A Unified Framework with Difference of Convex Decomposition and Sparse Topological Learning for Skeleton-Based Clinical Assessment

Quang Anh N.D. [1] , **Duc Minh Pham** [1] , **Minh-Anh Nguyen** [1] and **Hung Manh Ha** [1]
[1] International School, Vietnam National University, Hanoi, 100000, Vietnam
anhnd@vnuis.edu.vn, pmduc28@gmail.com @vnu.edu.vn hunghm@vnu.edu.vn

Abstract

xxx

1 Introduction

In recent decades, the global demand for physical rehabilitation services has increased dramatically, driven by demographic changes and the rising prevalence of chronic diseases [Jesus *et al.*, 2022]. It is estimated that in 2019, more than 2.4 billion people worldwide were living with health conditions that could benefit from rehabilitation services, a significant increase since 1990. This increase is largely due to factors such as an aging population and the increasing prevalence of non-communicable diseases, including musculoskeletal disorders, which have been identified as the leading cause of disability worldwide [Chen *et al.*, 2022]. Despite this growing demand, there remains a significant gap between the supply and demand for rehabilitation services, particularly in low and middle income countries where resources are often limited. This imbalance not only affects the quality of life of those requiring rehabilitation, but also imposes a significant economic burden on health systems globally. For example, a study by Soberg, H. L. *et al.* have examined rehabilitation demand and service provision highlighted the significant costs associated with rehabilitation in the first year after injury, underscoring the financial burden on health care infrastructure [Soberg *et al.*, 2022]. The integration of artificial intelligence (AI) into physical rehabilitation offers a promising avenue to address these challenges by improving the efficiency, accessibility, and personalization of care [Calderone *et al.*, 2024]. AI-based technologies, such as machine learning algorithms, can analyze complex patient data to develop individualized treatment plans, predict recovery trajectories, and monitor patient progress in real time. For example, AI applications have been used to improve motor function in patients with neurological disorders by tailoring interventions to specific needs. The application of AI in rehabilitation also extends to administrative and operational aspects, streamlining processes such as scheduling, documentation, and resource allocation. By automating these tasks, healthcare providers can devote more time and attention to direct patient care, ultimately improving service delivery and patient satisfaction.

2 Related Works

Several studies have highlighted the efficiency of deep learning in this domain. In 2020, Yao *et al.* proposed a deep learning framework for assessing the quality of physical rehabilitation exercises [Liao *et al.*, 2020a]. Their approach utilized performance metrics based on the log-likelihood of a Gaussian mixture model, combined with deep neural networks to generate quality scores. In recent years, Mennella *et al.* have introduced a novel system for home-based, remote, and unsupervised rehabilitation exercise monitoring, leveraging deep learning for real-time evaluation [Mennella *et al.*, 2023]. The system focuses on two components with range of motion (ROM) classification and compensatory pattern recognition, achieving mean accuracy of 89% and 98%, respectively, with a unique dataset of six resistance training exercises. Furthermore, Zhu *et al.* proposed a multipath convolutional neural network (MP-CNN), comprising a dynamic convolutional neural network (D-CNN) and a state transition probability CNN (S-CNN) [Zhu *et al.*, 2019]. The D-CNN uses Gaussian mixture models to capture sensor data distributions, while the S-CNN extracts transition probabilities using a modified Lempel–Ziv–Welch algorithm, achieving an average accuracy of 97.8% for recognition and 96.5% for evaluation. To further increase the advancement, this work further increase the robustness for skeleton-Based rehabilitation assessment, the main contributions of this work are as follows:

The main challenge lies in capturing the subtle biomechanical deviations that distinguish correct from compensatory movements. Traditional skeleton-based action recognition methods focus predominantly on discriminating between action categories, treating skeletal sequences as spatial-temporal graphs and applying graph convolutional networks to learn discriminative representations [Yan *et al.*, 2018b; Shi *et al.*, 2019]. However, movement quality assessment requires a fundamentally different paradigm: rather than learning categorical distinctions, we must capture fine-grained continuous variations in movement execution that correlate with clinical expertise. Existing approaches often rely solely on Euclidean geometric features, failing to capture the rich geometric structure inherent in skeletal motion. Furthermore, they typically employ single-stream architectures that cannot simultaneously model multi-scale spatial dependencies, geometric constraints, and topological relationships essential for comprehensive quality assessment.

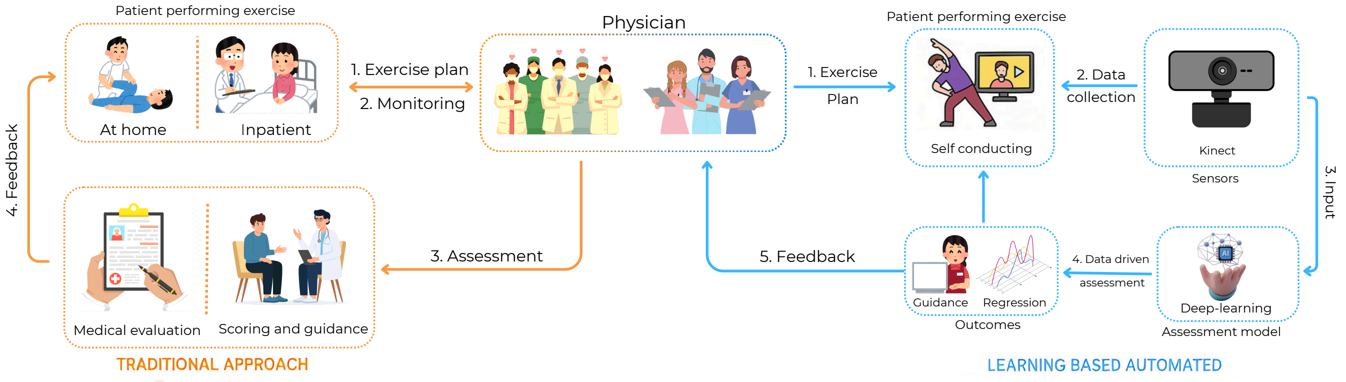

Figure 1: Enter Caption

Recent work in biomechanics and motor control theory suggests that movement quality emerges from the complex interplay of multiple factors: the geometric configuration of body segments, the differential relationships between joint positions, and the local topological structure of the kinematic chain [Bernstein, 1967]. This observation motivates our design of a multi-stream architecture that explicitly models these complementary aspects. Additionally, while graph neural networks have demonstrated success in capturing skeletal connectivity, they often treat all joints uniformly, neglecting the varying importance of different body regions in clinical assessment. Similarly, standard distance-based features employ simplistic metrics that do not exploit the underlying mathematical structure of distance functions, limiting their representational capacity.

3 Methodology

3.1 Problem Formulation and Notation

Using skeletal-based movement quality assessment in clinical rehabilitation presents a fundamental challenge in quantifying the quality of human motion based on temporal skeletal information. We begin by formally defining the skeletal representation and graph structure that underlies our approach. A human skeleton at time t is represented as a graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}, \mathbf{F}_t)$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denotes the set of N joints, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the natural skeletal connectivity defined by the human kinematic chain, and $\mathbf{F}_t \in \mathbb{R}^{N \times D}$ contains the feature vectors for all joints. Each joint v_i is associated with a feature vector $\mathbf{f}_{t,i} = [\mathbf{q}_{t,i}; \mathbf{p}_{t,i}] \in \mathbb{R}^D$, where $\mathbf{q}_{t,i} \in \mathbb{R}^4$ represents the quaternion orientation and $\mathbf{p}_{t,i} \in \mathbb{R}^3$ denotes the 3D spatial position in the coordinate system.

The adjacency matrix $\mathbf{A} \in \{0,1\}^{N \times N}$ encodes the skeletal connectivity, where $A_{ij} = 1$ if joints v_i and v_j are connected by a bone, and $A_{ij} = 0$ otherwise. We denote the degree matrix as $\mathbf{D}$, where $D_{ii} = \sum_j A_{ij}$, and we define the normalized adjacency matrix following by [Kipf and Welling, 2017] as $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. The temporal sequence of skeletons forms a spatial-temporal graph $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_T\}$. Given a temporal sequence of 3D skeleton data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times N \times D}$, where T denotes the number of temporal frames, N represents the number of skeletal joints, and D is the feature dimension encompassing both joint positions and orientations, our objective is to learn a mapping function $f : \mathbb{R}^{T \times N \times D} \rightarrow \mathbb{R}^K$ that predicts clinical assessment scores $\mathbf{y} \in \mathbb{R}^K$ reflecting movement quality across K evaluation criteria.

3.2 Overview of the Proposed Approach

We propose a novel three-stream GNN architecture that addresses the limitations of existing approaches through complementary feature extraction pathways, each designed to capture distinct aspects of movement quality. Our method, termed DC^2-STG (Difference of Convex-inspired Sparse Topological Graph Convolutional Network), integrates three parallel processing streams: (*i*) a skeleton-geometry stream with Selective Kernel (SK) attention [Li *et al.*, 2019] for adaptive multi-scale spatial feature learning, (*ii*) a geometric relationship stream employing difference-of-convex decomposition [Le Thi and Pham Dinh, 2005; Le Thi and Pham Dinh, 2018] with sparse regularization inspired by DC programming principles, and (*iii*) a sparse topological stream that captures local neighborhood structures through multi-scale k-nearest neighbor analysis.

The three streams operate in parallel on the input skeletal sequence, extracting features at different levels of abstraction. These features are subsequently fused through a learnable fusion layer, followed by temporal modeling via Gated Recurrent Units (GRU) to capture the sequential dynamics of movement execution. The entire architecture is trained end-to-end using movement quality scores annotated by clinical experts for the target regression assessment task. The proposed DC^2-STG architecture overview is illustrated in Figure **??**.

3.3 Skeleton-Geometry Processing Stream

The first stream processes the raw skeletal features through a sequence of spatial-temporal graph convolution layers augmented with adaptive multi-scale feature selection. We employ Shift-GCN [Cheng *et al.*, 2020] as the fundamental building block, which extends standard graph convolution by proposed an temporal shift operations to model motion dynamics effectively without additional parameters. To enable adaptive multi-scale spatial feature extraction, we utilized SK attention [Li *et al.*, 2019] following the graph convolution layers. The SK attention mechanism dynamically fuses features from multiple parallel convolution branches with different receptive fields, allowing the network to adaptively select the appropriate scale for each spatial location. Specifically, given M parallel convolution branches with kernel sizes $\mathcal{K} = \{k_1, k_2, \dots, k_M\}$, where we use $\mathcal{K} = \{1, 3, 5\}$, we obtain M feature representations through temporal convolutions along the graph-structured data:

$$\mathbf{U}_m = \mathrm{Conv}_{k_m}\left(\mathbf{X}^{(L)}\right), \quad m = 1, 2, \dots, M, \tag{1}$$

where $\mathbf{X}^{(L)}$ denotes the output from the final Shift-GCN layer. After that, we performed an feature fusion process through a two-stage process by compute global spatial-temporal statistics via global average pooling:

$$\mathbf{s} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{\tau N}\sum_{t=1}^{\tau}\sum_{i=1}^{N}\mathbf{U}_m(t, i, :), \tag{2}$$

where $\mathbf{s} \in \mathbb{R}^C$ aggregates information across all branches, time steps, and spatial locations. Subsequently, we compute branch-specific attention weights through a compact feature extraction and expansion process:

$$\mathbf{z} = \sigma(\mathbf{W}_{\text{fc}}\mathbf{s}), \quad \mathbf{a}_m = \mathrm{softmax}_m(\mathbf{W}_m\mathbf{z}), \tag{3}$$

where $\mathbf{W}_{\text{fc}} \in \mathbb{R}^{d \times C}$ reduces dimensionality to $d = \max(C/r, L)$ with reduction ratio $r = 8$ and minimum dimension $L = 32$, $\mathbf{W}_m \in \mathbb{R}^{C \times d}$ expands features for branch m, and the softmax operation ensures $\sum_{m=1}^{M} a_{m,c} = 1$ for each channel c. The final output combines all branches weighted by their attention:

$$\mathbf{X}_{\text{skel}} = \sum_{m=1}^{M}\mathbf{a}_m \odot \mathbf{U}_m, \tag{4}$$

where $\odot$ denotes element-wise multiplication with broadcasting. This design allows the network to emphasize features from the most informative receptive field scale for each spatial location and temporal instant, which is important for capturing both fine-grained local movements and coarse-grained global postural configurations.

3.4 DC-Inspired Geometric Relationship Stream

In this second stream, we propose a geometric feature extraction module inspired by the modeling intuition of Difference of Convex (DC) programming principles by Tao Pham Dinh & Hoai An Le Thi [Le Thi and Pham Dinh, 2005; Le Thi and Pham Dinh, 2018]. Rather than implementing the full iterative DC Algorithm (DCA) [Pham Dinh and Le Thi, 1997], we adopt the conceptual framework of DC decomposition for feature design and employ computationally efficient proximal gradient optimization [Parikh and Boyd, 2014; Beck and Teboulle, 2009] for sparse weight learning. This design is motivated by the observation that movement quality assessment inherently involves geometric constraints that benefit from the mathematical structure provided by convex decomposition.

The main ideas following by the explicit geometric modeling within the DC-inspired stream, which decomposes pairwise joint distances into complementary convex and concave components and applies proximal gradient optimization [Parikh and Boyd, 2014; Beck and Teboulle, 2009] with soft thresholding to learn sparse importance weights. By decomposing distance metrics into squared Euclidean and logarithmic components-functions with complementary growth properties-we enable the model to capture both local fine-grained deviations and global coarse-grained relationships simultaneously. This dual representation is particularly well-suited for clinical movement quality assessment, which fundamentally relies on geometric constraints at multiple scales.

Difference of Convex Inspire Feature Decomposition Giving 3D joint positions of X, the geometric relationships between joints are fundamentally characterized by pairwise distances. Inspired by DC programming [Le Thi and Pham Dinh, 2005], we decompose geometric features into two complementary functions that capture distinct mathematical properties of spatial relationships.

Definition 1 (DC-Inspired Feature Space)**.** *Let $\mathcal{P} = \{(\mathbf{p}_i, \mathbf{p}_j) : i < j, i, j \in [N]\}$ be the set of all unique joint pairs. We define the DC-inspired feature space as the Cartesian product $\Phi = \Phi_1 \times \Phi_2$, where:*

$$\Phi_1 = \{\phi_1(d) : d \in \mathbb{R}_+\}, \quad \Phi_2 = \{\phi_2(d) : d \in \mathbb{R}_+\}, \tag{5}$$

with $\phi_1(d) = d^2$ and $\phi_2(d) = \log(d + 1)$ being transformations mapping distances to complementary feature representations.

For any pair of joints (i, j), we define the pairwise distance as $d_{ij}^t = \|\mathbf{p}_{t,i} - \mathbf{p}_{t,j}\|_2$. We then construct two feature vectors by applying the transformations:

$$\phi_1(d_{ij}^t) = (d_{ij}^t)^2, \quad \phi_2(d_{ij}^t) = \log(d_{ij}^t + 1), \tag{6}$$

where the squared distance emphasizes local geometric deviations with quadratic sensitivity to large distances, while the logarithmic transformation provides robustness to outliers and better captures relative proportions between joint pairs through its compressive nature.

Lemma 1 (Properties of DC Feature Functions)**.** *The feature functions ϕ_1 and ϕ_2 have complementary mathematical properties:*

1. *$\phi_1(d) = d^2$ is strictly convex with $\nabla^2\phi_1(d) = 2 > 0$*
2. *$\phi_2(d) = \log(d + 1)$ is strictly concave with $\nabla^2\phi_2(d) = -\frac{1}{(d+1)^2} < 0$*
3. *ϕ_1 exhibits superlinear growth: $\lim_{d\to\infty}\frac{\phi_1(d)}{d} = \infty$*
4. *ϕ_2 exhibits sublinear growth: $\lim_{d\to\infty}\frac{\phi_2(d)}{d} = 0$*
5. *The functions are not affinely related, providing complementary geometric information*

Proposition 1 (Expressiveness of DC Features)**.** *The DC-inspired feature decomposition $\{\phi_1(d), \phi_2(d)\}$ provides complementary information about pairwise distances. Specifically, ϕ_1 and ϕ_2 are functionally independent: there exists no continuous function g such that $\phi_2(d) = g(\phi_1(d))$ for all*

$d \geq 0$. *This independence ensures that the feature space $\Phi_1 \times \Phi_2$ captures geometric relationships that neither function alone can represent.*

For computational efficiency, we consider only the upper triangular elements of the distance matrix, yielding $M = \frac{N(N-1)}{2}$ unique pairwise features. At time t, we construct the DC feature vector, operated by Eq (7).

$$\begin{aligned}\boldsymbol{\psi}_t = [&\phi_1(d_{12}^t), \ldots, \phi_1(d_{N-1,N}^t),\\ &\phi_2(d_{12}^t), \ldots, \phi_2(d_{N-1,N}^t)]^\top \in \mathbb{R}^{2M},\end{aligned} \tag{7}$$

representing a concatenation of squared distances and log-distances for all joint pairs. This decomposition naturally separates local and global geometric information while maintaining mathematical tractability for optimization.

Sparse Feature Weighting via Proximal Optimization

To learn the importance of different geometric relationships, we utilized an learnable weights $\mathbf{w} \in \mathbb{R}^{2M}$ that modulate the DC features, the weighted feature vector becomes $\tilde{\boldsymbol{\psi}}_t = \mathbf{w} \odot \boldsymbol{\psi}_t$. To prevent overfitting and encourage sparsity in the learned weights-allowing the model to select only the most clinically relevant geometric relationships-we formulate the weight learning as an elastic net regularized optimization problem [Zou and Hastie, 2005]:

$$\min_{\mathbf{w}} \mathcal{L}_{\text{task}}(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 + \mu \|\mathbf{w}\|_1, \tag{8}$$

where λ controls the ℓ_2 regularization strength for smoothness, and μ controls the ℓ_1 regularization strength for sparsity. By inspirations of DC programming [Le Thi and Pham Dinh, 2005], where the objective can be viewed through a DC lens via the equivalence $\|\mathbf{w}\|_1 = \max_{\mathbf{v} \in [-1,1]^{2M}} \mathbf{v}^\top \mathbf{w}$, expressing the non-smooth ℓ_1 norm through a max-over-convex-functions representation. Rather than implementing the full iterative DCA procedure [Pham Dinh and Le Thi, 1997], which would require solving a convex subproblem at each iteration, we employ the computationally efficient proximal gradient method [Parikh and Boyd, 2014; Beck and Teboulle, 2009] with soft thresholding to retains the sparsity-inducing benefits of DC optimization while being compatible with modern DL frameworks and gradient-based training.

Theorem 1 (Convergence of Proximal Gradient with Soft Thresholding). *Let $\mathcal{L}(\mathbf{w}) = \mathcal{L}_{task}(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 + \mu \|\mathbf{w}\|_1$ be the objective function with $\lambda, \mu > 0$. Assume $\mathcal{L}_{task}$ is L-smooth (i.e., $\nabla \mathcal{L}_{task}$ is L-Lipschitz continuous). The proximal gradient update:*

$$\mathbf{w}^{(k+1)} = \mathcal{S}_\theta \left((1 - 2\lambda\eta)\mathbf{w}^{(k)} - \eta \nabla \mathcal{L}_{task}(\mathbf{w}^{(k)}) \right), \tag{9}$$

where $\mathcal{S}_\theta(x) = sign(x) \max(|x| - \theta, 0)$ is the soft thresholding operator with $\theta = \frac{\mu\eta}{1+2\lambda\eta}$, converges to a stationary point of $\mathcal{L}(\mathbf{w})$ when the step size satisfies $\eta < \frac{1}{L+2\lambda}$.

Corollary 1 (Sparsity Guarantee). *The soft thresholding operator $\mathcal{S}_\theta$ induces exact zeros in the weight vector $\mathbf{w}$. Specifically, for any coordinate i, if $|(1 - 2\lambda\eta)w_i^{(k)} - \eta \nabla_i \mathcal{L}_{task}(\mathbf{w}^{(k)})| < \theta$, then $w_i^{(k+1)} = 0$. The expected number of non-zero weights decreases monotonically with increasing μ.*

In practice, we integrate soft thresholding into the forward pass as a regularization mechanism:

$$w_i^{\text{soft}} = \text{sign}(w_i) \cdot \max \left(|w_i| - \frac{\mu}{2\lambda + \epsilon}, 0 \right), \tag{10}$$

where $\epsilon = 10^{-8}$ is a small constant for numerical stability. The weights $\mathbf{w}$ are implemented as learnable parameters and are jointly optimized with other network parameters through backpropagation, with the soft thresholding providing an implicit sparsity-inducing regularization that selects clinically relevant geometric features.

Feature Normalization and Graph Convolution

After DC feature extraction and sparse weighting, we apply layer-wise feature normalization to stabilize training by $\hat{\boldsymbol{\psi}}_t = \tilde{\boldsymbol{\psi}}_t - \boldsymbol{\mu}_\psi / \boldsymbol{\sigma}_\psi + \epsilon$ where $\boldsymbol{\mu}_\psi$ and $\boldsymbol{\sigma}_\psi$ are the mean and standard deviation computed across both batch and temporal dimensions, then features are then projected to a higher-dimensional space through a MLP with layer norm by following:

$$\mathbf{h}_t^{\text{DC}} = \text{MLP}_{\text{DC}}(\hat{\boldsymbol{\psi}}_t) = \mathbf{W}_2 \cdot \sigma(\text{LN}(\mathbf{W}_1 \hat{\boldsymbol{\psi}}_t + \mathbf{b}_1)) + \mathbf{b}_2, \tag{11}$$

where $\mathbf{W}_1 \in \mathbb{R}^{2H \times 2M}$, $\mathbf{W}_2 \in \mathbb{R}^{H \times 2H}$ are learnable projection matrices, $\text{LN}(\cdot)$ denotes layer normalization, and H is the hidden dimension. This produces a temporal sequence of DC features $\mathbf{H}^{\text{DC}} = \{\mathbf{h}_1^{\text{DC}}, \ldots, \mathbf{h}_\tau^{\text{DC}}\} \in \mathbb{R}^{\tau \times H}$.

To enable graph-based spatial processing of these global geometric features, we expand the temporal features across the joint dimension and apply Shift-GCN layers by Eq (12).

$$\mathbf{X}_{\text{DC}} = \text{Shift-GCN}_2(\text{Shift-GCN}_1(\text{Expand}(\mathbf{H}^{\text{DC}}))), \tag{12}$$

where $\text{Expand} : \mathbb{R}^{\tau \times H} \to \mathbb{R}^{\tau \times N \times H}$ replicates the features across all N joints, creating tensors suitable for graph convolution that propagates the global geometric information through the skeletal structure.

3.5 Sparse Topological Neighborhood Stream

The third stream captures local topological structures by analyzing k-Nearest Neighbor (k-NN) relationships in the positional space. Unlike the skeleton connectivity $\mathcal{E}$ defined by anatomical constraints, k-NN topology reflects the actual spatial configuration and can reveal abnormal postures or compensatory movements where joints that are typically distant become spatially proximate [Bronstein *et al.*, 2017].

Multi-scale k-NN Feature Extraction

For each temporal frame t, we compute the pairwise Euclidean distance matrix $\mathbf{D}_t \in \mathbb{R}^{N \times N}$ between all joint positions, where $D_t(i,j) = \|\mathbf{p}_{t,i} - \mathbf{p}_{t,j}\|_2$. For each joint i, we identify its k nearest neighbors for multiple values of $k \in \mathcal{K} = \{k_1, k_2, k_3\}$, where typically $k_1 < k_2 < k_3$ (*e.g.*, $\mathcal{K} = \{2, 3, 4\}$) to capture multi-scale neighborhood structures ranging from immediate neighbors to broader local contexts.

For each scale $k \in \mathcal{K}$ and joint i, let $\mathcal{N}_k(i) = \{j_1, j_2, \ldots, j_k\}$ denote the set of k nearest neighbors. We compute three statistical features characterizing the local topology:

$$\mu_k^{(i)} = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} d_{ij}^t, \tag{13}$$

$$\sigma_k^{(i)} = \sqrt{\frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} (d_{ij}^t - \mu_k^{(i)})^2}, \tag{14}$$

$$\delta_k^{(i)} = \min_{j \in \mathcal{N}_k(i)} d_{ij}^t. \tag{15}$$

The mean distance $\mu_k^{(i)}$ reflects the overall spatial extent of the neighborhood, the standard deviation $\sigma_k^{(i)}$ captures the uniformity of neighbor distribution (low values indicate clustered neighbors while high values suggest dispersed configurations), and the minimum distance $\delta_k^{(i)}$ identifies the closest spatial relationship.

Definition 2 (Local Topological Descriptor)**.** *For a joint i at time t, we define the multi-scale topological descriptor as the concatenation:*

$$\mathbf{f}_t^{(i,\text{topo})} = [\mu_{k_1}^{(i)}, \sigma_{k_1}^{(i)}, \delta_{k_1}^{(i)}, \mu_{k_2}^{(i)}, \sigma_{k_2}^{(i)}, \delta_{k_2}^{(i)}, \mu_{k_3}^{(i)}, \sigma_{k_3}^{(i)}, \delta_{k_3}^{(i)}]^\top \in \mathbb{R}^{3|\mathcal{K}|}, \tag{16}$$

which characterizes the local topological context at multiple spatial scales.

Lemma 2 (Stability of k-NN Features)**.** *Let $\mathbf{P}_t$ and $\mathbf{P}_t'$ be two configurations of joint positions such that $\|\mathbf{p}_{t,i} - \mathbf{p}_{t,i}'\|_2 < \delta$ for all $i \in [N]$. Then the k-NN features are Lipschitz continuous with respect to position perturbations:*

$$|\mu_k^{(i)}(\mathbf{P}_t) - \mu_k^{(i)}(\mathbf{P}_t')| \leq 2\delta, \quad |\delta_k^{(i)}(\mathbf{P}_t) - \delta_k^{(i)}(\mathbf{P}_t')| \leq 2\delta. \tag{17}$$

Theorem 2 (Local Topological Preservation)**.** *The multi-scale k-NN descriptor $\mathbf{f}_t^{(i,topo)}$ preserves local topological structure. If two configurations $\mathbf{P}_t$ and $\mathbf{P}_t'$ satisfy $\mathbf{f}_t^{(i,topo)}(\mathbf{P}_t) = \mathbf{f}_t^{(i,topo)}(\mathbf{P}_t')$ for all $i \in [N]$ and all $k \in \mathcal{K}$, then the local neighborhoods $\mathcal{N}_k(i)$ have identical distance distributions. This ensures that local geometric relationships relevant for movement quality assessment are preserved.*

Learnable Scale Importance Weighting

For the model to adapt to the relative importance of different neighborhood scales-recognizing that different movement types may require different levels of spatial context-we introduce learnable importance weights $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]^\top$ implemented as trainable parameters. We apply a softmax constraint to ensure valid probability weights by $\alpha_k = \frac{\exp(\tilde{\alpha}_k)}{\sum_{k' \in \mathcal{K}} \exp(\tilde{\alpha}_{k'})}$, $\sum_{k \in \mathcal{K}} \alpha_k = 1$ where $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^{|\mathcal{K}|}$ are the unconstrained parameters. The weighted topological features become:

$$\mathbf{f}_t^{(i,\text{topo})} = \sum_{j=1}^{|\mathcal{K}|} \alpha_j \cdot [\mu_{k_j}^{(i)}, \sigma_{k_j}^{(i)}, \delta_{k_j}^{(i)}]^\top. \tag{18}$$

These per-joint features are then processed through Shift-GCN layers to capture spatial dependencies in the topological space:

$$\mathbf{X}_{\text{topo}} = \text{Shift-GCN}_2(\text{Shift-GCN}_1(\mathbf{F}_{\text{topo}})), \tag{19}$$

where $\mathbf{F}_{\text{topo}} \in \mathbb{R}^{\tau \times N \times 9}$ aggregates the topological features across time. The output is spatially pooled via global average pooling over the joint dimension to obtain temporal representations:

$$\mathbf{h}_t^{\text{topo}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_{\text{topo}}(t, i, :) \in \mathbb{R}^H. \tag{20}$$

3.6 Multi-Stream Fusion and Temporal Modeling

The three streams produce complementary feature representations of skeleton-geometry features $\mathbf{X}_{\text{skel}} \in \mathbb{R}^{\tau \times N \times C}$, DC-inspired geometric features $\mathbf{X}_{\text{DC}} \in \mathbb{R}^{\tau \times N \times H}$, and topological features $\mathbf{H}^{\text{topo}} \in \mathbb{R}^{\tau \times H}$, then we reshape the skeleton and DC features by flattening the spatial dimension, then concatenate all three streams:

$$\mathbf{X}_{\text{cat}} = [\omega(\mathbf{X}_{\text{skel}}); \omega(\mathbf{X}_{\text{DC}}); \mathbf{H}^{\text{topo}}] \in \mathbb{R}^{\tau \times D_{\text{cat}}}, \tag{21}$$

where ω denotes flattening operations and $D_{\text{cat}} = N \cdot C + N \cdot H + H$ is the concatenated dimension. These multi-stream features are fused through a learnable projection network:

$$\mathbf{H}_{\text{fused}} = \mathbf{W}_{\text{fuse2}} \cdot \sigma(\varphi(\text{LN}(\mathbf{W}_{\text{fuse1}} \mathbf{X}_{\text{cat}} + \mathbf{b}_1))) + \mathbf{b}_2, \tag{22}$$

where φ represents the Dropout layer, and $\mathbf{W}_{\text{fuse1}} \in \mathbb{R}^{2H \times D_{\text{cat}}}$, $\mathbf{W}_{\text{fuse2}} \in \mathbb{R}^{H \times 2H}$ are learnable fusion weights that reduce dimensionality to H while enabling non-linear interaction between streams.

The fused features $\mathbf{H}_{\text{fused}} \in \mathbb{R}^{\tau \times H}$ capture comprehensive spatial characteristics at each time step. To model the temporal dynamics of movement execution-which is crucial as movement quality depends not only on instantaneous postures but also on the smoothness and coordination of transitions-we employ GRU layers:

$$\mathbf{h}_t = \text{GRU}(\mathbf{H}_{\text{fused}}(t, :), \mathbf{h}_{t-1}), \tag{23}$$

where $\mathbf{h}_t \in \mathbb{R}^H$ represents the hidden state at time t. The GRU processes the sequence from $t = 1$ to $t = \tau$, accumulating temporal context through its gating mechanisms. The final hidden state $\mathbf{h}_\tau$ encapsulates the entire sequence's temporal dynamics and is passed through a linear output layer:

$$\hat{\mathbf{y}} = \mathbf{W}_{\text{out}} \mathbf{h}_\tau + \mathbf{b}_{\text{out}}, \tag{24}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{K \times H}$ and $\mathbf{b}_{\text{out}} \in \mathbb{R}^K$ produce the predicted assessment scores $\hat{\mathbf{y}} \in \mathbb{R}^K$ for K clinical evaluation criteria.

4 Experiments Environment

4.1 Dataset

To evaluate the performance of our proposed model, we utilize the KIMORE dataset [Capecci *et al.*, 2019], a well-organized collection of data made for studying movement and clinical scores in physical rehabilitation. Released in 2019, KIMORE includes detailed skeletal motion data from people doing specific rehabilitation exercises, providing joint positions and movement paths, along with clinical scores. The dataset has five different exercises labeled from Ex1 to Ex5, each focusing on certain motor skills and varying in difficulty. This range helps us test the model across different types of movements.

4.2 Environment

All experiments were conducted on a system running Windows 11 Pro with an Intel Core i5-12400F processor and 32GB of RAM. The model was implemented using Python 3.11.9 and PyTorch 2.3.0, leveraging GPU acceleration with an NVIDIA GeForce RTX 3060 and CUDA 12.6. Training and evaluation were performed using the Adam optimize with learning rate of 0.001 and Mean Squared Error (MSE) loss, utilizing the hardware parallel processing capabilities to efficiently process the KIMORE dataset skeleton-based graph data.

Hyperparameters are configured as follows: hidden dimension $H = 128$, number of Shift-GCN layers per stream $L_{\text{stream}} = 3$, GRU layers $L_{\text{GRU}} = 3$, dropout rate $p = 0.15$, initial learning rate $\eta_0 = 10^{-3}$, weight decay $\lambda_{\text{wd}} = 10^{-2}$, DC-DCA regularization parameters $\lambda = 10^{-3}$ and $\mu = 10^{-2}$, SK attention reduction ratio $r = 8$ with kernel sizes $\{1, 3, 5\}$, and k-NN scales $\mathcal{K} = \{2, 3, 4\}$. The batch size is set to 32, and training proceeds for a maximum of 1500 epochs with early stopping if validation RMSE does not improve for 100 consecutive epochs.

4.3 Training Objective and Optimization

The model is trained end-to-end using an simple mean squared error loss between predicted and ground-truth clinical scores:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{B}\sum_{b=1}^{B} \|\hat{\mathbf{y}}_b - \mathbf{y}_b\|_2^2, \tag{25}$$

where B is the batch size. The total loss incorporating the elastic net regularization becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda\|\mathbf{w}\|_2^2 + \mu\|\mathbf{w}\|_1, \tag{26}$$

where the regularization terms are applied only to the DC-inspired stream weights $\mathbf{w}$ to encourage sparse selection of clinically relevant geometric features.

We employ the AdamW optimizer with decoupled weight decay. The learning rate is adaptively adjusted using ReduceLROnPlateau scheduling, which monitors validation RMSE and reduces the learning rate by a factor $\gamma = 0.5$ when no improvement is observed for a patience period $P = 50$ epochs, ensuring convergence to high-quality local minima.

5 Results and Discussion

We compare the performance of our proposed model against several state-of-the-art (SOTA) approaches on the KIMORE dataset [Capecci *et al.*, 2019], as summarized in Table **??**. The results indicate that our model consistently achieves the lowest error values across all five exercises for the Root Mean Squared Error (RMSE), Mean Absolute Deviation (MAD), and Mean Absolute Percentage Error (MAPE) metrics demonstrating its superior accuracy and robustness in rehabilitation exercise assessment.

A Proof of Lemma 1

Proof. For $\phi_1(d) = d^2$, we have $\frac{d^2\phi_1}{dd^2} = 2 > 0$, establishing strict convexity. The superlinear growth follows from $\lim_{d\to\infty} \frac{d^2}{d} = \lim_{d\to\infty} d = \infty$.

For $\phi_2(d) = \log(d+1)$, the second derivative is $\frac{d^2\phi_2}{dd^2} = -\frac{1}{(d+1)^2} < 0$, establishing strict concavity. The sublinear growth follows from L'Hôpital's rule with $\lim_{d\to\infty} \frac{\log(d+1)}{d} = \lim_{d\to\infty} \frac{1}{d+1} = 0$.

The complementarity of these functions-quadratic amplification versus logarithmic compression-ensures that ϕ_1 emphasizes large-scale structural deviations while ϕ_2 captures fine-grained local relationships. This decomposition, inspired by DC programming principles where objectives are expressed as differences of convex functions [Le Thi and Pham Dinh, 2005] but unlike a genuine DC decomposition-which requires both functions to be convex, our formulation uses a convex–concave pair strictly for feature enrichment, providing an principled framework for geometric feature extraction with ϕ_2 itself is concave rather than convex in the distance variable. □

B Proof of Proposition 1

Proof. Suppose by contradiction that $\phi_2(d) = g(\phi_1(d))$ for some continuous function g, *i.e.*, $\log(d+1) = g(d^2)$. Differentiating both sides with respect to d by:

$$\frac{1}{d+1} = g'(d^2)\cdot 2d. \tag{27}$$

This would require $g'(x) = \frac{1}{2\sqrt{x}(\sqrt{x}+1)}$ for $x = d^2 \geq 0$. However, this expression is not defined at $x = 0$ (where $d = 0$), as it involves $\sqrt{x}$ in the denominator. Therefore, no continuous function g exists that relates ϕ_2 to ϕ_1 globally, establishing their functional independence.

The practical implication is that weighted combinations $w_1\phi_1(d) + w_2\phi_2(d)$ can adapt to different distance scales and sensitivities relevant for movement quality assessment, with the learned weights selecting the appropriate balance between quadratic and logarithmic emphasis. □

C Proof of Theorem 1

Proof. The proximal gradient method for the composite objective $\mathcal{L}(\mathbf{w}) = f(\mathbf{w}) + g(\mathbf{w})$ with $f = \mathcal{L}_{\text{task}} + \lambda\|\mathbf{w}\|_2^2$ (smooth part) and $g = \mu\|\mathbf{w}\|_1$ (non-smooth part) takes the form [Parikh and Boyd, 2014]:

$$\mathbf{w}^{(k+1)} = \text{prox}_{\eta g}\left(\mathbf{w}^{(k)} - \eta\nabla f(\mathbf{w}^{(k)})\right), \tag{28}$$

where the proximal operator of g is defined as:

$$\text{prox}_{\eta g}(\mathbf{z}) = \arg\min_{\mathbf{w}}\left\{\frac{1}{2\eta}\|\mathbf{w} - \mathbf{z}\|_2^2 + \mu\|\mathbf{w}\|_1\right\}. \tag{29}$$

Table 1: Performance comparison on Kimore Exercises 1–5. The best results are highlighted in **bold**.

Model	**MAD ↓**					**RMSE ↓**					**MAPE ↓**				
	Ex1	Ex2	Ex3	Ex4	Ex5	Ex1	Ex2	Ex3	Ex4	Ex5	Ex1	Ex2	Ex3	Ex4	Ex5
Song *et al.* [Song *et al.*, 2020]	0.977	1.282	1.033	0.715	1.536	2.165	3.345	1.679	2.018	3.198	2.605	5.202	2.968	3.599	4.959
Zhang *et al.* [Zhang *et al.*, 2020]	1.757	3.139	1.737	1.202	1.853	2.916	4.140	2.615	1.836	2.916	5.034	10.456	5.774	3.797	6.531
Liao *et al.* [Liao *et al.*, 2020b]	1.141	1.528	0.845	0.468	0.847	2.534	3.738	1.640	0.792	1.941	2.589	3.976	2.023	1.354	2.312
Yan *et al.* [Yan *et al.*, 2018a]	0.889	2.096	0.604	0.842	1.218	2.017	3.262	0.799	1.331	1.951	2.339	6.136	1.727	2.325	3.802
Li *et al.* [Li *et al.*, 2018]	1.378	1.877	1.452	0.675	1.662	2.344	2.823	2.034	1.078	2.575	3.491	5.298	4.188	2.130	3.752
Du *et al.* [Du *et al.*, 2015]	1.271	2.199	1.123	0.880	1.864	2.440	4.297	1.925	1.676	3.158	3.228	6.001	3.421	3.108	5.620
S. Deb *et al.* [Deb *et al.*, 2022]	0.799	0.774	0.369	0.347	0.621	2.024	2.130	0.856	0.644	1.181	1.926	1.472	0.728	1.222	1.591
RAST-G [Lim *et al.*, 2025]	0.225	0.227	0.231	**0.221**	0.220	0.267	0.268	0.274	**0.264**	**0.264**	0.364	0.370	0.385	0.346	0.349
FTF-HGCN [Zhang *et al.*, 2025]	0.622	0.491	**0.206**	0.204	0.390	1.378	0.748	**0.398**	0.515	0.698	1.508	0.952	**0.536**	**0.483**	1.113
STGCN-Seq [Kourbane *et al.*, 2025]	0.543	0.511	0.213	0.204	0.488	1.492	1.124	0.337	0.218	0.724	1.362	0.766	0.620	0.514	1.412
Jleli *et al.* [Jleli *et al.*, 2024]	0.482	0.521	0.389	0.478	0.489	1.026	1.102	1.206	1.054	0.996	1.112	1.205	1.098	0.981	1.108
Ours	**0.145**	**0.144**	0.269	0.268	**0.0001**	**0.303**	**0.361**	0.693	0.647	0.001	**0.330**	**0.330**	0.710	0.620	**0.001**

Table 2: Performance comparison (MAD ↓) on UI-PRMD Exercises 1–10. The best results are highlighted in **bold**.

Model	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6	Ex7	Ex8	Ex9	Ex10
Song *et al.* [Song *et al.*, 2020]	0.011	0.006	0.010	0.014	0.013	0.009	0.017	0.017	0.008	0.038
Zhang *et al.* [Zhang *et al.*, 2020]	0.022	0.003	0.016	0.016	0.003	0.003	0.021	0.025	0.027	0.066
Liao *et al.* [Liao *et al.*, 2020b]	0.011	0.028	0.029	0.012	0.019	0.012	0.018	0.023	0.023	0.042
Li *et al.* [Li *et al.*, 2018]	0.011	0.029	0.036	0.014	0.017	0.019	0.027	0.025	0.027	0.047
Du *et al.* [Du *et al.*, 2015]	0.030	0.077	0.137	0.036	0.064	0.047	0.193	0.073	0.065	0.160
S. Deb *et al.* [Deb *et al.*, 2022]	0.009	0.006	0.013	0.006	0.008	0.006	0.011	0.016	0.008	0.031
FTF-HGCN [Zhang *et al.*, 2025]	0.008	0.005	0.011	0.004	0.006	0.006	0.009	0.010	0.007	**0.010**
STGCN-Seq [Kourbane *et al.*, 2025]	**0.006**	0.008	**0.009**	0.006	**0.003**	**0.004**	0.009	0.013	**0.006**	0.028
Sardari *et al.* [Sardari *et al.*, 2024]	0.014	0.007	0.011	0.006	0.008	0.006	0.010	0.011	0.008	0.038
Mourchid *et al.* [Mourchid and Slama, 2023]	0.011	0.009	0.013	0.009	0.009	0.013	0.022	0.020	0.013	0.014
Yao *et al.* [Yao *et al.*, 2023]	0.015	0.012	0.015	0.008	0.009	0.010	0.011	0.018	0.010	0.044
PhysioFormer [Marusic *et al.*, 2024]	0.010	0.009	0.011	0.009	0.008	0.012	0.019	0.018	0.013	0.013
Ours	0.009	**0.006**	0.013	**0.006**	0.008	0.006	**0.011**	**0.010**	0.008	0.051

This proximal operator has a closed-form solution given by the soft thresholding operator [Donoho, 1995] with $\text{prox}_{\eta g}(\mathbf{z}) = \mathcal{S}_{\eta\mu}(\mathbf{z})$. For the gradient step, we have:

$$\nabla f(\mathbf{w}) = \nabla \mathcal{L}_{\text{task}}(\mathbf{w}) + 2\lambda \mathbf{w}. \tag{30}$$

Substituting into the proximal update:

$$\begin{aligned} \mathbf{w}^{(k+1)} &= \mathcal{S}_{\eta\mu}\left(\mathbf{w}^{(k)} - \eta(\nabla \mathcal{L}_{\text{task}}(\mathbf{w}^{(k)}) + 2\lambda \mathbf{w}^{(k)})\right) \qquad (31) \\ &= \mathcal{S}_{\eta\mu}\left((1 - 2\lambda\eta)\mathbf{w}^{(k)} - \eta \nabla \mathcal{L}_{\text{task}}(\mathbf{w}^{(k)})\right). \qquad (32) \end{aligned}$$

Rescaling by $(1 - 2\lambda\eta)^{-1}$ and adjusting the threshold gives the stated form with $\theta = \frac{\mu\eta}{1+2\lambda\eta}$.

Convergence follows from standard proximal gradient convergence theory of [Beck and Teboulle, 2009], since f is $(L + 2\lambda)$-smooth and g is convex (though non-smooth), the sequence $\{\mathbf{w}^{(k)}\}$ converges to a stationary point when $\eta < \frac{1}{L+2\lambda}$, ensuring the descent property $\mathcal{L}(\mathbf{w}^{(k+1)}) \leq \mathcal{L}(\mathbf{w}^{(k)})$ holds. □

D Proof of Lemma 2

Proof. By the triangle inequality, for any two joints i and j:

$$\begin{aligned} |d_{ij}^t - d_{ij}'^t| &= \big| \|\mathbf{p}_{t,i} - \mathbf{p}_{t,j}\|_2 - \|\mathbf{p}'_{t,i} - \mathbf{p}'_{t,j}\|_2 \big| \qquad (33) \\ &\leq \|(\mathbf{p}_{t,i} - \mathbf{p}_{t,j}) - (\mathbf{p}'_{t,i} - \mathbf{p}'_{t,j})\|_2 \qquad (34) \\ &\leq \|\mathbf{p}_{t,i} - \mathbf{p}'_{t,i}\|_2 + \|\mathbf{p}_{t,j} - \mathbf{p}'_{t,j}\|_2 < 2\delta. \qquad (35) \end{aligned}$$

For the mean feature:

$$\begin{aligned} |\mu_k^{(i)}(\mathbf{P}_t) - \mu_k^{(i)}(\mathbf{P}'_t)| &= \left| \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} (d_{ij}^t - d_{ij}'^t) \right| \\ &\leq \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} |d_{ij}^t - d_{ij}'^t| < \frac{1}{k} \cdot k \cdot 2\delta \\ &= 2\delta. \end{aligned} \tag{36}$$

The minimum distance follows similarly since $\min$ is 1-Lipschitz continuous. This stability property ensures that small perturbations in joint positions do not drastically change the topological features, providing robustness for clinical assessment. □

E Proof of Theorem 2

Proof. The equality of k-NN features across all scales implies:

1. Identical mean distances: $\sum_{j \in \mathcal{N}_k(i)} d_{ij}^t = \sum_{j \in \mathcal{N}_k(i)} d_{ij}'^t$ for all i and k
2. Identical standard deviations: The variance of distances within each k-neighborhood matches
3. Identical minimum distances: $\min_{j \in \mathcal{N}_k(i)} d_{ij}^t = \min_{j \in \mathcal{N}_k(i)} d_{ij}'^t$

Together, these constraints ensure that the local distance structure within k-neighborhoods is identical between the two configurations. While this does not guarantee global isometry (which would require all $\binom{N}{2}$ pairwise distances to match), it ensures that the local coordination patterns between nearby joints are preserved.

For clinical movement quality assessment, local joint coordination is more diagnostically relevant than global body configuration [Bernstein, 1967], making this local preservation property both sufficient and appropriate for the rehabilitation exercise evaluation task. □

F Evaluation Metrics

We evaluated the DC^2-STG model using three standard metrics of RMSE, MAD, and MAPE metrics as expressed in Eq (37) - Eq (38).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2} \tag{37}$$

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}| \tag{38}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y - \hat{y}|}{y} \times 100 \tag{39}$$

where y represents actual values, $\hat{y}$ denotes predictions, and n is the sample size. RMSE emphasizes larger errors through squaring, making it sensitive to significant deviations in prediction precision. MAD treats all errors equally by measuring average absolute differences without amplification. MAPE expresses errors as percentages of actual values, enabling comparison across exercises with different motion scales.

Ethical Statement

There are no ethical issues.

Acknowledgments

References

[Beck and Teboulle, 2009] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[Bernstein, 1967] Nikolai Bernstein. *The co-ordination and regulation of movements*. Pergamon Press, 1967.

[Bronstein *et al.*, 2017] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[Calderone *et al.*, 2024] Antonio Calderone, Daniele Latella, Mario Bonanno, Angelo Quartarone, Sepideh Mojdehdehbaher, Antonio Celesti, and Rocco Salvatore Calabrò. Towards transforming neurorehabilitation: The impact of artificial intelligence on diagnosis and treatment of neurological disorders. *Biomedicines*, 12(10):2415, 2024.

[Capecci *et al.*, 2019] Michela Capecci, Maria Grazia Ceravolo, Francesco Ferracuti, Stefano Iarlori, Andrea Monteriu, Luca Romeo, and Federica Verdini. The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7):1436–1448, 2019.

[Chen *et al.*, 2022] Nuo Chen, Daniel YT Fong, and Joseph YH Wong. Secular trends in musculoskeletal rehabilitation needs in 191 countries and territories from 1990 to 2019. *JAMA Network Open*, 5(1):e2144198, 2022.

[Cheng *et al.*, 2020] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020.

[Deb *et al.*, 2022] S. Deb, M. F. Islam, S. Rahman, and S. Rahman. Graph convolutional networks for assessment of physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:410–419, 2022.

[Donoho, 1995] David L Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[Du *et al.*, 2015] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015.

[Jesus *et al.*, 2022] Tiago S. Jesus, Michel D. Landry, and Helen Hoenig. Is physical rehabilitation need associated with the rehabilitation workforce supply? an ecological study across 35 high-income countries. *International Journal of Health Policy and Management*, 11(4):434–441, 2022.

[Jleli *et al.*, 2024] Mohamed Jleli, Bessem Samet, and Ashit Kumar Dutta. Artificial intelligence-driven remote monitoring model for physical rehabilitation. *Journal of Disability Research*, 2024.

[Kipf and Welling, 2017] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[Kourbane *et al.*, 2025] Ikram Kourbane, Panagiotis Papadakis, and Mihai Andries. Optimized assessment of physical rehabilitation exercises using spatiotemporal, sequential graph-convolutional networks. *Computers in Biology and Medicine*, 186:109578, 2025.

[Le Thi and Pham Dinh, 2005] Hoai An Le Thi and Tao Pham Dinh. The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1):23–46, 2005.

[Le Thi and Pham Dinh, 2018] Hoai An Le Thi and Tao Pham Dinh. Dc programming and dca: thirty years of developments. *Mathematical Programming*, 169(1):5–68, 2018.

[Li *et al.*, 2018] C. Li, Q. Zhong, D. Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018.

[Li *et al.*, 2019] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019.

[Liao *et al.*, 2020a] Y. Liao, A. Vakanski, and M. Xian. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):468–477, 2020.

[Liao *et al.*, 2020b] Y. Liao, A. Vakanski, and M. Xian. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):468–477, 2020.

[Lim *et al.*, 2025] Suhyeon Lim, Ye-Eun Kim, and Andrew J. Choi. AI-Based Stroke Rehabilitation Domiciliary Assessment System with ST$_GCNAttention, 92025$.

[Marusic *et al.*, 2024] Aleksa Marusic, Sao Mai Nguyen, and Adriana Tapus. PhysioFormer: A Spatio-Temporal Transformer for Physical Rehabilitation Assessment. In *HRI '23: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Odense, Denmark, October 2024. Lecture Notes in Artificial Intelligence series (LNAI).

[Mennella *et al.*, 2023] C. Mennella, U. Maniscalco, G. De Pietro, and M. Esposito. A deep learning system to monitor and assess rehabilitation exercises in home-based remote and unsupervised conditions. *Computers in Biology and Medicine*, 166:107485, 2023.

[Mourchid and Slama, 2023] Youssef Mourchid and Rim Slama. D-stgcnt: A dense spatio-temporal graph conv-gru network based on transformer for assessment of patient physical rehabilitation. *Computers in Biology and Medicine*, 165:107420, 2023.

[Parikh and Boyd, 2014] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[Pham Dinh and Le Thi, 1997] Tao Pham Dinh and Hoai An Le Thi. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

[Sardari *et al.*, 2024] Sara Sardari, Sara Sharifzadeh, Alireza Daneshkhah, Seng W Loke, Vasile Palade, Michael J Duncan, and Bahareh Nakisa. Lightpra: A lightweight temporal convolutional network for automatic physical rehabilitation exercise assessment. *Computers in Biology and Medicine*, 173:108382, 2024.

[Shi *et al.*, 2019] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019.

[Soberg *et al.*, 2022] Helene Lundgaard Soberg, Høgne Østerholt Moksnes, Audny Anke, Olav Røise, Cecilie Røe, Eline Aas, and Nada Andelic. Correction: Rehabilitation needs, service provision, and costs in the first year following traumatic injuries: protocol for a prospective cohort study. *JMIR Research Protocols*, 11(3):e37723, 2022.

[Song *et al.*, 2020] Y. F. Song, Z. Zhang, C. Shan, and L. Wang. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1915–1925, 2020.

[Yan *et al.*, 2018a] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Yan *et al.*, 2018b] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Yao *et al.*, 2023] Long Yao, Qing Lei, Hongbo Zhang, Jixiang Du, and Shangce Gao. A contrastive learning network for performance metric and assessment of physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3790–3802, 2023.

[Zhang *et al.*, 2020] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1112–1121, 2020.

[Zhang *et al.*, 2025] Shaohui Zhang, Qiuying Han, Peng Wang, and Junjie Li. Frame topology fusion-based hierarchical graph convolution for automatic assessment of physical rehabilitation exercises. *Scientific Reports*, 15(1):26720, 7 2025.

[Zhu *et al.*, 2019] Zheng-An Zhu, Yu-Chi Lu, Chih-Hsuan You, and Chih-Kang Chiang. Deep learning for sensor-based rehabilitation exercise recognition and evaluation. *Sensors*, 19(4):887, 2019.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.