Classification of Movement Disorders Using Video Recordings of Gait with Attention-based Graph Convolutional Networks

Wei Tang*, Peter M.A. van Ooijen, *Senior Member, IEEE*, Deborah A. Sival and Natasha M. Maurits, *Senior Member, IEEE*

***Abstract*—Early Onset Ataxia (EOA) and Developmental Coordination Disorder (DCD) are two pediatric movement disorders characterized by similar phenotypic traits, often complicating clinical differential diagnostics. Despite the recognized reliability of current clinical scales like the Scale for the Assessment and Rating of Ataxia (SARA), their dependence on specialist expertise, time-consuming nature, and inherent subjectivity can potentially limit their efficacy in assessing movement disorders, thereby underscoring the need for more objective, and efficient diagnostic methods. This study introduces a novel approach that utilizes 2D video recording in the coronal plane coupled with pose estimation to differentiate gait patterns in children with EOA, DCD, and healthy controls (HC). An attention-based Graph Convolutional Network (A-GCN) was proposed for the classification process, achieving an f1-score of 76% at the group level. The model incorporates channel-wise attention to stress the semantic nuances of body joints, and temporal attention to highlight important sequences in gait patterns. These mechanisms enhance the model's ability to accurately classify EOA and DCD. Our results demonstrate the potential of this method to improve diagnosis and understanding of movement disorders, thereby paving the way for more targeted treatment strategies. The code is available at https://github.com/jiudaa/Attention-basedGCN-EOA.git.**

***Index Terms*— Early Onset Ataxia (EOA), Developmental Coordination Disorder (DCD), Graph Convolutional Network (GCN), deep learning.**

I. INTRODUCTION

Early Onset Ataxia (EOA) and Developmental Coordination Disorder (DCD) are both pediatric movement disorders which feature symptoms of motor incoordination. Gait analysis is an important tool in assessing the presentation and severity of the disorder. However, symptom overlap often complicates differential diagnosis. In a clinical setting, a common method of assessing ataxia is through semi-quantitative rating scales, specifically the Scale for the Assessment and Rating of Ataxia (SARA) [1]. While SARA is accepted for clinical use, it requires experienced specialists for interpretation, is time consuming, and has inherent subjectivity that may impact its results [2]. These characteristics limit its ability to evaluate movement disorders, emphasizing the need for more efficient and objective diagnostic tools.

Use of wearable inertial sensors could address these limitations offering a more objective assessment of motor incoordination [3]. However, these devices require additional preparation and calibration, hindering their widespread applicability in clinical settings. Alternatively, pose estimation techniques, like AlphaPose [4], can extract body joint position and movement from plain videos. This technique could help discern differences between EOA, DCD, and healthy control groups by analyzing the motion trajectories of various skeleton joint points. This approach is not only simple and convenient but also holds significant practical value due to its applicability in real-world clinical settings. Moreover, insights gained using this methodology could enhance our understanding of these disorders and aid in the early detection and treatment planning.

In recent years, various studies have used pose estimation for the analysis and identification of neurological diseases [5-7]. Lu et al. [5] proposed a double-Features double-motion network to assess Parkinson's Disease motor severity with SORT [8] and SPIN [9] for extracting 3D body joint locations. Wang et al. [6] used support vector machine to detect abnormal gait with 2D skeleton data from AlphaPose [4]. GCNs [10-12] have also proven effective in managing skeleton data, primarily due to their aptitude for accurately modeling the topological connections inherent to the human body. Yan et al. [10] introduced a learnable edge importance weighting strategy aimed at implementation of graph-based convolution as Spatio-Temporal Graph Convolutional Networks (ST-GCN) for skeleton based action recognition. Guo et al. [11] proposed a two-stream spatial-temporal graph convolutional network (2s-ST-AGCN) for video assessment of Parkinson's Disease gait motor disorder. However, a significant limitation of the existing literature is the lack of extensive studies on how these techniques can be integrated into routine clinical practice. The current project aims to create a movement disorder assessment method with pose estimation and graph based convolution. This approach leverages the capabilities of free hand single camera video footage to classify and differentiate between EOA and DCD, providing a simpler, yet effective alternative to existing methodologies.

II. DATASET

This study was conducted at the University Medical Center Groningen in The Netherlands, in compliance with local research ethics and integrity standards. A total of 84 children participated in the experiment. Informed consent was obtained from all participating children over the age of 12, and from the parents or guardians of all participants. As part of their diagnostic evaluations, the EOA and DCD patients underwent a range of tests at the Department of (Pediatric) Neurology. These evaluations potentially included radiologic (MRI), metabolic, electromyography, muscle ultrasound, laboratory, and genetic tests to rule out other potential underlying neurological disorders.

*Author supported by a grant from the China Scholarship Council.

Wei Tang, Peter M.A. van Ooijen, Deborah. A. Sival, and Natasha M. Maurits are with the University of Groningen, University Medical Center Groningen, the Netherlands; e-mail: w.tang@umcg.nl.

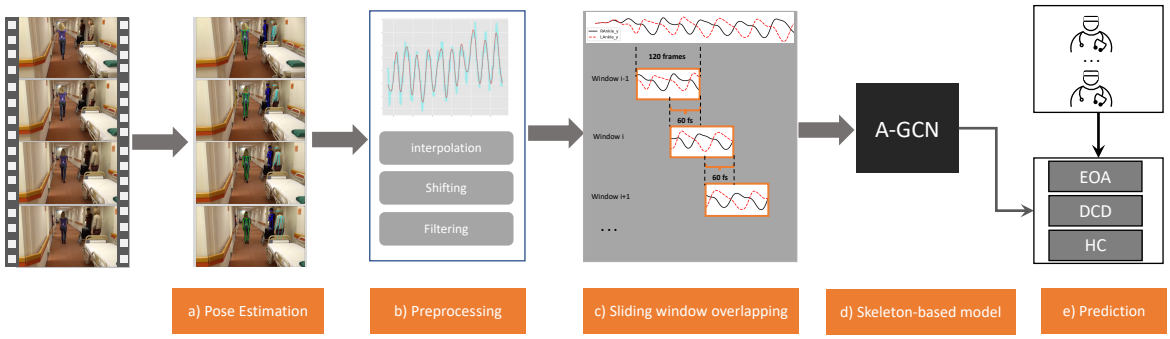

Figure 1. The proposed pipeline: (a) pose estimation to obtain the skeleton data, (b) preprocessing of skeleton data, (c) data augmentation using sliding window overlapping, (d) skeleton-based modelling, and (e) prediction of the movement disorders.

All children were asked to walk in a straight line along a corridor, make a 180-degree turn, and return to their starting position. This activity was in accordance with the guidelines for the regular gait task outlined in the SARA. A single experimenter stationed at the end of the corridor recorded the children's movements using a 2D video camera with different resolutions and frame rates, but resampled to a resolution of 1280 × 720 and 30 frames per second. Each video was divided into segments, categorized into four distinct types: (1) walking towards the camera, (2) walking away from the camera, (3) standing still, and (4) making a turn. The study's primary focus was on video segments featuring the children walking towards or away from the camera. Any frames in which body parts were obscured or lost were excluded from further analysis. This was due to the limitations of the 2D pose estimation algorithm used, which was unable to accurately detect skeletons in such frames. Finally, 240 video data segments from 84 participants (32 EOA, 34 healthy controls, 18 DCD) were utilized for skeleton dataset analysis.

III. METHODS

Figure 1 outlines the entire process for quantitatively assessing gait using freehand 2D camera recordings.

A. Skeleton Extraction

We employed AlphaPose to estimate the 2D positions of skeleton keypoints. A Yolo-v3 detector, pretrained on the MSCOCO dataset, was used for detecting individuals in the video frame. Subsequently, resnet50 was used to determine the final locations of each keypoint. Since videos may contain multiple other people, detecting and tracking the participant was crucial. To do this, PoseFlow was used to match the skeleton to the same subject throughout the recording. For each frame, the model provided 2D coordinates (in pixels) and a prediction confidence probability for all 17 keypoints (nose and bilateral eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles).

B. Preprocessing

We applied several preprocessing steps to the extracted skeleton data. First, to address missing keypoints in some frames, linear interpolation was used to estimate their positions by considering the values of neighboring keypoints. Second, to maintain consistency across frames, coordinates were transformed so that the mid-shoulder keypoint served as the origin in each frame, removing any positional offset caused by participant movement and providing a common reference point. Third, to minimize noise and fluctuations, average filtering was applied, calculating the average value of keypoints within a window of 9 frames, effectively smoothing the data and attenuating the effects of outliers or abrupt changes. Fourth, we segmented the data into smaller, more manageable portions using a sliding window approach, where each window contained 120 frames and had an overlap of 60 frames with the subsequent window. This facilitated the extraction of local features from the time series data, allowing for a more comprehensive analysis of gait patterns. Moreover, it also resulted in the generation of more samples, thereby bolstering the training process of our model.

C. Attention-based Graph Convolutional Network

Graph Convolution: When considering graph convolution, we can conceptualize the human skeleton as a graph, denoted as $G = (V, E)$ with V a set of N vertices (or joints) represented as $\{v_{t1}, v_{t2}, \ldots, v_{tN}\}$. As our model input, the feature vector at a node, Fv_{ti}, comprises *x* and *y* coordinates as well as the estimation confidence of the *i*-th joint at frame *t*. The edge set, E, is formulated as an adjacency matrix A in $\mathbb{R}_{NxN}$. Each element, a_{ij}, in this matrix signifies the connection strength between vertices v_i and v_j. The graph convolution operation in layer $l + 1$ collects local neighborhood information to refresh the node features, which can be represented as:

$$F(l+1) = \sigma(D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}F(l)W(l)) \quad (1)$$

Where $D_i = \sum_j \tilde{A}_{ij}$, $\tilde{A} = A + I$ is the adjacency matrix of an undirected graph, *I* is the identity matrix, W(l) is a learnable weight matrix, and σ(·) is set as the ReLU activation function.

Channel-wise Attention (CWA): The channel-wise attention mechanism refines the interaction between different joints [13]. This attention mechanism focuses on the fact that some joints and connections might be more important than others.

$$CWA(x) = g(x) * P(Softmax(\theta(x))) \quad (2)$$

An embedding function, $\theta(x)$, is applied to the input features *x* using a 1x1 convolution. The function $\theta(x)$ captures

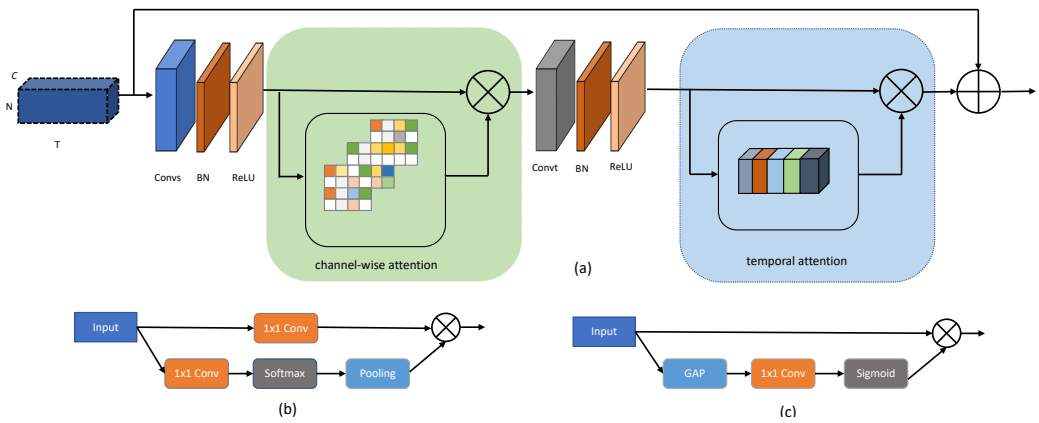

Figure 2. The detailed implementation of our proposed A-GCN model, presented in three key components: (a) an overview of the single layer A-GCN structure, (b) the fundamental block of the channel-wise attention module, and (c) the essential element of our temporal attention module.

varying aspects and features of the input data. A softmax operation is then applied to the output matrix, yielding a probabilistic understanding of the interaction among channels. Following this, a pooling operation *P(*)* is performed to reduce the dimensionality and extract dominant features. Finally, conducting an element-wise multiplication of this attention matrix with the features from a 1x1 convolution *g(x)* helps modulate the channel-wise feature responses according to their respective importance as shown in Figure 2 (b).

Temporal Attention (TA): The temporal attention employs an attention mechanism tailored for time series data in neural networks. By using a 1D convolution layer followed by a sigmoid function, this TA class enables the model to learn to focus on crucial time steps within the sequence data as:

$$\mathrm{TA}(x) = x \cdot Sigmoid(\varphi(P(x))) \qquad (3)$$

The process begins with a global average pooling operation *P(x)* which aims to reduce the dimensionality of the data. Subsequently, a 1D convolution layer, denoted by *φ(*)*, is used to discover local temporal patterns in the sequence data. The sigmoid function then assigns a set of probability-based weights to each time step, reflecting their respective importance. In the final step, these attention weights are used to modulate the original input features, effectively amplifying features deemed important while suppressing those identified as less relevant as shown in Figure 2 (c).

IV. Experimental Results

A. Implementation Details

Our experiments were run on the GPU Nibbler HPC cluster of the University Medical Center Groningen. To mitigate the potential for overfitting given our limited dataset size, we used a five-fold cross-validation. Clips from the same participant were not split between training and testing sets. We used the Adam optimizer for training, with a batch size of 8, a learning rate of 0.001 that decayed by half every 50 epochs, up to a maximum of 200 epochs. The total model training time was about nine hours for one model. To enhance the performance of the model, we incorporated a voting mechanism where each sub-clip was labeled in accordance with the overall label of the patient and the final prediction was determined by the majority vote across the sub-clips. This approach not only augmented the dataset, but also bolstered the system's robustness. Results reported are the best obtained for each model.

B. Comparing with State-of-the-art Methods

To assess the efficacy of our proposed A-GCN model, we compared it with the following models: 1) Conv1D: A 1D convolutional neural network that includes three 1D convolutional layers. 2) LSTM: A Long Short-Term Memory network with a fully-connected layer. 3) ST-GCN (Spatial Temporal Graph Convolutional Network) [10]: utilizes graph convolution and was initially applied for action recognition. 4) CTR_GCN [12]: A Channel-wise Topology Refinement Graph Convolution Network designed to effectively learn and refine topologies for different channels within GCN.

TABLE I

THE COMPARISON WITH STATE-OF-THE-ARTS MODELS.

Models	Mean F1	EOA	HC	DCD
Conv1D	0.64	0.66	0.72	0.49
LSTM	0.66	0.70	0.71	0.51
ST-GCN	0.72	**0.79**	0.72	0.62
CTR_GCN	0.71	0.74	0.74	0.61
A-GCN	**0.76**	**0.79**	**0.81**	**0.68**

A-GCN, with an average F1-score of 0.76 at the group level with 0.79 for EOA, 0.81 for HC and 0.68 for DCD, outperformed the other models in distinguishing between EOA, DCD and HC (Table I). These results suggest that the incorporation of the attention mechanism in A-GCN enhances the model's ability, outperforming other popular deep learning models applied to our dataset.

C. Ablation Studies

In our ablation study, we conducted a series of experiments with networks of varying complexities and configurations. Three types of networks were considered with different depths, specifically, 4-layer, 7-layer, and 10-layer models. These networks were further evaluated in three different conditions: without channel-wise attention, without temporal attention, and without both channel-wise and temporal attention.

It was shown that the depth of the network and the incorporation of attention mechanisms play an important role in the model's performance (Table II). When comparing the networks based on depth, it was found that the 7-layer network performed optimally, showing the best balance between complexity and accuracy. The 4-layer network, being less complex, did not have the capacity to capture intricate patterns in the data, leading to a lower performance. Conversely, the 10-layer network, although more complex, did not improve the performance, indicating a potential overfitting scenario. Examining the role of attention mechanisms, it's clear that both channel-wise and temporal attentions contributed positively to the model's performance. When these attention mechanisms were removed individually or jointly from the model, a drop in performance was observed. The experiments underscore the value of the attention mechanism and emphasize the importance of model depth optimization.

TABLE II

THE ABLATION STUDY RESULTS.

Models	F1 score
4 layer A-GCN	0.73
4 layer w/o CWA	0.71
4 layer w/o TA	0.73
4 layer w/o CWA&TA	0.71
7 layer A-GCN	**0.76**
7 layer w/o CWA	0.75
7 layer w/o TA	0.72
7 layer w/o CWA&TA	0.72
10 layer A-GCN	0.72
10 layer w/o CWA	0.70
10 layer w/o TA	0.69
10 layer w/o CWA&TA	0.69

V. DISCUSSION & CONCLUSION

This study presents a novel, non-invasive method for assessing and differentiating between EOA and DCD in pediatric populations, by applying computer vision and deep learning techniques. Leveraging the ease of 2D video recording and the power of pose estimation, our approach offers an accessible and efficient alternative to traditional, more invasive diagnostic tools, which often require specialist expertise and are time-consuming. The effectiveness of our method was evaluated and compared against established methods. Despite the limitations imposed by the small dataset, our model, equipped with attention mechanisms and optimally configured with a 7-layer depth, demonstrated competitive performance. These findings reinforce the value of pose estimation and GCN in movement disorder classification tasks and show a promising future for deep learning in clinical settings.

In conclusion, this work supports objective and efficient diagnosis of complex pediatric movement disorders like EOA and DCD. However, further work is needed to validate this approach with larger, more diverse datasets and to investigate its generalizability to other motor disorders. In addition, as a future direction, we aim to make the network more interpretable for clinical usage. This would include development of tools and techniques to visualize and understand the decision-making process of the model, which can further enhance trust in the predictions and inform clinical decision making. Despite these challenges, we are optimistic about the potential of this methodology to make a substantial contribution to the field of neurology and, most importantly, to the lives of the patients affected by these disorders.

VI. ACKNOWLEDGEMENTS

This work was supported by a grant from the China Scholarship Council under grant number 202006150040.

REFERENCES

[1] T. Schmitz-Hubsch, S. T. du Montcel *et al.*, "Scale for the assessment and rating of ataxia: development of a new clinical scale," *Neurology,* vol. 66, no. 11, pp. 1717-20, Jun 13, 2006.

[2] K. Burk, and D. A. Sival, "Scales for the clinical evaluation of cerebellar disorders," *Handb Clin Neurol,* vol. 154, pp. 329-339, 2018.

[3] Z. T. Dominguez-Vega, D. Dubber, J. W. J. Elting *et al.*, "Instrumented classification of patients with early onset ataxia or developmental coordination disorder and healthy control children combining information from three upper limb SARA tests," *Eur J Paediatr Neurol,* vol. 34, pp. 74-83, Sep, 2021.

[4] Fang H S, Xie S, et al. "Rmpe: Regional multi-person pose estimation." Proceedings of the IEEE international conference on computer vision. pp. 2334-2343, 2017.

[5] M. Lu, K. Poston, A. Pfefferbaum *et al.*, "Vision-based Estimation of MDS-UPDRS Gait Scores for Assessing Parkinson's Disease Motor Severity," *Med Image Comput Comput Assist Interv,* vol. 12263, pp. 637-647, Oct, 2020.

[6] Y. Wang, Q. Zou, Y. Tang *et al.*, "SAIL: A Deep-Learning-Based System for Automatic Gait Assessment From TUG Videos," *IEEE Transactions on Human-Machine Systems,* vol. 52, no. 1, pp. 110-122, 2022.

[7] W. Tang, P. M. A. van Ooijen, D. A. Sival *et al.*, "2D Gait Skeleton Data Normalization for Quantitative Assessment of Movement Disorders from Freehand Single Camera Video Recordings," *Sensors (Basel),* vol. 22, no. 11, Jun 2, 2022.

[8] F. Yang, Y. Wu, S. Sakti *et al.*, "Make Skeleton-based Action Recognition Model Smaller, Faster and Better," in Proceedings of the ACM Multimedia Asia, 2019, pp. 1-6.

[9] Kolotouros N, Pavlakos G, et al. "Learning to reconstruct 3D human pose and shape via model-fitting in the loop." Proceedings of the IEEE/CVF international conference on computer vision. pp. 2252-2261, 2019.

[10] Yan S, Xiong Y, Lin D. "Spatial temporal graph convolutional networks for skeleton-based action recognition." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

[11] R. Guo, X. Shao, C. Zhang *et al.*, "Multi-Scale Sparse Graph Convolutional Network For the Assessment of Parkinsonian Gait," *IEEE Transactions on Multimedia,* vol. 24, pp. 1583-1594, 2022.

[12] Chen Y, Zhang Z, Yuan C, et al. "Channel-wise topology refinement graph convolution for skeleton-based action recognition." Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13359-13368, 2021.

[13] Wang Q, Wu B, Zhu P, et al. "ECA-Net: Efficient channel attention for deep convolutional neural networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534-11542, 2020.