic patient movement). Each exercise was repeated 10 times in both manners. The list of exercises performed is given in Table 3.

The rehabilitation exercises for our custom RGB dataset were recorded using a OnePlus 11R Android smartphone equipped with a 50MP rear camera. Videos were captured at a resolution of 720p and a frame rate of 30 FPS, with the smartphone mounted on a tripod positioned approximately 2.5–3 m from the subject. A total of 128 videos were recorded. Data processing was subsequently performed on an HP Z6 workstation equipped with an NVIDIA Quadro RTX5000 graphics card, 128 GB of RAM, and an Intel Xeon processor. Figure 1 displays sample video frames of a subject performing different rehabilitation exercises.

An overview of our proposed approach is depicted in Fig. 2. A detailed description of the different modules follows.

3.2 Data notation

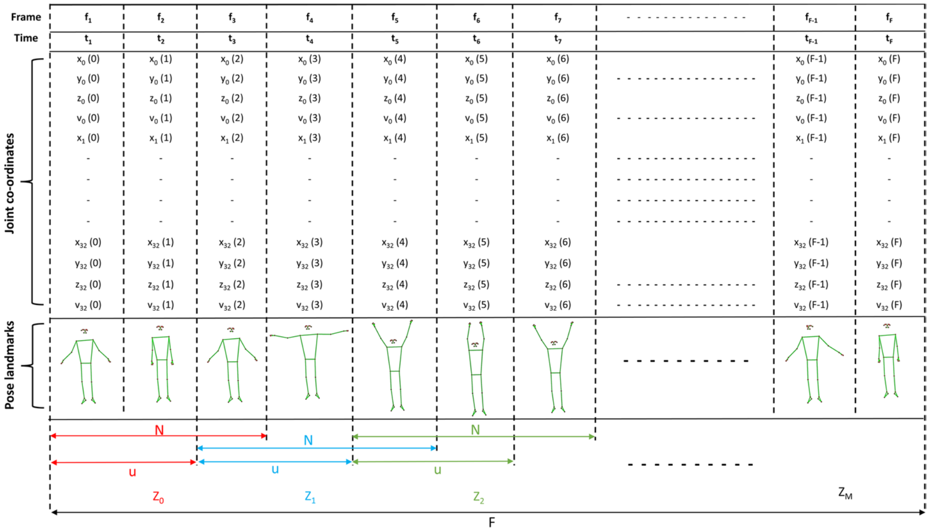
We use tensor notation to represent our dataset with healthy or correct movements of each exercise: $\chi \in \mathbb{R}^{S\times R\times F}$, where S is the number of subjects, R represents the number of repetitions of an exercise, and F is the number of video frames for each repetition. An element of χ is denoted by X_{srf}, where $1 \leq s \leq S$, $1 \leq r \leq R$ and $1 \leq f \leq F$ and $X_{srf} \in \mathbb{R}^D$ represents a feature vector of length D. Using similar tensor notation, we denote the patient data or incorrect movements of each exercise with $\mathrm{y} \in \mathbb{R}^{S\times R\times F}$, and $Y_{srf} \in \mathbb{R}^D$ represents a feature vector of length D.

3.3 Data processing

3.3.1 Joint coordinate extraction

We have used the MediaPipe open-source framework for joint coordinate extraction. MediaPipe Pose is a machine learning solution for high-fidelity body pose tracking from RGB video frames. Its real-time performance on devices like mobile phones makes it a viable option for home-based rehabilitation systems.

The landmark model in MediaPipe Pose predicts the location (3D coordinates) of 33 body landmarks. Additionally, it estimates the visibility of each landmark in every input frame to indicate whether it is present or hidden by other body parts or objects. Each landmark is represented by a vector $[x_i,\ y_i, z_i,\ v_i]$ with three spatial coordinates and one visibility factor. Combining all 33 landmarks generates a feature vector of length $D = 132$ for each frame of the video: x = $[x_0(f), y_0(f), z_0(f), v_0(f), \ldots, x_{32}(f), y_{32}(f), z_{32}(f), v_{32}(f)]$, where the subscripts represent landmarks and f is the frame index or time-point. The output of MediaPipe Pose Landmark model is shown in Fig. 4.



Frame	f_1	f_2	f_3	f_4	f_5	f_6	f_7	- - - - - - - - - - - - -	f_{F-1}	f_F
Time	t_1	t_2	t_3	t_4	t_5	t_6	t_7		t_{F-1}	t_F
Joint co-ordinates	x_0 (0)	x_0 (1)	x_0 (2)	x_0 (3)	x_0 (4)	x_0 (5)	x_0 (6)		x_0 (F-1)	x_0 (F)
	y_0 (0)	y_0 (1)	y_0 (2)	y_0 (3)	y_0 (4)	y_0 (5)	y_0 (6)	- - - - - - - - - - - - - - - - -	y_0 (F-1)	y_0 (F)
	z_0 (0)	z_0 (1)	z_0 (2)	z_0 (3)	z_0 (4)	z_0 (5)	z_0 (6)		z_0 (F-1)	z_0 (F)
	v_0 (0)	v_0 (1)	v_0 (2)	v_0 (3)	v_0 (4)	v_0 (5)	v_0 (6)	- - - - - - - - - - - - - - - - -	v_0 (F-1)	v_0 (F)
	x_1 (0)	x_1 (1)	x_1 (2)	x_1 (3)	x_1 (4)	x_1 (5)	x_1 (6)		x_1 (F-1)	x_1 (F)
	-	-	-	-	-	-	-	- - - - - - - - - - - - - - - - -	-	-
	-	-	-	-	-	-	-	- - - - - - - - - - - - - - - - -	-	-
	-	-	-	-	-	-	-	- - - - - - - - - - - - - - - - -	-	-
	-	-	-	-	-	-	-	- - - - - - - - - - - - - - - - -	-	-
	x_{32} (0)	x_{32} (1)	x_{32} (2)	x_{32} (3)	x_{32} (4)	x_{32} (5)	x_{32} (6)		x_{32} (F-1)	x_{32} (F)
	y_{32} (0)	y_{32} (1)	y_{32} (2)	y_{32} (3)	y_{32} (4)	y_{32} (5)	y_{32} (6)		y_{32} (F-1)	y_{32} (F)
	z_{32} (0)	z_{32} (1)	z_{32} (2)	z_{32} (3)	z_{32} (4)	z_{32} (5)	z_{32} (6)	- - - - - - - - - - - - - - - - -	z_{32} (F-1)	z_{32} (F)
	v_{32} (0)	v_{32} (1)	v_{32} (2)	v_{32} (3)	v_{32} (4)	v_{32} (5)	v_{32} (6)	- - - - - - - - - - - - - - - - -	v_{32} (F-1)	v_{32} (F)

Fig. 4 Output of MediaPipe Pose Landmark model with 3D spatial coordinates and visibility value for each of the 33 body landmarks

3.3.2 Action performance metric

In the absence of clinical scores for each exercise, the movement quality can be estimated by measuring the deviation of the patient's performance from that of healthy participants performing the same exercise. This estimation can be done using both model-less and model-based approaches. Model-less metrics use distance functions to quantify the deviation between reference and test data sequences. Common distance functions include Euclidean Distance, Mahalanobis Distance, and Dynamic Time Warping (DTW). While the primary advantage of these distance functions is their flexibility, they are limited by their inability to derive a model of the rehabilitation data. In contrast, model-based metrics employ probabilistic approaches to model the movement data. These methods often use Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) and evaluate performance based on log-likelihood values. The stochastic nature of probabilistic models makes them advantageous in that they can handle the variability inherent in human movements.

For a fair comparison with state-of-the-art approaches to AQA [34], we adopted the Gaussian mixture model-based metric (GMM) for movement assessment. GMMs are probabilistic models that assume the data is generated from a mixture of several Gaussian distributions, each representing a different component of the overall data distribution. Due to the probabilistic nature of this approach, it can effectively handle the natural variability in human movements. The log-likelihood values of GMMs can be leveraged to evaluate the patient's movement quality. Higher log-likelihood values suggest that the patient's movements are similar to those of healthy individuals, indicating better movement quality. Conversely, lower log-likelihood values indicate greater deviation and potentially lower movement quality.

As a parametric probability density function, a GMM is represented as a weighted sum of K Gaussian component densities and is given by the following equation:

$$p\left(x\right)=\sum{}_{k=1}^{K}\Phi_K\mathcal{N}\left(x\mid\mu_k,\sum{}_k\right) \tag{1}$$

where **x**, an element of $X_{srf}\in\mathbb{R}^D$, represents the joint-coordinate feature vector for one frame, $\mathcal{N}\left(x\mid\mu_k,\sum{}_k\right)$ represents a Gaussian function with parameters ϕ_k, μ_k, $\boldsymbol{\Sigma}_k$, which represent the mixing coefficient, mean vector, and covariance matrix of the k^{th} Gaussian component, respectively. The mixture weights satisfy the constraint that $\sum{}_{k=1}^{K}\Phi_k=1$.

The complete GMM is parameterized by the mean vectors, covariance matrices, and mixing coefficients of all component densities. These parameters can be collectively represented as $\theta=\{\phi_k,\mu_k,\boldsymbol{\Sigma}_k\}$, $k=1..K$.

By far, the most popular and well-established method for estimating the parameters of a GMM is maximum likelihood (ML) estimation using Expectation-Maximization (EM) algorithm. Here, the GMM parameters are estimated using the joint coordinates extracted from healthy movements as training feature vectors. Then, the trained GMM with parameters θ is used to estimate the performance metrics of patient movement based on log-likelihood [34], as shown below:

$$\updownarrow\left(\mathrm{Y}_{sr}|\theta\right)=\sum{}_{f=1}^{F}log\left\{\sum{}_{k=1}^{K}\Phi_k\mathcal{N}\left(Y_{srf}\mid\mu_k,\sum{}_k\right)\right\} \tag{2}$$

where $\mathbf{Y}_{srf} = \mathbf{y} \in \mathbb{R}^D$ represents the joint-coordinate feature vector for patient s, repetition r, and frame f.

3.3.3 Scoring function

The log-likelihood values range from large positive to large negative numbers. To make these values more understandable for patients, especially in a home environment, it is necessary to scale the scores to a more intuitive range. For example, a score of 83 out of 100 is more understandable than a value in an arbitrary range.

Let $\mathbf{L} = \{\mathbf{L}healthy, \mathbf{L}patient\}$, where both $\mathbf{L}healthy$ and $\mathbf{L}patient$ represent the log-likelihood values estimated by the GMM for each of the R repetitions of an exercise in both correct and incorrect forms: $[\ell_1, \ell_2, ., \ell_R]$. Consequently, the min-max normalization [54] of these scores is given by the following equation:

$$\ell_r' = \frac{\ell_r - \min\{L\}}{\max\{L\} - \min\{L\}} \tag{3}$$

For enhanced intuition and effective comparison, Fig. 5 shows the distribution of normalized scores for a sample exercise (E5 - standing shoulder scaption) performed correctly and incorrectly.

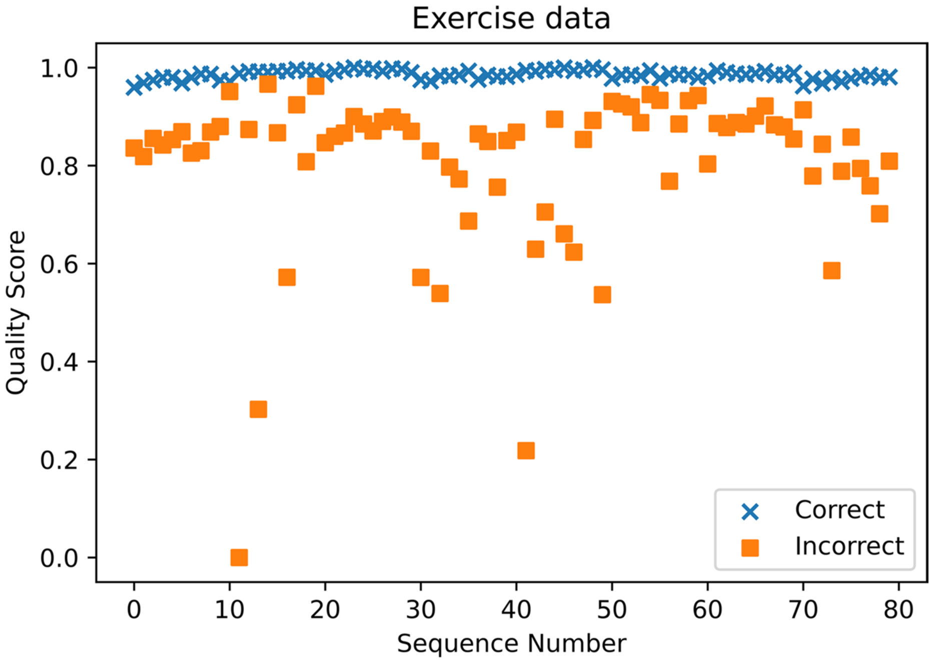


Fig. 5 Distribution of normalized quality assessment scores for a sample exercise of standing shoulder scaption

3.4 Deep learning model architecture

We propose utilizing a spatio-temporal transformer encoder-based architecture to assess the quality of exercises by analyzing movement patterns from video sequences. The RGB video sequence of an exercise has two types of inherent dependencies: spatial dependency, which represents the relationship between the joint-coordinates within a frame, and temporal dependency, which denotes relationship of joint-coordinates across frames. Standard spatio-temporal transformer models include spatial and temporal transformer encoders that capture relationships within a frame and dependencies over time, respectively.

For improved and accurate quality assessment of movement, our work proposes to capture both local (fine) and global (coarse) patterns in two phases. The proposed model comprises three key components. First, a data segmentation module divides the video sequence into smaller segments of contiguous frames using a sliding window mechanism. Second, a dual attention block, consisting of both spatial and temporal transformer encoders, captures spatio-temporal patterns in local video frames within a window. Lastly, a temporal transformer encoder captures global temporal patterns across local video segments.

3.4.1 Problem statement

Given an input sequence $X = [\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_F]$ of length F, we aim to predict a corresponding motion quality assessment score u, where each element $x_f \in \mathbb{R}^D$ represents a D-dimensional feature vector corresponding to joint coordinates extracted from an individual frame f, and u is a scalar value that represents the target normalized quality score. We define this prediction task as a mapping $f: X \rightarrow u$ that leverages a spatio-temporal transformer encoder-based model to learn dependencies within and across elements of X. The goal of the transformer model is to learn the mapping $f(X) = u$ by minimizing a loss function $\mathcal{L}(u, \widehat{u})$, where $\widehat{u} = f(X)$ represents the predicted quality assessment score.

3.4.2 Data segmentation

Instead of processing the entire video sequence at once, we divide it into smaller segments of contiguous frames using a sliding window mechanism. Figure 4 in Sect. 3.3.1 shows the joint-coordinate data output by the MediaPipe Pose for a single repetition of an exercise. This can be represented as a $2D$ matrix: $\mathbf{X}_{sr} = X = [\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_F]^{\mathrm{T}} \in \mathbb{R}^{F \times D}$ where F is the number of frames and each $\mathbf{x}_f$ represents the joint-coordinate feature vector of one frame with length D.

As depicted in Fig. 4, X is split into local segments of contiguous frames using an overlapping sliding window mechanism. For a window size of N and a stride of u, an exercise repetition with F frames, the number of windows generated, M, is given by the following equation:

$$M = \left\lfloor \frac{F - N}{u} \right\rfloor + 1$$

For each window $m \in [0, M-1]$, the sliced data matrix or video segment is given by $x_m = X_{[m\times u]:[(m\times u)+N]} \in \mathbb{R}^{N\times D}$. If required, zero-padding is used to produce fixed-length sequence from the last segment [55, 56].

3.4.3 Dual attention block

The data segments generated by the sliding window technique include small video segments that contain both spatial and temporal information within each local window. To capture these short-range spatio-temporal patterns, we have devised a Dual Attention Block that builds on the standard spatio-temporal transformer architecture by introducing two distinct attention mechanisms, one dedicated to spatial attention and the other to temporal attention. This separation allows the model to more effectively capture intricate dependencies in data with both spatial and temporal dimensions, as depicted in Fig. 6.

Input embedding and positional encoding For any transformer model, it is crucial to transform the Raw input data into a format suitable for processing. We assume the transformer uses a constant latent vector of size d_{model} through all of its layers [57]. First, we used input embedding to map each input feature vector to match the high-dimensional vector with a trainable linear projection. for a generic input vector $\mathbf{x} \in \mathbb{R}^d$, the embedding is given by $\mathbf{z}=\mathbf{x}E$, where $\mathbf{z} \in \mathbb{R}^{dmodel}$ and $E \in \mathbb{R}^{d\times dmodel}$ represents the embedding matrix

Next, we added positional encodings to the input embeddings to retain positional information in the sequence. These encodings have the same dimension, d_{model}, as the input embeddings, which allows them to be summed together. Hence, the final embedding that combines both input embedding (E) and positional encoding (E_{pos}) is given by $\mathbf{z}=\mathbf{x}E+E_{pos}$.

Spatial attention This module processes a multi-frame, joint coordinate data segment with a Spatial transformer encoder to capture Spatial dependencies within and across multiple frames. To focus on the contextual relationship in Spatial dimension, the multiframe data

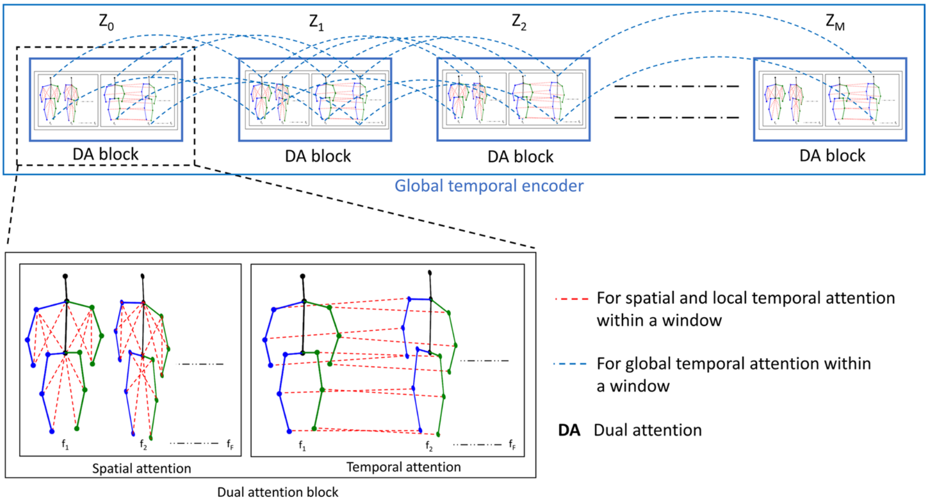


Fig. 6 Local spatial and temporal attention in addition to global temporal attention

segment, $x_m \in \mathbb{R}^{N\times D}$, is first transposed to $x_m^{'} \in \mathbb{R}^{D\times N}$ and embedded into the latent feature space $z_m^{'} \in \mathbb{R}^{D\times d_{model}}$. Next, a self-attention mechanism is used to capture Spatial relationships among joints. The Spatial dependencies between joints in different contexts are extracted using multi-head attention, as described below. The output of the Spatial attention block that provides a Spatial summary of joints across multiple frames is given by $\overline{z}_m \in \mathbb{R}^{D\times d_{model}}$.

Temporal attention A temporal transformer encoder is used to capture Temporal dependencies between frames in a segment. Here, the input multi-frame segment, $x_m \in \mathbb{R}^{N\times D}$, is directly embedded into the latent feature space $z_m \in \mathbb{R}^{N\times d_{model}}$. Across frames, self-attention is applied to capture dependencies over time, focusing on Temporal relationships. The Temporal summary of joints in the multi-frame segment as calculated by the Temporal attention block is represented by $\widetilde{z}_m \in \mathbb{R}^{N\times d_{model}}$.

Transformer encoder Standard transformer encoder is a powerful architecture widely used in many computer vision tasks. It consists of multiple layers (N_{layer}), each containing two main components: multi-head attention and feed-forward neural networks. Multi-head self-attention is a sophisticated mechanism in the transformer model designed to capture intricate dependencies within a sequence. It enhances the standard self-attention mechanism by utilizing multiple parallel attention heads. Each head performs self-attention independently with its own set of query (Q), key (K), and value (V) matrices. For a single attention head, the self-attention mechanism computes the attention scores using the scaled dot-product formula:

$$Attention\,(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_{model}}}\right)V$$

where d_{model} is the dimensionality of the keys. The softmax function ensures that the attention scores are normalized, giving a probability distribution over the values.

In multi-head self-attention, the input embeddings are linearly projected into H different subspaces (heads) using learned projection matrices: $Q_h = XW_h^Q,\ K_h = XW_h^K$, and $V_h = XW_h^V$, where $W_h^Q,\ W_h^K$ and W_h^V are the projection matrices for the h-th head.

Each head h computes self-attention as follows:

$$Head_h = Attention(QW_h^Q,\ KW_h^K,\ VW_h^V)$$

The outputs from all heads are then concatenated and passed through a final linear projection:

$$MultiHead(Q,K,V\ ) = Concat(Head_1,\ \ldots,\ Head_H)W^O$$

where W^O is the output projection matrix. Following the multi-head self-attention, the encoder layer includes a position-wise feed-forward neural network (FFN). This feed-forward network introduces non-linearity and allows the model to transform the representations further. Both sub-layers (multi-head self-attention and feed-forward neural network) are followed by a residual connection and layer normalization.

Weighted residual transformer encoder Residual connections are a simple yet powerful way to mitigate the vanishing gradient problem, as they allow gradients to flow directly through the network. To enable our model to capture complex dependencies and relationships within the input sequence, we propose a weighted residual transformer block. It introduces an additional, weighted skip connection across both the sub-layers—multi-head attention and feed-forward network—of the transformer. The effect of this weighted residual connection can be represented using a linear transformation:

$$\widehat{z} = z_{out} + z_{in} W_R$$

where $\mathbf{z}_{in}$ and $\mathbf{z}_{out}$ represent the input and output of a standard transformer block and $W_R \in \mathbb{R}^{d_{model} \times\, d_{model}}$ is the learnable weight matrix. In addition to mitigating the vanishing gradient problem, the proposed enhancement encourages feature reuse and enhanced information propagation.

Spatio-temporal feature aggregation The outputs of both Spatial and Temporal attention modules are concatenated to generate the final output of a dual attention block, as given below:

$$A_m = Concat(\overline{z}_m,\ \widetilde{z}_m)$$

Furthermore, the outputs of all the dual attention blocks (that process individual window segments) are concatenated as:

$$A = \ Concat(A_0,\ A_1,\ A_2,\ A_3,\ \ldots\ldots,\ A_{M-1})$$

where $A_m \in \mathbb{R}^{(N+D)\times\ d_{model}}$ and $A \in \mathbb{R}^{M(N+D)\times\ d_{model}}$.

3.4.4 Global attention block

The global attention block processes the sequence of short-range spatio-temporal features extracted by the dual attention blocks from local segments to capture global temporal dependencies. As the standard transformer encoder lacks inherent knowledge of order, positional encodings are added to the dual attention feature vectors to retain the temporal order of the sequence. Hence, the input to the global attention module is given by $A + E_{pos}$.

The self-attention mechanism in a global temporal transformer encoder allows each segment (or window) to attend to every segment in the sequence. This global scope ensures that the model captures long-range dependencies across the entire sequence. Furthermore, the use of multiple attention heads allows the model to capture different aspects of temporal relationships, each focusing on different parts of the sequence. The temporal summary of the sequence of segments as calculated by the attention block is represented by $B \in \mathbb{R}^{M(N+D)\times\ d_{model}}$.

After the self-attention operation, the output is flattened and passed through a feed-forward network, which consists of a fully connected or dense layer with N_{nodes}. Finally, an output layer with a single node helps estimate assessment scores.

4 Experiments

4.1 Datasets

In addition to our own dataset, we conducted extensive experiments using rehabilitation exercises from two public benchmark datasets.

4.1.1 KIMORE dataset

This dataset includes RGBD videos recorded using the Kinect system, with ground truth scores for people performing various rehabilitation exercises. It is divided into two groups: a control group comprising experts and non-experts, and a patient group with individuals suffering from pain and postural disorders (such as stroke, back pain, and Parkinson's). The dataset contains a heterogeneous population of 78 people, including 44 healthy subjects (12 experts and 32 non-experts) and 34 patients with chronic motor disabilities. The exercises are ($E1$) lifting the arms up, ($E2$) lateral tilt of the trunk with arms extended, ($E3$) trunk rotation, ($E4$) pelvis rotations on the transverse plane, and ($E5$) squatting.

4.1.2 UI-PRMD dataset

This dataset was recorded using two motion capture systems, Vicon and Kinect. It includes 10 subjects performing 10 different rehabilitation exercises, both correctly and incorrectly, with ground truth scores generated using trained GMM models. For our model training and evaluation, we used the data recorded via the Vicon system in the UI-PRMD dataset. The training and evaluation focused on positional data, as this dataset contains both positional and angular data. The exercises included are ($E1$) deep squats, ($E2$) hurdle step, ($E3$) inline lunge, ($E4$) side lunge, ($E5$) sit-to-stand, ($E6$) standing active straight leg raise, ($E7$) standing shoulder abduction, ($E8$) standing shoulder extension, ($E9$) standing shoulder internal–external rotation, and ($E10$) standing shoulder scaption.

4.2 Evaluation metrics

We use the following metric to evaluate and compare our proposed approach.

- Mean Absolute Error (MAE) is a commonly used evaluation metric for regression tasks. It measures the average magnitude of errors between predicted values and actual values without considering their direction.

$$MAE = \frac{1}{n}\sum {}_{i=1}^{n}\left|l_i - \widehat{l_i}\right| \tag{4}$$

- Root Mean Squared Error (RMSE) is another commonly used evaluation metric for regression tasks. This is the square root of the mean of squared errors. This metric is helpful as it penalizes errors due to the squared terms.

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(l_i - \widehat{l_i}\right)^2} \tag{5}$$

where l_i is the predicted value and $\widehat{l_i}$ is the actual or true value.

4.3 Implementation details

All experiments were conducted on an HP Z6 workstation equipped with an NVIDIA Quadro RTX5000 graphics card, an Intel Xeon CPU, and 128 GB of RAM. Our model was implemented in Python using PyTorch (v2.3.1) and Tensorflow (v2.10.0), with additional dependencies including NumPy (v1.23) and SciPy (v1.9). We used the Adam optimizer and binary cross-entropy [34] as the loss function to train our models for 100 epochs across all datasets. Learning rates were set to 0.00001 and 0.0001 for the KIMORE and UI-PRMD datasets, respectively. Further parameter details for all datasets are provided in Table 4.

We used the K-fold Cross-Validation technique, as it provides a more accurate measure of model performance by reducing the variance associated with a single train-test split. It exposes the models to multiple train-test splits to help identify those that are not only robust but that generalize well to unseen data. We also employed Dropout as a regularization method to improve generalization and prevent overfitting. We determined the choice of hyperparameter values empirically based on validation performance.

5 Results

5.1 Comparison of custom RGB dataset

We compared the performance of our proposed approach with two other state-of-the-art models, one proposed by Liao et al. [34] and the Light PRA model by S. Sardari et al. [38], using our custom RGB camera dataset. The results, evaluated using Mean Absolute Error (MAE), are presented in Table 5.

Our proposed model demonstrated superior performance, outperforming the other models in six out of eight exercises. For the remaining two exercises, it achieved the second-best results. This consistent out-performance across the majority of exercises highlights the robustness and effectiveness of our approach. Additionally, the comparatively lower standard deviation values across most exercises indicate that our model not only achieves higher accuracy but also maintains consistent performance. This consistency is crucial for practical applications, as it ensures reliable assessment across different types of exercises and varying conditions.

5.2 Comparison of benchmark datasets

A comparison of our model with several state-of-the-art models on the KIMORE and UI-PRMD datasets is presented in Tables 6 and 7, respectively.

Table 4 Parameters of the best model found

Dataset	N_{nodes}	Dropout	d_{ff}	$\mathrm{d}_{\mathrm{model}}$	H	N_{layers}	N	u
Custom RGB	118	0.1	56	48	4	6	60	60
UI-PRMD	118	0.1	48	48	4	6	60	60
KIMORE	118	0.2	8	8	4	6	50	50

Table 5 Comparison of model performances on Custom RGB Camera Dataset using K-fold cross-validation ($k = 4$) and MAE as an evaluation metric. (Lower is better. The best and the second-best scores are in bold and italicized, respectively.)

(a) Our proposed model

Mean absolute error (MAE)

Exercise	Fold 1	Fold 2	Fold 3	Fold 4	Mean ± Std dev.
E1	0.03116	0.09433	0.05023	0.03815	**0.05347** ± **0.02836**
E2	0.04369	0.05819	0.05202	0.07601	**0.05748** ± **0.01371**
E3	0.04337	0.05282	0.05025	0.03293	*0.04484* ± *0.00889*
E4	0.06237	0.02815	0.01947	0.06887	**0.04472** ± **0.02454**
E5	0.06381	0.06435	0.06614	0.04823	*0.06063* ± *0.00833*
E6	0.06123	0.09878	0.04276	0.18664	**0.09735** ± **0.06393**
E7	0.06257	0.17540	0.11317	0.10735	**0.11462** ± **0.04640**
E8	0.07273	0.06012	0.04630	0.03015	**0.05232** ± **0.01830**
				Average	**0.06568** ± **0.01955**

(b) Using Liao et al. model [34]

Mean absolute error (MAE)

Exercise	Fold 1	Fold 2	Fold 3	Fold 4	Mean ± Std dev.
E1	0.05170	0.12049	0.18233	0.04168	0.09905 ± 0.06565
E2	0.04730	0.07365	0.07439	0.09446	*0.07245* ± *0.01934*
E3	0.01323	0.04267	0.03387	0.01746	**0.02681** ± **0.01382**
E4	0.08466	0.05597	0.05608	0.08854	*0.07131* ± *0.01772*
E5	0.05515	0.05255	0.07463	0.02031	**0.05066** ± **0.02251**
E6	0.10412	0.10822	0.09417	0.19730	*0.12595* ± *0.04793*
E7	0.07708	0.10924	0.15949	0.18819	*0.13350* ± *0.04979*
E8	0.20587	0.06637	0.06050	0.04191	*0.09366* ± *0.07553*
				Average	*0.08417* ± *0.03904*

(c) Using Light PRA model [38]

Mean absolute error (MAE)

Exercise	Fold 1	Fold 2	Fold 3	Fold 4	Mean ± Std dev.
E1	0.05903	0.12323	0.01215	0.00126	*0.04892* ± *0.05552*
E2	0.15886	0.18295	0.29473	0.17148	0.20201 ± 0.06260
E3	0.36529	0.22566	0.11384	0.05433	0.18978 ± 0.13688
E4	0.08771	0.23928	0.04990	0.11688	0.12344 ± 0.08195
E5	0.12238	0.08947	0.11729	0.17356	0.12568 ± 0.03505
E6	0.08164	0.17353	0.08529	0.29244	0.15822 ± 0.09905
E7	0.10226	0.25472	0.17973	0.18135	0.17951 ± 0.06225
E8	0.39292	0.23256	0.10216	0.61598	0.33591 ± 0.22137
				Average	0.17043 ± 0.09433

5.2.1 KIMORE dataset

Our proposed model demonstrated significant performance improvements across all five exercises on the KIMORE dataset recorded using a Kinect sensor. This indicates the model's strong ability to accurately assess and analyze movement patterns captured by the Kinect system. The superior performance can be attributed to the model's advanced architecture, which effectively captures spatio-temporal dependencies and leverages the dual attention module and weighted residual connections.

Table 6 Comparison of model performances on KIMORE dataset using MAE and RMS as evaluation metrics. (Lower is better. The best and the second-best scores are in bold and italicized, respectively.)

(a)

Mean Absolute Error (MAE)

Exercise	Our approach	LightPRA [38]	Jleli et al. [26]	Deb et al. [35]	Song et al. [58]	Zang et al. [59]	Liao et al. [34]
E1	**0.22**	*0.25*	0.48	0.80	0.98	1.76	1.14
E2	**0.23**	*0.28*	0.52	0.77	1.28	3.14	1.53
E3	**0.20**	*0.25*	0.39	0.37	1.11	1.74	0.85
E4	**0.23**	*0.30*	0.48	0.35	0.72	1.20	0.47
E5	**0.21**	*0.28*	0.49	0.62	1.54	1.85	0.85
Average	**0.217**	*0.272*	0.472	0.582	1.126	1.938	0.925

(b)

Root Mean Square (RMSE)

Exercise	Our approach	LightPRA [38]	Jleli et al. [26]	Deb et al. [35]	Song et al. [58]	Zang et al. [59]	Liao et al. [34]
E1	**0.09**	*0.25*	1.03	2.02	2.17	2.92	2.53
E2	**0.13**	*0.32*	1.10	2.12	3.35	4.14	3.74
E3	**0.11**	*0.19*	1.21	0.56	1.93	2.62	1.56
E4	**0.07**	*0.30*	1.05	0.64	2.02	1.84	0.79
E5	**0.11**	*0.27*	1.00	1.18	3.20	2.92	1.91
Average	**0.105**	*0.266*	1.077	1.305	2.531	2.885	2.108

Table 7 Comparison of model performances on UI-PRMD dataset using MAE as evaluation metric. (Lower is better. The best and the second-best scores are in bold and italicized, respectively.)

Mean Absolute Error (MAE)

Exercise	Our approach	LightPRA [38]	Long et al. [36]	Liao et al. [34]	Deb et al. [35]	Song et al. [58]	Zang et al. [59]	Li et al. [60]	Pan et al. [63]
E1	**0.009**	0.014	0.015	0.011	**0.009**	0.011	0.022	0.011	*0.010*
E2	0.012	*0.007*	0.012	0.028	**0.006**	**0.006**	0.008	0.029	0.012
E3	**0.008**	0.011	0.015	0.039	0.013	0.010	0.016	0.056	0.015
E4	*0.008*	**0.006**	0.008	0.012	**0.006**	0.014	0.016	0.014	0.010
E5	0.014	**0.008**	*0.009*	0.019	**0.008**	0.013	**0.008**	0.017	**0.001**
E6	0.011	**0.006**	0.010	0.018	**0.006**	0.009	*0.008*	0.019	0.012
E7	**0.009**	*0.010*	0.011	0.038	0.011	0.017	0.021	0.027	0.013
E8	0.018	**0.011**	0.018	0.023	*0.016*	0.017	0.025	0.025	*0.013*
E9	0.016	**0.008**	*0.010*	0.023	**0.008**	**0.008**	0.027	0.027	0.014
E10	**0.012**	0.038	0.044	0.042	*0.031*	0.038	0.066	0.047	*0.016*
Average	*0.0118*	0.0119	0.0150	0.0253	**0.0114**	0.0143	0.0217	0.0272	0.0126

5.2.2 UI-PRMD dataset

Against the ten exercises on the UI-PRMD dataset, our model achieved the best performance in four exercises and the second-best in one exercise. While our model's performance was on par with the Light PRA model [38], the model by Deb et al. [35], which employed spatio-temporal graph convolutional networks (STGCN), showed superior performance across most exercises. Nevertheless, our model achieved the second-best average performance across all exercises, underscoring its competitive edge in the field.

5.3 Ablation study

We conducted comprehensive ablation studies to investigate the effect of various enhancements proposed in our work, namely, the dual attention module, generic residual connections, and weighted residual connections across the transformer blocks. These ablation experiments were performed using both the UI-PRMD and KIMORE datasets. The results are shown in Tables 8 and 9, respectively.

Our findings reveal that incorporating either dual attention, simple residual connections, or weighted residual connections into the standard transformer model results in performance improvements. Specifically, dual attention enhances the model by capturing short-range spatio-temporal dependencies, while residual connections facilitate the flow of gradients and the reuse of contextual information, thereby stabilizing training.

The best performance, however, is achieved when both dual attention and weighted residual connections are combined. This combination effectively captures intricate spatio-temporal patterns and maximizes the reuse of contextual information across transformer layers, leading to superior results compared to individual enhancements. The observed improvements underscore the importance of these enhancements in refining the model's ability to accurately assess movement quality.

5.4 Computational efficiency

We compared the computational efficiency of our approach with that of two state-of-the-art models, LightPRA [38] and the model by Liao et al. [34], the results of which are presented in Table 10. The comparison was performed in terms of inference time, total training time, and total trainable parameters, using the UI-PRMD dataset.

Table 8 Ablation study of our model on UI-PRMD dataset with based on mean absolute error (MAE) criteria with the exclusion of weighted residual, residual, and double attention

Exercise	Mean Absolute Error (MAE)					
Standard Transformer	x	x	x	✓	✓	✓
Dual Attention	✓	✓	✓	x	x	x
Residual (identity)	x	✓	x	x	✓	x
Residual (weighted)	✓	x	x	✓	x	x
E1	0.009162	0.014350	0.013155	0.017612	0.016773	0.019876
E2	0.012189	0.010720	0.012060	0.014249	0.013659	0.011852
E3	0.007955	0.006812	0.007204	0.018079	0.010019	0.016253
E4	0.008131	0.007851	0.008487	0.009819	0.010081	0.009236
E5	0.014075	0.014869	0.012651	0.016596	0.017124	0.014075
E6	0.010640	0.013489	0.011577	0.015411	0.015073	0.015367
E7	0.009104	0.009573	0.012434	0.016500	0.014539	0.015704
E8	0.017917	0.019105	0.019104	0.019554	0.020976	0.023452
E9	0.016382	0.014701	0.016163	0.014782	0.014662	0.016246
E10	0.012317	0.012364	0.014707	0.014592	0.014308	0.013893
Average	**0.011787**	0.012383	0.012754	0.0157194	0.0147214	0.015595

Table 9 Ablation study of our model on the KIMORE dataset based on mean absolute error (MAE) criteria with the exclusion of weighted residual, residual, and double attention

Exercise	Mean Absolute Error (MAE)					
Standard Transformer	x	x	x	✓	✓	✓
Dual Attention	✓	✓	✓	x	x	x
Residual (identity)	x	✓	x	x	✓	x
Residual (weighted)	✓	x	x	✓	x	x
E1	0.216002	0.216698	0.216080	0.215494	0.214594	0.219524
E2	0.225947	0.226520	0.222956	0.225192	0.227421	0.234730
E3	0.201067	0.199292	0.201037	0.201684	0.199615	0.205828
E4	0.229890	0.235219	0.230001	0.231579	0.228296	0.230012
E5	0.212730	0.212610	0.219763	0.221024	0.217787	0.211705
Average	**0.217127**	0.218068	0.217967	0.218995	0.217543	0.220359

Table 10 Comparison of computational cost of models on UI-PRMD dataset based on training time, inference time, and number of model parameters (Lower is better. Best scores are in bold)

Exercise	Our approach		Liao et al. [34]		LightPRA [38]	
	Inference time (in sec)	Total training time (in sec)	Inference time (in sec)	Total training time (in sec)	Inference time (in sec)	Total training time (in sec)
m01	0.1272	2663.97	0.2082	219.04	0.4267	746.96
m02	0.1302	2635.76	0.1677	134.12	0.4197	655.77
m03	0.1298	2639.22	0.1667	212.23	0.4235	654.10
m04	0.1278	2643.65	0.1685	222.54	0.4330	663.31
m05	0.1274	2673.28	0.1562	266.63	0.4367	679.88
m06	0.1373	2645.22	0.1722	298.51	0.4397	678.37
m07	0.1378	2652.99	0.1682	274.28	0.4359	678.96
m08	0.1398	2660.18	0.1685	153.03	0.4330	677.49
m09	0.1356	2636.55	0.1665	151.02	0.4293	690.39
m10	0.1386	2651.53	0.1662	247.02	0.4126	670.43
Average	**0.1332**	2650.23	0.1709	**217.84**	0.4290	679.57
Parameters	4,789,215		5,688,081		**122,461**	

As shown in the table, our model achieves the lowest inference time. Although LightPRA [38] has the fewest trainable parameters, it exhibits a notably higher inference time. To ensure a fair comparison, all models were trained and tested on the same hardware system (HP Z6), and the code for the comparative models was sourced from their respective repositories.

6 Discussion

The comparative analysis in this study highlights the strengths of our proposed approach. The dual attention module enhances the model's ability to capture intricate spatio-temporal patterns, while the weighted residual connections facilitate the reuse of contextual information across layers, leading to improved performance and consistency. Despite the superior performance of the STGCN model on the UI-PRMD dataset, our model's second-best performance demonstrates its effectiveness and potential for further enhancements.

It is important to note that the UI-PRMD dataset was recorded using the expensive and impractical VICON system, as it requires users to wear a suit with multiple markers. Although the KIMORE dataset uses the more affordable KINECT system, which does not require specialized suits, our approach tops it by utilizing a simple RGB camera, which makes it a cost-effective and practical solution for home-based rehabilitation by eliminating the need for expensive equipment and complicated setups. This highlights the accessibility and convenience of our model for effective home rehabilitation.

Limitations Despite its several notable merits, our study has some potential limitations. First, while our model is designed to handle variable-length inputs, this is achieved through zero-padding, which introduces additional redundancy. Additionally, our model was trained and evaluated on data collected under standard conditions with good lighting and camera quality. Although it leverages MediaPipe, which was trained on diverse lighting, noise, and motion conditions to enhance adaptability in real-world environments, the model's robustness still needs to be assessed under more varied lighting and environmental conditions. Further, our current work focused on developing a deep learning-based model capable of accurately assessing motion quality from RGB camera footage, with evaluations performed on pre-recorded rehabilitation exercises. However, it currently lacks real-time implementation for on-device assessment and feedback. Furthermore, although the KIMORE dataset contains data from real patients, our current setup lacks an extensive rehabilitation dataset, particularly one with diverse patient data, which may limit the generalizability of our model.

Future work Our future efforts will be primarily directed toward collecting more comprehensive data with significant patient demographics, captured in diverse "in-the-wild" conditions using various camera quality devices. More importantly, we aim to expand this work by developing a complete end-to-end system with real-time, on-device assessment and feedback capabilities. We also plan to leverage transfer learning to fine-tune our model, enabling it to support new rehabilitation exercises not explicitly represented in the original datasets, thereby enhancing its adaptability and utility for a wider range of rehabilitation contexts. Furthermore, we plan to explore the potential of Generative AI models such as SCA-GAN [61] to synthesize realistic skeleton sequences that simulate rehabilitation exercises, thereby improving the diversity and robustness of rehabilitation datasets. These GAN models have diverse applications [64] and may assist in augmented meaningful data generation for quality assessment tasks [65]. Additionally, while the size of our proposed transformer model is comparable to existing models, we intend to explore optimization techniques, such as pruning and knowledge distillation, to create a lightweight version suitable for mobile deployment without sacrificing performance. Finally, we plan to investigate potential applications in other medical domains, including mental health rehabilitation, to broaden the model's impact and explore new avenues for patient support.

Broader impact Our findings suggest a promising advancement in rehabilitation technology, moving toward bridging the gap between clinical and home-based rehabilitation using an affordable RGB camera-based system for advanced motion quality assessment. This development is especially valuable for patients facing barriers to regular in-person supervision, such as geographic, financial, or logistical challenges. The robustness and state-of-the-art performance of our system indicate that it could serve as an accessible alternative to traditional, costly motion-capture systems such as Vicon or Kinect, which are not feasible for widespread home use. This innovation has the potential to improve adherence to prescribed rehabilitation routines by enabling patients to receive real-time feedback on their exercises from home. Additionally, it may help reduce healthcare costs associated with extended clinical supervision, while promoting better recovery outcomes by allowing patients to self-monitor and adjust their movements independently. More broadly, our approach could drive further advancements in tele-rehabilitation and personalized healthcare, making quality rehabilitation more accessible to a diverse patient population.

7 Conclusion

In this paper, we proposed a spatio-temporal transformer encoder-based model with dual attention and weighted residual connections to assess the quality of human motion using skeleton data. Our model processes skeleton data to predict the quality of actions and assign scores accordingly. The dual attention mechanism captures intricate spatio-temporal dependencies, while the weighted residual connections significantly enhance the model's overall performance, as demonstrated by our experiments. Our proposed model achieved state-of-the-art performance when tested on our recorded data, as well as on two publicly available datasets, UI-PRMD and KIMORE, outperforming other contemporary models. We recorded the data using a simple RGB camera and extracted joint coordinates using Mediapipe, an open-source framework. The model's exemplary performance on this setup underscores its robustness and effectiveness.

Given its straightforward and affordable RGB camera setup, our proposed system is particularly well-suited for home-based rehabilitation. It offers a simple, cost-effective, and reliable solution, making advanced motion assessment without expensive or cumbersome equipment accessible to a wider population. This approach has the potential to revolutionize home-based rehabilitation, providing high-quality assessments that are both convenient and practical for users.

Funding This research have not received any funding.

Data availability The datasets analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest None.

Ethical approval The required ethical approval was obtained from the local Institutional Ethics Committee (Humans) at Indian Institute of Technology Ropar.

References

1. Warburton DE, Nicol CW, Bredin SS (2006) Health benefits of physical activity: the evidence. CMAJ 174:801–809
2. Cameron ID (2002) Kurrle, 1: rehabilitation and older people. Med J Australia 177(7):387–391
3. Komatireddy R, Chokshi A, Basnett J, Casale M, Goble D, Shubert T (2014) Quality and quantity of rehabilitation exercises delivered by a 3-D motion controlled camera: a pilot study. Int J Phys Med Rehab 2(4):214
4. Miller KK, Porter RE, DeBaun-Sprague E, Van Puymbroeck M, Schmid AA (2017) Exercise after stroke: patient adherence and beliefs after discharge from rehabilitation, top. Stroke Rehabil 24(2):142–148
5. Bassett SF, Prapavessis H (2007) Home-based physical therapy intervention with adherence-enhancing strategies versus clinic-based management for patients with ankle sprains. Phys Ther 87(9):1132–1143
6. Jack K, McLean SM, Moffett JK, Gardiner E (2010) Barriers to treatment adherence in physiotherapy outpatient clinics: a systematic review. Man Ther 15(3):220–228
7. Schutzer KA, Graves BS (2004) Barriers and motivations to exercise in older adults. Prev Med 39:1056–1061
8. Yu BXB, Liu Y, Chan KCC, Yang Q, Wang X (2021) Skeleton based human action evaluation using graph convolutional network for monitoring Alzheimer's progression. Pattern Recognit 119:Art. no. 108095
9. Parmar P, Morris BT (2017) Learning to score Olympic events. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp 76–84
10. Roditakis K, Makris A, Argyros A (2021) Towards improved and interpretable action quality assessment with self-supervised alignment. In: Proc. 14th Pervasive Technol. Rel. Assistive Environ. Conf., pp 507–513
11. Xu A, Zeng L-A, Zheng W-S (2022) Likert scoring with grade decoupling for long-term action assessment. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 3232–3241
12. Pirsiavash H, Vondrick C, Torralba A (2014) Assessing the quality of actions. In: Proc. Eur. Conf. Comput. Vis. Zurich, Switzerland: Springer, pp 556–571
13. Pan J-H, Gao J, Zheng W-S (2019) Action assessment by joint relation graphs. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp 6330–6339
14. Tang Y et al (2020) Uncertainty-aware score distribution learning for action quality assessment. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 9836–9845
15. Pan J-H, Gao J, Zheng W-S (2022) Adaptive action assessment. IEEE Trans Pattern Anal Mach Intell 44(12):8779–8795
16. Wang J, Du Z, Li A, Wang Y (2020) Assessing action quality via attentive spatio-temporal convolutional networks. In: Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV). Nanjing, China: Springer, pp 3–16
17. Lei Q, Zhang H, Du J (2021) Temporal attention learning for action quality assessment in sports video. Signal Image Video Process 15(7):1575–1583
18. Yu X, Rao Y, Zhao W, Lu J, Zhou J (2021) Group-aware contrastive regression for action quality assessment. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 7899–7908
19. Parmar P, Reddy J, Morris B (2021) Piano skills assessment. In: Proc. IEEE 23rd Int. Workshop Multimedia Signal Process. (MMSP), pp 1–5
20. Pazhooman H, Alamri MS, Pomeroy RL, Cobb SC (2023) Foot kinematics in runners with plantar heel pain during running gait. Gait Posture 104:15–21
21. Liu D et al (2021) Towards unified surgical skill assessment. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 9517–9526
22. Wang T, Wang Y, Li M (2020) Towards accurate and interpretable surgical skill assessment: a video-based method incorporating recog nized surgical gestures and skill levels. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Springer, Lima, pp 668–678
23. Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Badia SBI (2019) Learning to assess the quality of stroke rehabilitation exercises. In: Proc. 24th Int. Conf. Intell. User Interface, pp 218–228
24. Hamaguchi T et al (2020) Support vector machine-based classifier for the assessment of finger movement of stroke patients undergoing rehabilitation. J Med Biol Eng 40(1):91–100
25. Pogorelc B, Bosni·c Z, Gams M (2012) Automatic recognition of gait-related health problems in the elderly using machine learning. Multimedia Tools Appl 58(2):333–354
26. Lei Q, Li H, Zhang H, Du J, Gao S (2023) Multi-skeleton structures graph convolutional network for action quality assessment in long videos. Appl Intell 53(19):21692–21705
27. Jleli M, Samet B, Dutta AK (2024) Artificial intelligence-driven remote monitoring model for physical rehabilitation. Journal of Disability Research 3(1):20230065

28. Liu S, Zhang A, Li Y, Zhou J, Xu L, Dong Z, Zhang R (2021) Temporal segmentation of fine-grained semantic action: a motion centered figure skating dataset. AAAI Conf Artif Intell 35:2163–2171
29. Elkholy A, Hussein ME, Gomaa W, Damen D, Saba E (2020) Efficient and robust skeleton-based quality assessment and abnor mality detection in human action performance. IEEE J Biomed Health Inform 24(1):280–291. https://doi.org/10.1109/JBHI.2019.2904321
30. Hakim T, Shimshoni I (2019) A-mal: Automatic motion assess ment learning from properly performed motions in 3d skeleton videos. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 1589–1598
31. Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, Zhang F, Chang CL, Yong M, Lee J, Chang WT (2019) Mediapipe: a framework for perceiving and processing reality. In: Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR) (Vol. 2019)
32. Vakanski A, Jun HP, Paul D, Baker R (2018) A data set of human body movements for physical rehabilitation exercises. Data 3(1):2
33. Capecci M, Ceravolo MG, Ferracuti F, Iarlori S, Monteriu A, Romeo L, Verdini F (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
34. 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
35. Deb S, Islam MF, Rahman S, Rahman S (2022) Graph convolutional networks for assessment of physical rehabilitation exercises. IEEE Trans Neural Syst Rehabil Eng 30:410–419
36. Yao L, Lei Q, Zhang H, Du J, Gao S (2023) A contrastive learning network for performance metric and assessment of physical rehabilitation exercises. IEEE Trans Neural Syst Rehabil Eng. https://doi.org/10.1109/TNSRE.2023.3317411
37. Mourchid Y, Slama R (2023) September. MR-STGN: Multi-Residual Spatio Temporal Graph Network Using Attention Fusion for Patient Action Assessment. In: 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP). IEEE, pp 1–6
38. Sardari S, Sharifzadeh S, Daneshkhah A, Loke SW, Palade V, Duncan MJ, Nakisa B (2024) Lightpra: a lightweight temporal convolutional network for automatic physical rehabilitation exercise assessment. Comput Biol Med 173:108382
39. Zhang Q, Cheng G, Kang N (2023) A network structure for rehabilitation training evaluation based on transformer. In: 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP). IEEE, pp 234–237
40. Kanade A, Sharma M, Muniyandi M (2023) Attention-guided deep learning framework for movement quality assessment. In: ICASSP 20232023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 1–5
41. Mourchid Y, Slama R (2023) D-STGCNT: a dense spatio-temporal graph conv-GRU network based on transformer for assessment of patient physical rehabilitation. Comput Biol Med 165:107420
42. Wang K, Zhang J (2023) Spatio-temporal transformer model for skeleton-based rehabilitation exercises assessment. In: 2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, pp 188–192
43. He T, Chen Y, Wang L, Cheng H (2024) An expert-knowledge-based graph convolutional network for skeleton-based physical rehabilitation exercises assessment. IEEE Trans Neural Syst Rehabil Eng. https://doi.org/10.1109/TNSRE.2024.3400790
44. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. ArXiv Preprint arXiv:1603.05027
45. Targ S, Almeida D, Lyman K (2016) Resnet in resnet: generalizing residual architectures. ArXiv Preprint arXiv:1603.08029
46. Huang G, Sun Y, Liu Z, Sedra D, Weinberger K (2016) Deep networks with stochastic depth. ArXiv Preprint arXiv:1603.09382
47. Sergey Zagoruyko NK (2016) Deep networks with stochastic depth. ArXiv Preprint arXiv:1605.07146
48. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
49. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
50. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
51. Dai J, He K, Sun J (2015) Instance-aware semantic segmentation via multi-task network cascades. ArXiv Preprint arXiv:1512.04412
52. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99

53. Shen F, Gan R, Zeng G (2016) Weighted residuals for very deep networks. In: 2016 3rd international conference on systems and informatics (Icsai) (pp. 936–941). IEEE
54. Singh D, Singh B (2020) Investigating the impact of data normalization on classification performance. Appl Soft Comput 97:105524
55. Jain H, Harit G (2019) An unsupervised sequence-to-sequence autoen coder based human action scoring model. In: Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP), p. 15. https://doi.org/10.1109/GlobalSIP45357.2019.8969424
56. OReilly MA, Whelan DF, Ward TE, Delahunt E, Cauleld BM (2017) Classi cation of deadlift biomechanics with wearable inertial measurementunits. J Biomech 58:155161. https://doi.org/10.1016/j.jbiomech.2017.04.028
57. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inform Proc Syst 30
58. Song YF, Zhang Z, Shan C, Wang L (2020) Richly activated graph convolutional network for robust skeleton-based action recognition. IEEE Trans Circuits Syst Video Technol 31(5):1915–1925
59. Zhang P, Lan C, Zeng W, Xing J, Xue J, Zheng N (2020) Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1112–1121
60. Li C, Zhong Q, Xie D, Pu S (2018) Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv preprint arXiv:1804.06055
61. Chen Y, Xia S, Zhao J, Zhou Y, Niu Q, Yao R, Zhu D, Chen H (2023) Adversarial learning-based skeleton synthesis with spatial-channel attention for robust gait recognition. Multimedia Tools Appl 82(1):1489–1504
62. Karlov M, Abedi A, Khan SS (2024) Rehabilitation exercise quality assessment through supervised contrastive learning with hard and soft negatives. Med Biol Eng Comput. https://doi.org/10.1007/s11517-024-03177-x
63. Pan Z, Zhang J (2023) A self-supervised framework with a modified hop-layer network for rehabilitation exercise assessment. In: 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML) (pp. 274–279). IEEE
64. Sengar SS, Hasan AB, Kumar S, Carroll F (2024) Generative artificial intelligence: a systematic review and applications. Multimedia Tools Appl. https://doi.org/10.1007/s11042-024-20016-1
65. Ismail-Fawaz A, Devanne M, Berretti S, Weber J, Forestier G (2024) Weighted average of human motion sequences for improving rehabilitation assessment. In: International Workshop on Advanced Analytics and Learning on Temporal Data. Springer Nature Switzerland, Cham, pp 131–146

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.