Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

CD-STGCN: A Novel Image-to-Graph Dual-Stream Spatio-Temporal Network for Structurally Interpretable Clinical Movement Quality Assessment

Diyar M. Witefee * and Aseel Hussein Zahi

Ministry of Education, General Directorate of Education in Babylon, Babylon 51001, Iraq;
e-mail : dyarz2017@gmail.com; aseel1371983@gmail.com
* Corresponding Author : Diyar M. Witefee

Abstract: In medical imaging, automated assessment of human movement quality remains challenging because many existing computer vision approaches rely on rigid anatomical priors that fail to capture complex spatio-temporal coordination patterns. This limitation is particularly relevant for neurological rehabilitation, where objective motion analysis is needed to monitor patient progress and support remote telerehabilitation. To address this problem, we propose Clustering-based Dual-Stream Spatio-Temporal Graph Convolutional Network (CD-STGCN), a novel image-to-graph framework that dynamically discovers latent coordination patterns through unsupervised joint clustering. Unlike conventional graph-based approaches with fixed anatomical structures, the proposed method learns adaptive functional joint relationships directly from movement data. The framework integrates 3D skeleton extraction from RGB/depth clinical recordings, followed by a dual-stream architecture that captures both spatial posture and temporal motion dynamics from skeletal sequences. Latent Motion Clustering, implemented using a Student's t-distribution kernel, enables data-driven functional joint grouping without explicit anatomical supervision, while dynamic graph construction generates sample-adaptive adjacency matrices. A multi-task learning strategy is further employed for simultaneous clinical score regression and exercise classification. Evaluation on the KIMORE dataset (n = 78 stroke survivors) demonstrates the effectiveness of the proposed framework. CD-STGCN achieved a mean absolute error of 0.098 (95% CI: 0.089–0.107), outperforming CTR-GCN (MAE = 0.121), corresponding to a 19% improvement (Cohen's d = 1.24, $p < 0.001$). The proposed method also achieved real-time performance at 83 FPS compared with 35 FPS reported by recent rehabilitation-oriented approaches. Ablation analysis showed that removing the Latent Motion Clustering module reduced performance by 20.4% (MAE increased from 0.098 to 0.118, $p < 0.001$). These findings provide proof-of-concept evidence that adaptive, data-driven coordination modeling may offer advantages over static anatomical priors for clinical movement assessment, although larger and more diverse studies are still required before clinical deployment.

Keywords: Clinical movement assessment; Deep learning; Graph convolutional networks; Medical image processing; Rehabilitation; Skeleton-based action recognition; Spatio-temporal modeling; Telerehabilitation.

Received: March, 25th 2026
Revised: May, 5th 2026
Accepted: May, 5th 2026
Published: May, 16th 2026
Curr. Ver.: May, 16th 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Globally, more than 80 million people are affected by neurological conditions such as stroke and Parkinson's disease, many of which require precise assessment of movement quality during rehabilitation [1], [2]. Objective movement evaluation is essential for tracking patient progress, personalizing treatment, and supporting remote monitoring in telerehabilitation settings [2], [3]. Present clinical practice commonly depends on standardized assessment scales, such as the Fugl–Meyer Assessment and the Unified Parkinson's Disease Rating Scale.

They are administered by trained physiotherapists with evaluations typically requiring 25–30 minutes per patient, depend heavily on specialist expertise, and remain subject to inter-rater variability [2], [3].

A forward development in depth sensing and pose estimation has permitted marker-less 3D skeleton extraction from medical RGB/depth data, creating new opportunities for automated movement analysis [4], [5]. Though translating skeletal motion sequences into clinically meaningful representations remains challenging, traditional biomechanical techniques fail to capture subtle pathological coordination patterns associated with neurological disorders [6]. Spatial-Temporal Graph Convolutional Networks (ST-GCNs), which have demonstrated strong performance in action recognition tasks [7], offer a promising direction for clinical movement analysis.

First, most ST-GCN variants used fixed anatomical priors based on physical proximity between joints. Such assumptions may be inadequate for neurological disorders, where abnormal compensatory strategies can cause anatomically distant joints to move in coordinated patterns [8], [9]. Stroke patients may recruit shoulder muscles during pelvic movement, producing coordination patterns that are difficult to capture using static graph structures. To address this limitation, we investigate clustering-based structural modeling that enables adaptive, data-driven joint grouping.

Second, many existing methods insufficiently model temporal dynamics, although clinical symptoms such as bradykinesia, tremor, and freezing emerge through complex spatio-temporal motion variations [10]. Capturing both posture and motion evolution is therefore critical for reliable movement assessment. In this study, we hypothesize that dual-stream feature extraction can better represent pathological movement dynamics than single-stream architectures [10], [11].

Third, clinical adoption requires models with interpretable structural representations rather than purely black-box predictions [12], [13]. While post-hoc explainability techniques such as SHAP or attention visualization can highlight influential features, they do not explicitly represent functional coordination patterns among body joints. Therefore, this work explores structural interpretability through latent motion clustering (LMC), enabling visualization of learned coordination patterns that may support clinical interpretation while remaining exploratory and dataset-specific [12]–[14].

To address these challenges, this paper proposes CD-STGCN, a Clustering-based Dual-Stream Spatio-Temporal Graph Convolutional Network for automated clinical movement assessment. Unlike conventional ST-GCN approaches with fixed anatomical hierarchies, the proposed framework learns adaptive coordination patterns directly from skeletal motion data. The proposed LMC module dynamically groups joints according to their kinematic relationships using a Student's t-distribution-based clustering mechanism, enabling pathology-sensitive graph construction without explicit anatomical supervision. The main contributions of this work are summarized as follows:

- A clustering-enhanced ST-GCN framework is proposed to model adaptive and pathology-sensitive joint coordination patterns through a Student's t-distribution-based LMC mechanism, reducing reliance on fixed anatomical priors.
- A dual-stream spatio-temporal feature extraction strategy. It is introduced to jointly capture spatial posture information and temporal motion dynamics for clinical movement assessment.
- A dynamic graph construction method is developed to generate sample-adaptive adjacency representations that reflect functional coordination patterns in neurological movement data.
- An interpretable graph-based clinical movement evaluation framework is presented to support visualization of learned coordination structures and facilitate exploratory analysis of pathological movement behavior in rehabilitation settings.

The rest of the paper is ordered as follows. Section 2 reviews related work on clinical movement analysis and graph-based action recognition, Section 3 describes the proposed CD-STGCN framework, Section 4 presents the experimental setup and results, Section 5 discusses the findings and limitations of the study, and Section 6 concludes the paper and outlines future research directions.

2. Related Work

Current literature in skeleton-based action recognition mainly depends on handcrafted motion descriptors, such as Lie group representations [15] and joint quadruple features [5] which promising results in constrained environments. Generally, they suffered from limited generalization capability and sensitivity to movement variability. Modern Subsequent deep learning methods introduced convolutional neural networks for skeleton action analysis [16]. Many of these approaches represented skeletal sequences as pseudo-images or vector sequences, thereby overlooking the inherent graph structure of human kinematics.

Spatial-Temporal Graph Convolutional Networks (ST-GCNs) were suggested to address these issues. ST-GCNs utilize explicitly modeled skeletons as graphs, where joints are represented as nodes and anatomical connections as edges [7]. Several extensions have since improved graph representation learning. AGCN [17] introduced adaptive topology learning. While MS-G3D [11] incorporated multi-scale disentangled graph convolutions for richer spatial-temporal modeling. CTR-GCN [18] further enhances channel-wise topology refinement. Despite these advances, most existing methods still rely on fixed anatomical priors or globally learned graph structures that are shared across samples. Such assumptions may be insufficient for pathological movement analysis, where compensatory motion patterns can produce abnormal coordination between anatomically distant joints [8], [9].

GNNs have also exhibited increasing potential in broader medical imaging applications, including medical image processing [19], histopathology [20], and musculoskeletal imaging [21], [22]On the other hand, in movement analysis, J. Zhu et al. [23] proposed a graph convolutional framework based on functional connectivity networks. However, the method relied on precomputed graph structures rather than dynamically learned coordination patterns. Hierarchical graph pooling approaches, such as differentiable pooling [14] and graph pooling [24] perform graph coarsening through learned assignments, but these techniques were primarily designed for static graph classification rather than spatial-temporal rehabilitation analysis.

In the rehabilitation assessment, conventional systems often rely on expensive marker-based motion capture and handcrafted biomechanical features [25], [26]. The introduction of low-cost depth sensors, such as Microsoft Kinect, enabled 3D skeleton markerless tracking and facilitated applications in balance assessment [27], gait analysis [28], and up-per-extremity rehabilitation evaluation [29]. For that reason, many existing methods still rely on simplified metrics that may fail to capture subtle pathological coordination patterns.

Table 1. Comparison among State-of-the-Art methods.

Method	Year	Key Innovation	Graph Structure	FPS	Main Limitation
AGCN [17]	2019	Bilateral adaptive topology	Learned global topology	38	Limited sample-specific adaptation
MS-G3D [11]	2020	Multi-scale disentangled convolution	Fixed hierarchical topology	42	Fixed scales may overlook compensatory motion patterns
CTR-GCN [18]	2021	Channel-wise topology refinement	Channel-adaptive topology	51	Relies on globally shared graph structures
D-STGCNT [30]	2024	Transformer + Conv-GRU	Full attention-based modeling	35	High computational complexity and limited real-time efficiency
BlockGCN [31]	2024	Spatial block decomposition	Fixed block-based graph	55	Rigid graph partitioning for pathological motion
DeGCN [32]	2024	Graph decomposition	Decomposed fixed topology	68	Limited adaptability to subject-specific coordination patterns

Recent deep learning approaches have further advanced rehabilitation-oriented movement analysis. CNN-based models have been explored for Parkinson's disease severity classification [33], while LSTM architectures have been applied to stroke exercise recognition [34]. However, these methods typically process skeletons as temporal vector sequences and do not explicitly model spatial relationships among joints. More recently, In [30] proposed

DSTGCNT, a transformer-based rehabilitation assessment framework that achieved improved predictive performance. However, the quadratic computational complexity of self-attention limits real-time applicability, with reported inference speed around 35 FPS.

Modern studies have also researched efficient graph operations. BlockGCN [31] Introduced spatial block decomposition to improve graph convolution efficiency. DeGCN [32] employed strategies of graph decomposition to execute lightweight real-time architectures. These studies, similar to CTR-GCN [18]. Table 1 outlines representative graph-based skeleton action recognition and rehabilitation assessment approaches, highlighting their architectural characteristics and limitations

Table 1 summarizes the most existing methods that either rely on fixed anatomical priors or employ globally shared adaptive topologies that are not explicitly optimized for pathological movement variability. Many approaches prioritize action recognition precision without addressing structural interpretability or clinically meaningful coordination analysis. These limitations motivate the development of a framework capable of learning adaptive, sample-specific coordination structures directly from rehabilitation motion data while maintaining computational efficiency for real-time clinical applications.

3. Proposed Method

The overall architecture of the proposed CD-STGCN framework is illustrated in Figure 1. The framework consists of five major stages: (1) medical image feature extraction, (2) dual-stream spatio-temporal feature extraction, (3) Latent Motion Clustering (LMC), (4) adaptive graph convolution, and (5) multi-task prediction heads for clinical score regression and exercise classification.

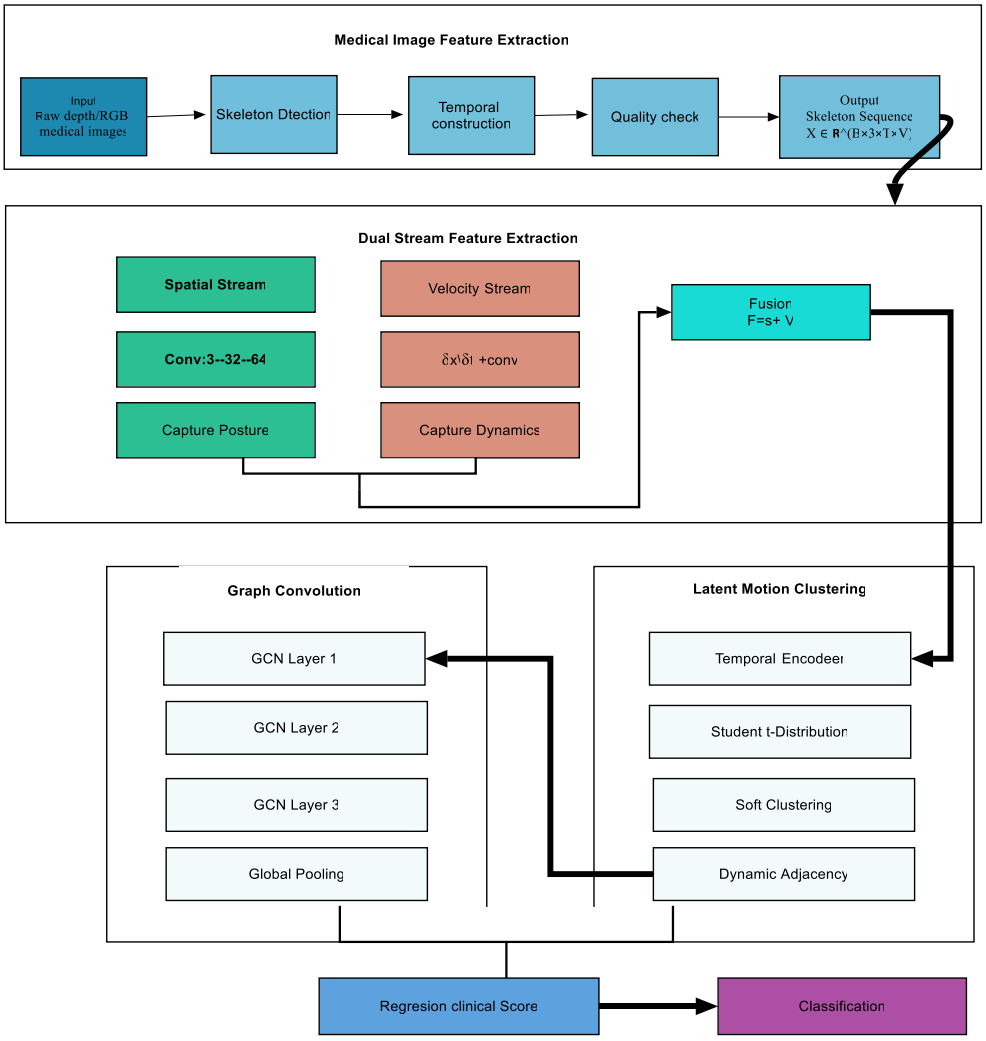

Figure 1. Overall architecture of the proposed CD-STGCN framework.

3.1. Medical Image Feature Extraction

Given raw RGB/depth medical images captured using Microsoft Kinect Azure and Intel RealSense devices, the preprocessing pipeline converts unstructured image sequences into structured skeleton-based spatial-temporal representations. The complete preprocessing workflow is illustrated in Figure 2.

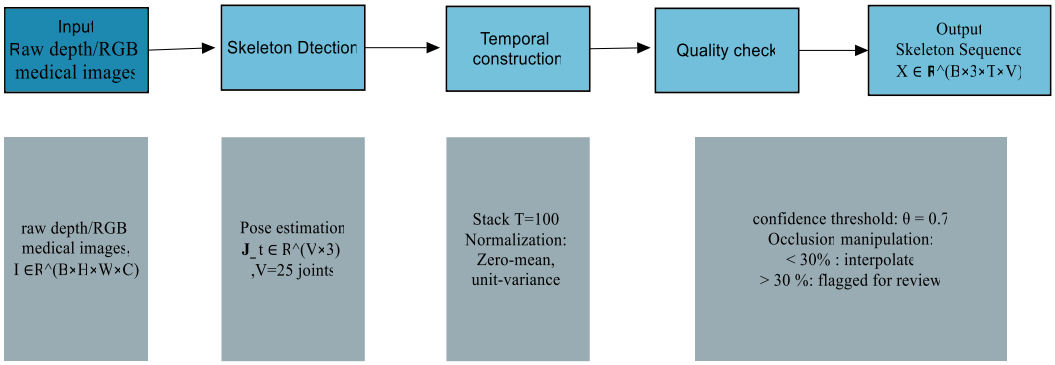

Figure 2. Medical image preprocessing and skeleton sequence construction pipeline.

Let the input image sequence be represented as:

$$I \in \mathbb{R}^{B\times H\times W\times C} \tag{1}$$

where B denotes the batch size, H and W represent image height and width, and C denotes the image channels.

Stage 1: Skeleton Detection and 3D Pose Estimation

For each frame I_t a pretrained pose estimation network extracts 3D joint coordinates:

$$J_t \in \mathbb{R}^{V\times 3} \tag{2}$$

where $V = 25$ represents the number of skeletal joints and the three channels correspond to (x, y, z) coordinates.

Stage 2: Temporal Sequence Construction

A temporal skeleton sequence is constructed by stacking $T = 100$ consecutive frames:

$$X = [J_1, J_2, \dots, J_T] \tag{3}$$

The skeletal coordinates are normalized using zero-mean and unit-variance normalization for each joint coordinate. Data augmentation is further applied using random temporal shifts (±5 frames) and Gaussian noise with standard deviation $\sigma = 0.01$.

Stage 3: Quality Assurance

To improve robustness against noisy clinical recordings, joint detections with confidence values below $\theta = 0.7$ are filtered. Absent joints with less than 30% occlusion are reconstructed using linear interpolation, while sequences with severe occlusion (>30%) are flagged for manual inspection. skeleton representation is defined as:

$$X \in \mathbb{R}^{B\times 3\times T\times V} \tag{4}$$

where T denotes the temporal dimension, and V represents the skeletal joints.

3.2. Dual-Stream Spatio-Temporal Feature Extraction

The proposed framework utilized a dual-stream representation to separately model spatial posture information and temporal motion dynamics. The spatial stream processes raw joint coordinates, and the temporal stream captures motion variations through the first-order velocity differences.

Spatial Stream

Given the skeleton sequence X, spatial posture features are extracted using temporal convolution layers:

$$F_s = \mathrm{Conv}_2\big(\mathrm{Conv}_1(X)\big) \tag{5}$$

where $\mathrm{Conv}_1: 3 \to 32$ and $\mathrm{Conv}_2: 32 \to 64$ use kernel size (3,1).

Velocity Stream

Temporal motion dynamics are computed using first-order frame differences:

$$V_t = X_t - X_{t-1} \tag{6}$$

The resulting velocity features are processed using identical convolutional layers:

$$F_v = \mathrm{Conv}_2\big(\mathrm{Conv}_1(X)\big) \tag{7}$$

Feature Fusion

The spatial and temporal representations are fused using element-wise addition:

$$F = F_s + F_v \tag{8}$$

where $F \in \mathbb{R}^{B\times 64\times T\times V}$ denotes the fused spatio-temporal feature representation.

3.3. Latent Motion Clustering (LMC)

The proposed LMC module dynamically learns functional coordination patterns directly from skeletal motion sequences without predefined anatomical grouping. The LMC module generates sample-adaptive adjacency matrices that capture latent coordination structures associated with pathological movement behavior.

Temporal Encoding

For each joint v, temporal features are encoded into compact latent representations:

$$z_v = \mathrm{Encoder}(F_v) \tag{9}$$

where $z_v \in \mathbb{R}^d$ denotes the encoded latent representation.

Student's t-Distribution-Based Clustering

The similarity between joint representations and learnable cluster centroids μ_j is computed using a Student's t-distribution kernel:

$$q_{ij} = \left(1 + \frac{\left\|z_i - \mu_j\right\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} \tag{10}$$

The soft assignment probabilities are normalized across clusters:

$$\hat{q}_{ij} = \frac{\exp(q_{ij})}{\sum_{k=1}^{K}\exp(q_{ik})} \tag{11}$$

where $K = 5$ denotes the number of latent coordination clusters.

Student's t-distribution is adopted due to its robustness to noisy and highly variable clinical motion measurements [14], [24].

Dynamic Adjacency Construction

Sample-specific adjacency matrices are constructed from cluster assignments using outer-product operations:

$$A^{(k)} = q_{:,k} q_{:,k}^T \tag{12}$$

where $A^{(k)} \in \mathbb{R}^{V\times V}$ represents the adjacency matrix for cluster k. Algorithm 1 summarizes the complete LMC procedure.

Algorithm 1. Latent Motion Clustering

INPUT: $F \in \mathbb{R}^{B\times 64\times T\times V}$, number of clusters $K = 5$

OUTPUT: $q \in \mathbb{R}^{B\times V\times K}$, adjacency matrices $\{A^{(k)}\}_{k=1}^{K}$

1: Extract joint-wise temporal features
2: Encode temporal trajectories into latent vectors
3: Compute Student's t-distribution similarities
4: Normalize cluster assignments using SoftMax
5: Construct cluster-specific adjacency matrices
6: Return adjacency representations

3.4. Adaptive Graph Convolution

Graph convolution layers update joint representations by aggregating information through dynamically learned adjacency structures. The adjacency matrices are first normalized as:

$$\tilde{A}^{(k)} = \Lambda_k^{-1/2} A^{(k)} \Lambda_k^{-1/2} \tag{13}$$

where Λ_k denotes the degree matrix.
A learnable modulation parameter β_k is further applied:

$$\hat{A}^{(k)} = \tilde{A}^{(k)} \odot \sigma(B_k) \tag{14}$$

Graph feature propagation is then computed as:

$$f_{\text{out}}^{(k)} = \hat{A}^{(k)} f_{\text{in}} W^{(k)} \tag{15}$$

where $W^{(k)}$ denotes the learnable graph convolution weights.
The outputs across clusters are aggregated as:

$$f_{\text{out}} = \sum_{k=1}^{K} f_{\text{out}}^{(k)} \tag{16}$$

3.5. Multi-Task Prediction Heads

The final graph representation is processed using a multi-task prediction framework consisting of parallel regression and classification branches. The overall prediction head architecture is shown in Figure 3.

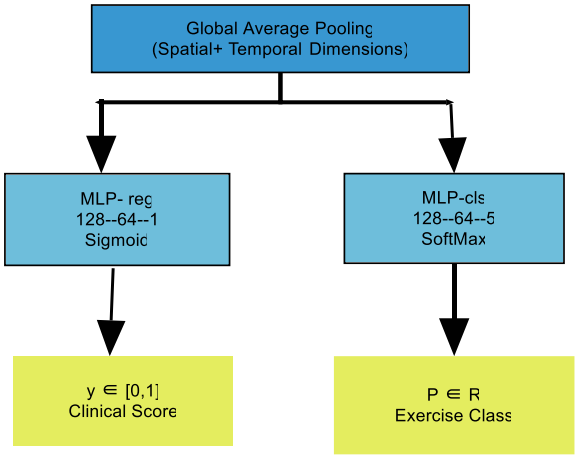

Figure 3. Multi-task prediction heads for clinical score regression and exercise classification.

Global feature representations are obtained using spatial-temporal global average pooling:

$$g = \text{GAP}\left(f_{\text{out}}^{(L)}\right) \tag{17}$$

where $L = 3$ denotes the final graph convolution layer and $g \in \mathbb{R}^{128}$

Clinical Score Regression

The regression branch predicts continuous movement quality scores:

$$\hat{y} = \mathrm{MLP}_{\mathrm{reg}}(g) \tag{18}$$

The regression objective uses mean squared error (MSE):

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{B}\sum_{b=1}^{B}(y_b - \hat{y}_b)^2 \tag{19}$$

Exercise Classification

The classification branch predicts exercise categories:

$$\hat{y}_{\mathrm{cls}} = \mathrm{SoftMax}\big(\mathrm{MLP}_{\mathrm{cls}}(g)\big) \tag{20}$$

The classification loss is defined using categorical cross-entropy:

$$\mathcal{L}_{\mathrm{cls}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{21}$$

where $C = 5$ denotes the number of rehabilitation exercise classes.

Multi-Task Optimization

The total optimization objective combines regression and classification losses:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}} + \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}} \tag{22}$$

where $\lambda_{\mathrm{reg}} = 1.0$ and $\lambda_{\mathrm{cls}} = 0.5$ are empirically selected weighting coefficients based on validation experiments.

4. Experimental Results and Analysis

This section presents the empirical evaluation of the proposed CD-STGCN framework. The experiments are designed to investigate: (1) the effectiveness of dynamic clustering compared with fixed anatomical priors, (2) the contribution of spatial and temporal feature streams, (3) the behavior of the learned clustering structures under healthy and pathological movement patterns, and (4) the computational efficiency of the proposed architecture. The experiments were repeated through five separate runs with random initialization. The evaluation metrics Mean Absolute Error (MAE), Pearson correlation (r), Accuracy, Precision, Recall, and F1-score are measured to achieve the precision of the proposed method.

4.1. Dataset

Experimental results were obtained using the KIMORE rehabilitation dataset. It contains recordings from 78 stroke survivors (age: 61.3±12.7 years; 52% male) performing five rehabilitation exercises: Lateral Tilt, Trunk Rotation, Pelvic Rotation, Sit-to-Stand, and Bending. Each exercise was performed at a self-selected speed with 2–3 repetitions per subject. The clinical movement quality scores were provided by three board-certified physiotherapists with a minimum of five years of experience in neurological rehabilitation. The assessments were independently performed using validated clinical instruments, including the Fugl–Meyer Assessment and the Exercise Accuracy Assessment Questionnaire (EAAQ). Repetition was evaluated on a 0–100 scale considering movement quality, pain indicators, and compensatory movement patterns. The final reference score corresponds to the arithmetic mean of the three independent ratings. Scores of training models were linearly normalized to the [0,1] interval, where 1 represents optimal movement quality, and 0 indicates severe motor impairment.

To ensure fair comparison, all baseline methods (ST-GCN, CTR-GCN, MS-G3D, 2s-AGCN, DeGCN, and D-STGCNT) were trained from scratch on the KIMORE dataset using identical preprocessing procedures, training protocols, and hyperparameter search spaces. Classification-oriented architectures were adapted for clinical regression by replacing the final

classification layers with regression heads optimized using MSE loss. To avoid subject-level data leakage, a subject-independent train/validation/test split strategy was employed, ensuring that all repetitions from the same participant were assigned exclusively to a single partition. Although KIMORE provides a clinically relevant benchmark for rehabilitation assessment, the relatively limited cohort size ($n = 78$) should be considered when interpreting generalizability. Therefore, the reported results primarily reflect dataset-specific performance within this stroke rehabilitation setting.

4.2. Validating the Dynamic Clustering Hypothesis

4.2.1. Sample-Adaptive Graph Construction

The proposed LMC module replaces fixed anatomical graph connections with dynamically learned coordination structures. Instead of relying solely on physical joint proximity, the model learns sample-specific connectivity patterns directly from motion dynamics. Three stages that the clustering process consists of: First, temporal encoding transforms each joint trajectory into a latent representation z_v through adaptive temporal pooling. Second, a Student's t-distribution kernel computes soft assignment based on kinematic similarity rather than anatomical distance evaluation. Third, cluster-specific adjacency matrices are computed using Equation (12). Joints exhibiting similar temporal behavior receive stronger connectivity weights. This formulation enables the graph structure to adapt to pathological coordination patterns that may not follow standard anatomical relationships.

4.2.2. Empirical Validation of Sample-Adaptive Clustering

The LMC module was replaced with fixed anatomical adjacency matrices corresponding to the standard ST-GCN topology, while all remaining architectural components were kept identical. As illustrated in Table 2, replacing adaptive clustering with fixed anatomical graphs resulted in a 20.4% performance degradation (MAE: 0.118 vs. 0.098, $p<0.001$). Outputs suggest that sample-adaptive graph construction improves representation flexibility for pathological movement analysis.

The trustworthiness between CD-STGCN predictions and expert physiotherapist assessments was examined using the two-way random single-measure intraclass correlation coefficient ICC(2,1). This formulation models both raters and subjects as random effects, supporting generalization across clinicians [35]. ICC(2,1) was computed using mean squares derived from a two-way ANOVA formulation:

$$ICC(2,1) = \frac{MS_{subject} - MS_{error}}{MS_{subject} + (R-1)MS_{error} + \frac{R}{n}(MS_{rater} - MS_{error})} \quad (23)$$

where $R = 3$ indicates the number of raters and $n = 78$ represents the number of subjects.

The proposed framework achieved an ICC value of 0.89 (95% CI: 0.84–0.94), indicating strong agreement with expert ratings of the dataset. This value is comparable again to ranges commonly reported in rehabilitation assessment literature for physiotherapist inter-rater reliability. However, these comparisons remain indirect, as no explicit rater-to-rater variability analysis was conducted within the current cohort.

Table 2. Component study on the KIMORE dataset

Configuration	MAE	R
Full CD-STGCN	0.098	0.91
w/o Velocity Stream	0.112	0.87
w/o LMC (fixed anatomy)	0.118	0.85
w/o Cluster Modulation	0.105	0.89
w/o Multi-Task Learning	0.107	0.88
Single-Stream (spatial)	0.125	0.84

The ablation analysis further shows that both the velocity stream and adaptive clustering module contribute substantially to performance. In particular, the degradation observed in the single-stream configuration indicates that temporal motion dynamics remain important for rehabilitation-oriented movement assessment. Table 3 summarizes the exercise

classification performance associated with the multi-task learning framework. The model achieved an overall classification accuracy of 96.5%±0.8%, with consistently high Precision, Recall, and F1-score across all exercise categories.

Table 3. Multi-Task Exercise Classification Performance

Exercise Type	Accuracy (%)	Precision	Recall	F1-Score
Overall Mean	(96.5 \pm 0.8)	0.96	0.95	0.95
Lateral Tilt (Es1)	(97.2 \pm 0.4)	0.97	0.97	0.97
Trunk Rotation (Es2)	(95.8 \pm 1.1)	0.95	0.94	0.94
Pelvic Rotation (Es3)	(96.1 \pm 0.9)	0.96	0.96	0.96
Sit-to-Stand (Es4)	(97.5 \pm 0.5)	0.98	0.97	0.97
Bending (Es5)	(95.9 \pm 1.2)	0.94	0.95	0.94

The relatively consistent classification performance across exercises suggests that the proposed dual-stream representation and adaptive clustering mechanism are capable of capturing discriminative spatio-temporal movement characteristics across multiple rehabilitation activities.

4.3. Cluster Behavior and Learned Coordination Patterns

4.3.1. Emergent Structure Without Anatomical Supervision

An important characteristic of the proposed Latent Motion Clustering (LMC) module is its ability to discover functionally coherent joint groupings without predefined anatomical supervision. Table 4 summarizes the dominant joint assignments learned by the clustering mechanism.

Table 4. Learned functional cluster structure

Cluster	Primary Joints	Functional Role	Avg. Confidence
1	SpineBase, SpineMid, Neck, Head	Trunk/Core stabilization	0.87
2	ShoulderLeft, ElbowLeft, WristLeft, HandLeft	Left upper extremity	0.91
3	ShoulderRight, ElbowRight, WristRight, HandRight	Right upper extremity	0.90
4	HipLeft, KneeLeft, AnkleLeft, FootLeft	Left lower extremity	0.89
5	HipRight, KneeRight, AnkleRight, Foot-Right	Right lower extremity	0.88

Although clustering was learned directly from motion dynamics rather than anatomical labels, the resulting structures showed alignment with kinematically coherent body regions. The soft-assignment mechanism additionally allows partial joint membership across multiple clusters, enabling the representation of overlapping coordination behavior frequently observed in rehabilitation movements.

4.3.2. Visualization of Learned Coordination Patterns

Figure 4 illustrates representative connectivity patterns generated by the LMC module for healthy and pathological movement samples. These visualizations are intended as exploratory representations of learned coordination behavior rather than validated clinical biomarkers. n the healthy movement example (Figure 4A), the learned graph structure exhibits strong localized connectivity within anatomically related regions, with relatively limited cross-limb interactions. In contrast, the pathological movement example (Figure 4B) shows increased cross-regional connectivity, particularly between hip and shoulder regions, suggesting altered coordination behavior during compensatory movement execution.

These observations indicate that the adaptive graph construction mechanism is capable of capturing motion relationships beyond fixed anatomical adjacency, potentially improving sensitivity to pathological coordination changes.

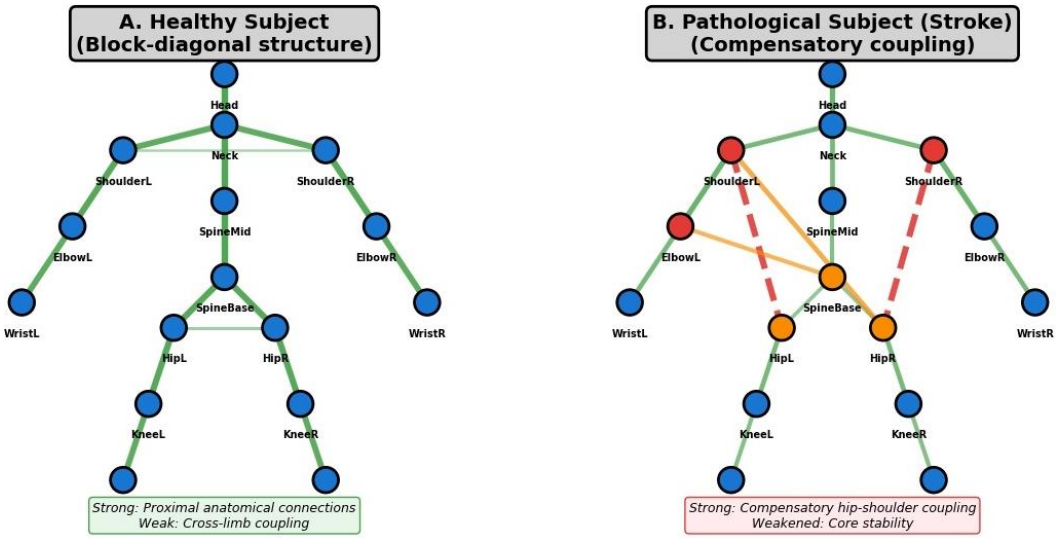

Figure 4. Visualization of learned coordination patterns for healthy and pathological movement samples.

4.3.3. Adjacency Matrix Analysis

Figure 5 shows adjacency matrix visualizations generated from the learned graph structures; such matrices provide a quantitative representation of the connectivity patterns produced by the LMC module. In Figure 5(a), the healthy subject demonstrates a relatively clear block-diagonal structure. Indicating stronger connectivity within functional joint groups and weaker interactions across unrelated regions. Figure 5(b) shows increased off-diagonal connectivity in the pathological sample, reflecting abnormal cross-regional coordination patterns.

The accentuated connectivity values related to learned adjacency weights are derived from the soft cluster assignment probabilities obtained by the Student's t-distribution kernel. These distribution values represent the strength of coordinated kinematic activity between joints within the learned graph representation rather than Pearson correlation coefficients or direct anatomical relationships. Figure 5(c) shows the difference between pathological and healthy connectivity patterns, highlighting regions with increased coordination strength in the pathological movement sample. The results imply that the adaptive graph representation captures movement relationships that may not be explicitly represented in static anatomical graph structures.

4.3.4. Cluster Number Sensitivity

The subtlety of the proposed architecture with respect to the number of clusters K was examined, as clarified in Table 5. The best performance outcomes from using $K = 5$, which provided a balance between representational capacity and generalization. Reducing the number of clusters to $K = 3$ resulted in decreased performance (MAE = 0.108, suggesting useless representational granularity for modeling distinct functional body regions. Increasing the number of clusters beyond five also reduced performance, likely due to increased model complexity relative to the limited dataset size.

Table 5. Performance sensitivity to the number of cluster K

Clusters (K)	MAE	R	Interpretation
K = 3	0.108	0.87	Limited representational capacity
K = 5 (optimal)	0.098	0.91	Balanced functional representation
K = 7	0.101	0.90	Redundant cluster structure
K = 10	0.103	0.89	Increased risk of overfitting

The results indicate that moderate cluster granularity provides sufficient flexibility for modeling functional coordination patterns without using excessive structural complexity.

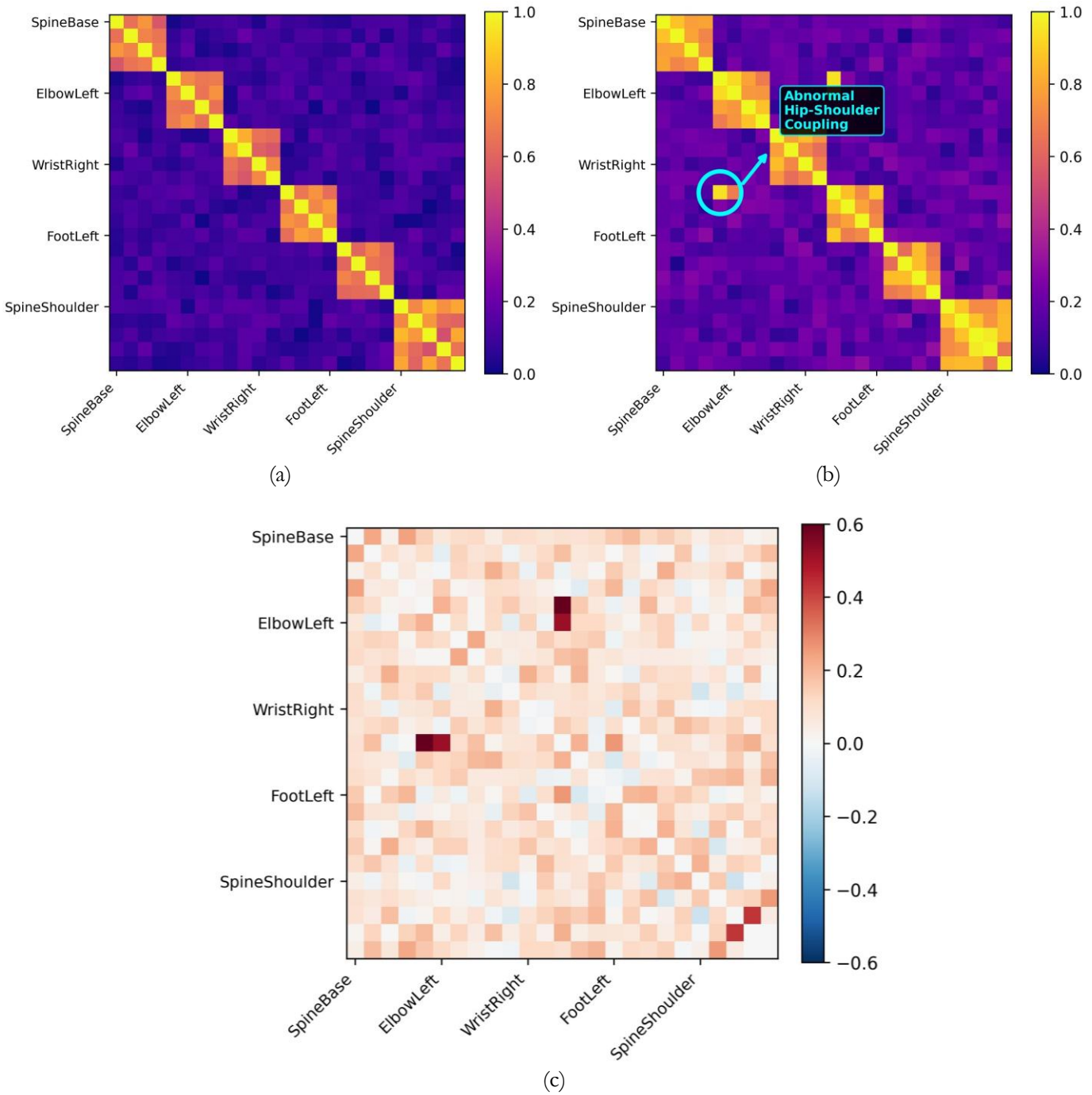

Figure 5. Learned coordination patterns using sample-adaptive adjacency matrices (**a**) Healthy subject; (**b**) Pathological subject (stroke); (**c**) Difference.

4.4. Comparison with Alternative Architectures

CD-STGCN architecture was compared with transformer-based rehabilitation assessment architectures, particularly D-STGCNT [30] which is a settled competitive predictive performance (MAE = 0.108), the proposed framework achieved a prediction error (MAE = 0.098), which is considered acceptable while maintaining substantially higher inference efficiency compared with transformer-based attention mechanisms. The proposed adaptive clustering method submits lower computational complexity while preserving the ability to model dynamic movement relationships. The resulting inference speed reached 83 FPS, corresponding to approximately 2.37× faster processing compared to D-STGCNT.

Figure 6 illustrates the relationship between CD-STGCN predictions and expert clinical assessment scores on the KIMORE dataset. The predicted scores denoted strong linear

agreement with the reference assessments (r=0.91). Most samples were distributed near the ideal prediction line, with larger variance observed in the intermediate score range (0.4−0.7), which may reflect increased ambiguity in moderately impaired movement patterns.

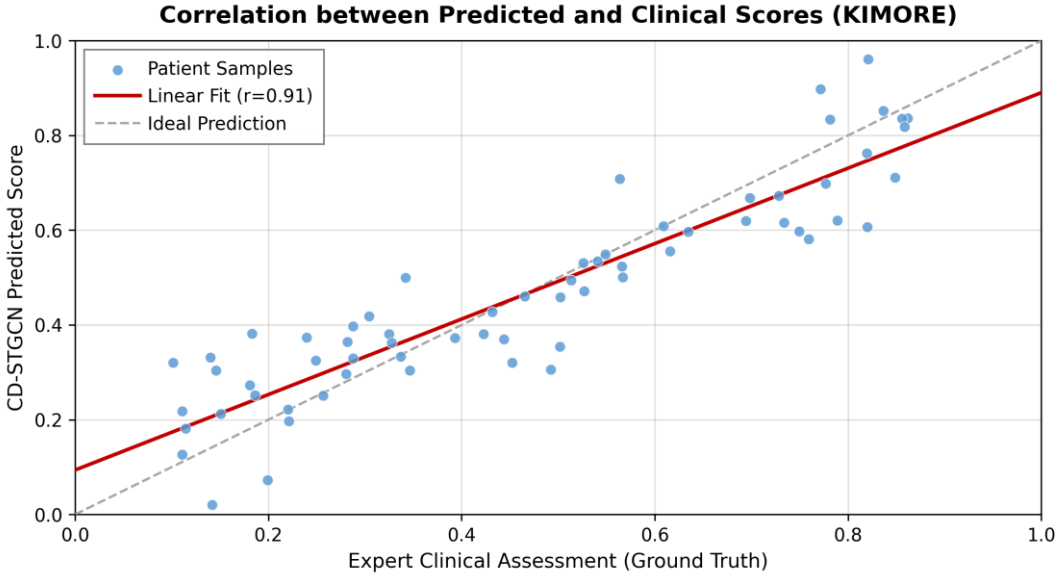

Figure 6. Correlation between predicted movement quality scores and expert clinical assessments.

The observed correlation and ICC values imply strong agreement between model predictions and expert ratings within the evaluated dataset. However, these findings should be interpreted as dataset-specific evidence derived from the KIMORE cohort rather than definitive evidence of generalized clinical reliability.

4.5. Comparative Analysis with State-of-the-Art Methods

Table 6 compares the proposed framework with recent graph-based skeleton action recognition and rehabilitation assessment methods. The comparison includes prediction error, graph representation strategy, velocity modeling capability, sample-adaptive graph construction, and real-time inference performance.

Table 6. Comparison with state-of-the-art skeleton-based rehabilitation assessment methods

Method	Year	MAE	Graph Type	Velocity Modeling	Sample-Adaptive	FPS
2s-AGCN [17]	2019	0.135	Global adaptive	No	No	38
MS-G3D [11]	2020	0.128	Fixed multi-scale	No	No	42
CTR-GCN [18]	2021	0.121	Channel-wise adaptive	No	No	51
D-STGCNT [30]	2024	0.108	Attention-based	Implicit	No	35
DeGCN [32]	2024	0.115	Graph decomposition	No	No	68
CD-STGCN (ours)	2026	0.098	Dynamic clustering-based	Yes	Yes	83

The proposed framework achieved the lowest prediction error while maintaining real-time inference capability. Opposite to prior approaches that rely on fixed or globally adaptive graph structures. CD-STGCN incorporates sample-adaptive graph construction through latent motion clustering, enabling dynamic modeling of subject-specific coordination patterns. The combined use of dual-stream temporal modeling and adaptive graph construction appears to contribute to improved rehabilitation-oriented movement representation, particularly for pathological motion patterns that may not conform to fixed anatomical connectivity assumptions.

5. Discussion

The CD-STGCN framework presents as a graph-based framework for automated clinical movement assessment that moves beyond fixed anatomical priors toward adaptive, data-driven coordination modeling. The experimental results demonstrate that dynamically learned graph structures can provide more flexible representations of pathological movement behavior compared with conventional fixed-topology approaches. The ablation analysis in Table 2 highlights the importance of the proposed Latent Motion Clustering (LMC) module. Replacing adaptive clustering with fixed anatomical adjacency matrices resulted in a substantial increase in prediction error (MAE: 0.118 vs. 0.098), suggesting that static anatomical connectivity may limit the representation of compensatory movement behavior commonly observed in neurological rehabilitation. These findings are consistent with the observation that pathological movement patterns do not always follow anatomically localized coordination structures. In particular, the increased cross-regional connectivity observed in the adjacency visualizations suggests that stroke-related compensatory behavior may involve coordinated recruitment of anatomically distant body regions.

The learned clustering structures indicate outgrowths without explicit anatomical supervision. Despite the resulting clusters broadly aligning with clinically intuitive body regions, the soft-assignment mechanism additionally captures overlapping coordination behavior across multiple functional groups. This characteristic may be relevant in rehabilitation settings where movement compensation emerges gradually rather than through discrete anatomical transitions. The degradation noticed in the single-stream and velocity-removed configurations suggests that temporal motion dynamics provide complementary information beyond static posture representation alone. The finding is clinically relevant because many neurological impairments, including bradykinesia, freezing, and compensatory stabilization behavior, are inherently expressed through temporal movement characteristics rather than isolated joint positions.

A lower prediction error was achieved in the proposed framework while maintaining substantially higher inference efficiency. The reported real-time performance of 83 FPS specifies that adaptive graph construction can be integrated into computationally efficient rehabilitation systems without requiring high-complexity attention mechanisms; this may be advantageous for future deployment in real-time rehabilitation monitoring environments, particularly on resource-constrained hardware. The observed agreement between model predictions and expert clinical scores is ($r=0.91$, ICC = 0.89), further suggesting that the learned representations capture clinically relevant movement characteristics within the dataset cohort; However, these findings should be interpreted cautiously. The current study evaluates performance on a relatively limited dataset consisting of 78 stroke survivors, and no direct intra-study rater-to-rater comparison was performed. The reported agreement metrics should be considered dataset-specific proof-of-concept evidence rather than definitive validation of generalized clinical reliability.

Several limitations should also be taken into consideration. First, the relatively limited cohort size may restrict generalizability across broader neurological populations, rehabilitation protocols, and demographic groups, Second, the proposed image-to-graph pipeline remains dependent on the quality of upstream 3D skeleton estimation. Severe sensor noise, occlusion, or pose estimation failure may propagate errors into the graph construction stage and affect downstream clustering behavior, and Third, although the proposed visualizations provide interpretable representations of learned coordination patterns, these structures should not be interpreted as validated physiological biomarkers without further prospective clinical validation.

Future work should therefore investigate large-scale multi-center validation across diverse neurological conditions and rehabilitation environments. Incorporating self-supervised or domain-adaptive learning strategies may further improve robustness to sensor variability and occlusion. In addition, future studies should evaluate the proposed framework in unconstrained home-based rehabilitation settings to better assess real-world deployment feasibility for automated telerehabilitation systems.

6. Conclusion

This paper proposed CD-STGCN, a clustering-based dual-stream spatio-temporal graph convolutional framework for automated clinical movement assessment. Unlike conventional

graph-based approaches that rely on fixed anatomical priors, the proposed method dynamically learns latent coordination structures directly from skeletal motion data through the Latent Motion Clustering module. The framework combined dual-stream spatial-temporal feature extraction, adaptive graph construction, and multi-task learning. Experimental evaluation on the KIMORE dataset demonstrated competitive predictive performance (MAE = 0.098, r=0.91) while maintaining real-time inference capability at 83 FPS. Ablation studies further showed that adaptive clustering and temporal motion modeling contribute substantially to rehabilitation-oriented movement representation.

The learned graph structures additionally provide an interpretable visualization of coordination behavior, illustrating the potential of adaptive graph modeling for capturing pathological movement patterns that may not be represented through fixed anatomical connectivity alone. Nevertheless, the current findings should be interpreted as dataset-specific proof-of-concept evidence derived from a relatively limited stroke rehabilitation cohort. Overall, the proposed framework demonstrates the feasibility of combining adaptive graph learning and spatio-temporal motion modeling for automated rehabilitation assessment. Further large-scale validation across diverse neurological conditions and clinical environments remains necessary before broader clinical deployment can be considered.

Author Contributions: Conceptualization: D.M.W. and A.H.Z.; Methodology: D.M.W.; Software: A.H.Z.; Validation: D.M.W. and A.H.Z.; Formal analysis: A.H.Z.; Investigation: D.M.W.; Resources: D.M.W.; Data curation: A.H.Z.; Writing—original draft preparation: D.M.W.; Writing—review and editing: D.M.W.; Visualization: A.H.Z.; Supervision: D.M.W.; Project administration: D.M.W.; Funding acquisition: D.M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The KIMORE dataset used in this study is publicly available and was originally published by Capecci et al. in IEEE Transactions on Neural Systems and Rehabilitation Engineering (2019, DOI: 10.1109/TNSRE.2019.2923060). The dataset repository is maintained at: KIMORE Dataset Repository. During the preparation of this study, temporary server access issues were encountered; therefore, dataset access was additionally facilitated through publicly available processed distributions and direct correspondence with the original authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] W. M. Carroll, "The global burden of neurological disorders," *Lancet Neurol.*, vol. 18, no. 5, pp. 418–419, May 2019, doi: 10.1016/S1474-4422(19)30029-8.

[2] F. S. Sarfo, U. Ulasavets, O. K. Opare-Sem, and B. Ovbiagele, "Tele-Rehabilitation after Stroke: An Updated Systematic Review of the Literature," *J. Stroke Cerebrovasc. Dis.*, vol. 27, no. 9, pp. 2306–2318, Sep. 2018, doi: 10.1016/j.jstrokecerebrovasdis.2018.05.013.

[3] S. C. Cramer, B. M. Young, A. Schwarz, T. Y. Chang, and M. Su, "Telerehabilitation Following Stroke," *Phys. Med. Rehabil. Clin. N. Am.*, vol. 35, no. 2, pp. 305–318, May 2024, doi: 10.1016/j.pmr.2023.06.005.

[4] A. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind, "The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance," *J. Rehabil. Med.*, vol. 7, no. 1, pp. 13–31, Jan. 1975, doi: 10.2340/1650197771331.

[5] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A Human Activity Recognition System Using Skeleton Data from RGBD Sensors," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–14, 2016, doi: 10.1155/2016/4351435.

[6] C. G. Goetz *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Mov. Disord.*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008, doi: 10.1002/mds.22340.

[7] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12328.

[8] R. Dickstein and N. Abulaffio, "Postural sway of the affected and nonaffected pelvis and leg in stance of hemiparetic patients," *Arch. Phys. Med. Rehabil.*, vol. 81, no. 3, pp. 364–367, Mar. 2000, doi: 10.1016/S0003-9993(00)90085-6.

[9] L. P. Cobb *et al.*, "Movement System Diagnosis: A Consensus-Based Position for Pediatric Movement Specialists," *Pediatr. Phys. Ther.*, vol. 37, no. 2, pp. 257–264, Apr. 2025, doi: 10.1097/PEP.0000000000001186.

[10] Z. Zhang, "Microsoft Kinect Sensor and Its Effect," *IEEE Multimed.*, vol. 19, no. 2, pp. 4–10, Feb. 2012, doi: 10.1109/MMUL.2012.24.

[11] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 140–149. doi: 10.1109/CVPR42600.2020.00022.

[12] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in Healthcare," in *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, Jun. 2020, pp. 1–2. doi: 10.1109/CyberSA49311.2020.9139655.

[13] X. Hu, "A Clinical Theory-Driven Deep Learning Model for Interpretable Autism Severity Prediction," *medRxiv*. Jan. 26, 2026. doi: 10.64898/2026.01.25.26344792.

[14] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical Graph Representation Learning with Differentiable Pooling," in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/e77dbaf6759253c7c6d0efc5690369c7-Paper.pdf

[15] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 588–595. doi: 10.1109/CVPR.2014.82.

[16] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *J. Neuroeng. Rehabil.*, vol. 9, no. 1, p. 21, Dec. 2012, doi: 10.1186/1743-0003-9-21.

[17] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 12018–12027. doi: 10.1109/CVPR.2019.01230.

[18] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 13339–13348. doi: 10.1109/ICCV48922.2021.01311.

[19] S. Parisot *et al.*, "Spectral Graph Convolutions for Population-Based Disease Prediction," in *Lecture Notes in Computer Science*, 2017, pp. 177–185. doi: 10.1007/978-3-319-66179-7_21.

[20] M. Y. Lu *et al.*, "AI-based pathology predicts origins for cancers of unknown primary," *Nature*, vol. 594, no. 7861, pp. 106–110, Jun. 2021, doi: 10.1038/s41586-021-03512-4.

[21] C. Lindner, S. Thiagarajah, J. Wilkinson, T. Consortium, G. Wallis, and T. Cootes, "Fully Automatic Segmentation of the Proximal Femur Using Random Forest Regression Voting," *IEEE Trans. Med. Imaging*, vol. 32, no. 8, pp. 1462–1472, Aug. 2013, doi: 10.1109/TMI.2013.2258030.

[22] K. A. Thomas *et al.*, "Automated Classification of Radiographic Knee Osteoarthritis Severity Using Deep Neural Networks," *Radiol. Artif. Intell.*, vol. 2, no. 2, p. e190065, Mar. 2020, doi: 10.1148/ryai.2020190065.

[23] J. Zhu, Z. Ye, M. Ren, and G. Ma, "Transformative skeletal motion analysis: optimization of exercise training and injury prevention through graph neural networks," *Front. Neurosci.*, vol. 18, Mar. 2024, doi: 10.3389/fnins.2024.1353257.

[24] F. M. Bianchi, D. Grattarola, and C. Alippi, "Spectral Clustering with Graph Neural Networks for Graph Pooling," *Proceedings of the 37 th International Conference on Machine Learning*. Dec. 29, 2020. [Online]. Available: http://arxiv.org/abs/1907.00481

[25] B. H. Dobkin, "Rehabilitation after Stroke," *N. Engl. J. Med.*, vol. 352, no. 16, pp. 1677–1684, Apr. 2005, doi: 10.1056/NEJMcp043511.

[26] G. Kwakkel, R. C. Wagenaar, B. J. Kollen, and G. J. Lankhorst, "Predicting Disability in Stroke—A Critical Review of the Literature," *Age Ageing*, vol. 25, no. 6, pp. 479–489, 1996, doi: 10.1093/ageing/25.6.479.

[27] R. A. Clark *et al.*, "Validity of the Microsoft Kinect for assessment of postural control," *Gait Posture*, vol. 36, no. 3, pp. 372–377, Jul. 2012, doi: 10.1016/j.gaitpost.2012.03.033.

[28] B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester, "Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease," *Gait Posture*, vol. 39, no. 4, pp. 1062–1068, Apr. 2014, doi: 10.1016/j.gaitpost.2014.01.008.

[29] W. Tao, T. Liu, R. Zheng, and H. Feng, "Gait Analysis Using Wearable Sensors," *Sensors*, vol. 12, no. 2, pp. 2255–2283, Feb. 2012, doi: 10.3390/s120202255.

[30] Y. Mourchid and R. Slama, "D-STGCNT: A Dense Spatio-Temporal Graph Conv-GRU Network based on transformer for assessment of patient physical rehabilitation," *Comput. Biol. Med.*, vol. 165, p. 107420, Oct. 2023, doi: 10.1016/j.compbiomed.2023.107420.

[31] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, and X.-S. Hua, "BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 2049–2058. doi: 10.1109/CVPR52733.2024.00200.

[32] W. Myung, N. Su, J.-H. Xue, and G. Wang, "DeGCN: Deformable Graph Convolutional Networks for Skeleton-Based Action Recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 2477–2490, 2024, doi: 10.1109/TIP.2024.3378886.

[33] Y.-W. Ma, J.-L. Chen, Y.-J. Chen, and Y.-H. Lai, "Explainable deep learning architecture for early diagnosis of Parkinson's disease," *Soft Comput.*, vol. 27, no. 5, pp. 2729–2738, Mar. 2023, doi: 10.1007/s00500-021-06170-w.

[34] S. Rani, S. Masood, and D. R. Rizvi, "A Bi-LSTM based approach for Compensation detection during robotic stroke rehabilitation therapy," *Procedia Comput. Sci.*, vol. 258, pp. 3145–3154, 2025, doi: 10.1016/j.procs.2025.04.572.

[35] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability.," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979, doi: 10.1037/0033-2909.86.2.420.