Quantitative Pose-Based Analysis of Movement Disorders in Pediatric NGLY1 and SLC13A5 Patients

Chengliang Dai[1,2] CHENGLIANG.DAI@UCB.COM
[1] *UCB, Slough, United Kingdom*
[2] *Imperial College London, London, United Kingdom*

Phil Scordis[1]
Prathyusha Teeyagura[3]
Rayann M. Solidum[3]
[3] *Stanford University, Stanford, CA, United States*

Jeff Broderick[4]
Julia Broderick[4]
Jane Broderick[4]
[4] *Beneufit, Inc., Kentfield, CA, United States*

Brenda E. Porter[3]

Editors: Accepted for publication at MIDL 2026

Abstract

Movement disorders have long relied on subjective clinical observation for diagnosis and monitoring. By contrast, computer vision tools such as OpenPose can turn video recordings into precise, time-resolved measurements of a patient's posture and movement. In this work, we apply a fully markerless, pose-based pipeline to classify abnormal movements in children with NGLY1 or SLC13A5 mutations. Our primary focus is on simple, physician-informed pose features that can be interpreted in clinical terms and used with conventional classifiers (Random Forest, SVM, etc.) on a very small dataset. We show that these handcrafted features capture clinically meaningful differences between movement-disorder phenotypes and can achieve useful classification performance. In addition, we include an exploratory comparison with a transformer model that is pre-trained on large-scale action-recognition data and then fine-tuned on our pose data. This experiment illustrates the potential performance ceiling of deep learning with extensive pretraining, but we emphasize that such models are less transparent and more data-hungry than the traditional approaches that form the core contribution of this study.

Keywords: Movement disorders, pose-based analysis, NGLY1, SLC13A5, pediatrics

1. Introduction

Movement disorders encompass a range of neurological conditions that impair motor control and lead to symptoms such as tremors, ataxia, and involuntary movements. Clinical evaluation has traditionally relied on expert visual assessment in the clinic. Although such assessments are grounded in deep clinical expertise, they remain subjective, resource-intensive, and difficult to reproduce or scale. For ultra-rare pediatric conditions such as NGLY1 deficiency and SLC13A5 disorder, these limitations are particularly problematic because patients are geographically dispersed and often cannot attend frequent in-person evaluations.

Recent advances in computer vision and artificial intelligence (AI) have opened the possibility of extracting quantitative measurements from ordinary video recordings. For instance, markerless pose-estimation algorithms infer joint locations frame by frame, enabling the computation of spatiotemporal kinematic descriptors without specialized motion-capture equipment. In this paper, we investigate whether pose-based features can be used to quantitatively characterize movement disorders in pediatric patients with NGLY1 deficiency or SLC13A5 disorder. Because our cohort is small and our clinical collaborators value interpretability, our main emphasis is on classical machine-learning models applied to carefully designed, physician-informed pose features. Specifically, we aim to (i) define a set of intuitive, angle-based features that correlate with clinician severity ratings and distinguish between broad categories of movement disorder, and (ii) evaluate a panel of conventional classifiers on these features, highlighting trade-offs between performance and interpretability.

An important design choice is that physician ratings are based on the original clinical videos, whereas the models are trained only on pose data extracted from those videos, since raw videos of pediatric patients are highly sensitive and cannot be readily shared across sites. As a secondary, exploratory analysis, we also adapt a transformer-based architecture that is pre-trained on a large public action-recognition dataset and fine-tuned on our clinical data. This experiment mainly serves to demonstrate what additional performance may be achievable with extensive pretraining.

2. Background and Related Work

2.1. Pose-based motion analysis in medicine

Markerless pose estimation and mesh recovery have transformed quantitative movement assessment by enabling extraction of 2D/3D joint coordinates from video. Representative frameworks include OpenPose (Cao et al., 2019), ViTPose (Xu et al., 2022), HRNet (Sun et al., 2019), VIBE (Kocabas et al., 2020), and HMR (Kanazawa et al., 2018), along with subsequent work building on these approaches. The resulting trajectories support computation of spatiotemporal kinematic descriptors (e.g., joint angles, angular velocities, and movement variability) without markers or wearable sensors.

For clinical use, the reliability of video-derived kinematics is critical. Multiple studies have therefore evaluated markerless pose pipelines against reference systems (e.g., instrumented gait analysis or marker-based motion capture) and reported encouraging agreement under standardized protocols. For example, OpenPose-based or OpenPose-derived pipelines have been assessed in adult gait/kinematic settings (Washabaugh et al., 2022; Stenum et al., 2021) and in pediatric contexts including toddlers with and without neurodevelopmental disabilities (Anderson et al., 2025). These validation efforts support the feasibility of using de-identified keypoint trajectories as quantitative inputs for downstream clinical modeling.

Building on these pose/mesh backbones, video-based kinematic analysis has been explored across neurological and developmental conditions. Beyond video, motion modeling has also been used in fetal MRI to estimate fetal motion and reduce motion artifacts (Xu et al., 2019; Zhang et al., 2020). These pose trajectories enable downstream learning-based clinical assessment.

2.2. AI-driven assessment of movement disorders

Given pose trajectories, prior work has developed ML/DL models to detect and quantify movement disorders. In Parkinson's disease, pose-based features combined with Graph Neural Networks have been used to classify tremor severity and assess bradykinesia by highlighting subtle movements that may be difficult to quantify by eye (Zhang et al., 2022; Quan et al., 2024). In gait analysis, pose-based metrics derived from joint trajectories, including step length, symmetry indices, and joint-angle entropy, have been used to identify gait abnormalities indicative of ataxia and other disorders (Tang et al., 2022). For infant and developmental disorders, early detection systems for cerebral palsy have analyzed spontaneous infant movements using pose-estimation models, showing that deviations from typical movement patterns can be detected from ordinary videos (Luo et al., 2022; Khan et al., 2018). Together, these studies show that pose-based analysis can yield clinically meaningful markers across a range of neurological conditions.

2.3. Transformers for motion modeling

A separate line of work has explored transformer architectures for modeling human motion. Pose Transformers (PoTr) (Martínez-González et al., 2021) introduced a non-autoregressive approach to motion prediction, thereby avoiding error accumulation in autoregressive models. Subsequent frameworks such as SPOTR (Nargund and Sra, 2023) and STPOTR (Mahdavian et al., 2023) disentangle spatial and temporal features, improving joint-trajectory prediction on large-scale motion datasets. Although these methods were developed primarily for general action recognition and motion synthesis, their ability to capture complex temporal dependencies suggests that, with sufficient data and careful pretraining, they could be applied to clinical video to assess diseases such as Parkinson's disease (Endo et al., 2022), for example by tracking disease progression or predicting changes in motor function. However, such models are typically less transparent than classical approaches and require larger datasets than are usually available in ultra-rare diseases.

2.4. Relevance to NGLY1 and SLC13A5 disorders

NGLY1 deficiency and SLC13A5 disorder are ultra-rare genetic syndromes characterized by poorly described motor delays and diverse movement-disorder phenotypes. Patients may exhibit both hyperkinetic movements (e.g., ataxia, chorea, myoclonus) and hypokinetic features (e.g., dystonia, bradykinesia), and these patterns can change with age. For such conditions, quantitative pose analysis offers three main advantages. First, it provides objective measurement: for example, tremor frequency and amplitude can be quantified from joint trajectories (Futrell et al., 2024), and similar principles can be extended to other movement patterns. Second, it enables longitudinal monitoring of therapy effects by tracking changes in movement features over time. Third, because these conditions are rare and patients are widely distributed geographically, pose-based analysis can support remote assessments from home-recorded videos, reducing the need for frequent in-person visits.

Within this context, our primary goal is to understand whether a small set of intuitive, physically meaningful pose features can already capture clinically relevant differences between movement-disorder phenotypes in NGLY1 and SLC13A5. We therefore concentrate

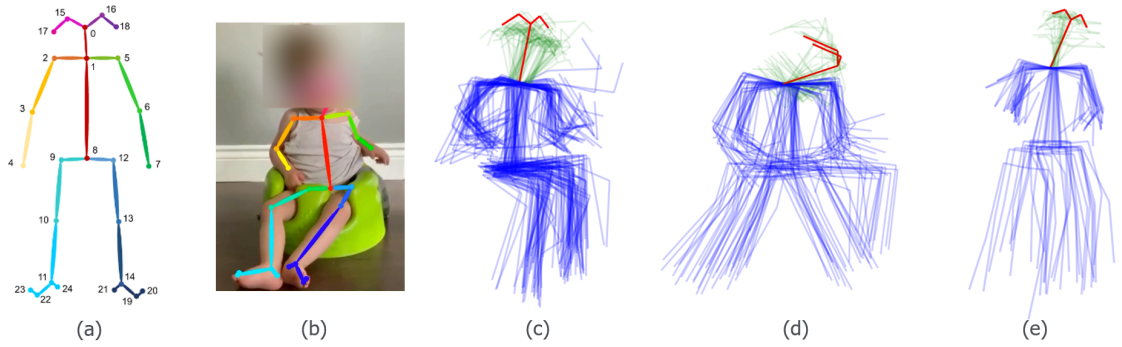

Figure 1: (a) OpenPose keypoint labels; (b) keypoints extracted from a representative frame; (c–e) keypoints extracted from video recordings from pediatric patients. The red lines highlight the mean neck angle.

on traditional classification models, which can be inspected through feature importance and related tools, and treat deep transformer models as exploratory benchmarks.

3. Methods

Our pipeline begins by extracting 2D skeleton data (Figure 1), then derives a set of angular and joint-angle features inspired by clinical rating scales such as ICARS (International Cooperative Ataxia Rating Scale) (Trouillas et al., 1997), UDRS (Unified Dystonia Rating Scale) (Comella et al., 2003), Burke–Fahn–Marsden Dystonia Rating Scale (BFMDRS) (Kuiper et al., 2016), and UMRS (Unified Myoclonus Rating Scale) (Frucht et al., 2002), from selected body parts, and finally applies a suite of classical classifiers to predict broad movement-disorder categories. In a separate, exploratory branch, we adapt a pre-trained pose transformer to the same task.

3.1. Pose features

The pose data are generated by the TRACER platform (Beneufit Inc.), which uses OpenPose as its core 2D pose-estimation backbone. TRACER also includes additional proprietary improvements for robustness and quality control (e.g., internal filtering and processing to better handle clinical videos). Pose quality was reviewed by neurologists. The labels of the keypoints are given in Figure 1(a). Pose-based features derived from these keypoints are used to quantify movement and relate it to the severity of the movement disorder. In contrast to the clinicians, who scored the disorders directly from the original videos, all our models operate only on 2D skeleton sequences. This design reflects privacy and data-governance constraints, as pose sequences provide a de-identified representation that is more amenable to research use and potential future data sharing.

Body parts were selected based on the quality and consistency of the extracted keypoints. For the present analysis, we focus on head, upper limb, and lower limb segments. Table 1 summarizes the keypoint pairs and triplets used to define angular-movement and joint-angle features.

Angular movement	Keypoint labels	Joint angle	Keypoint labels
Head	(0,1)	Neck	(0,1,2), (0,1,5)
Upper limbs	(2,3), (5,6)	Shoulder	(1,2,3), (1,5,6)
Lower limbs	(9,10), (12,13)	Elbow	(2,3,4), (5,6,7)

Table 1: OpenPose keypoint labels of selected body parts.

3.1.1. Angular-movement features

We compute segment-wise angular displacement using a fixed temporal lag of L frames (here $L = 10$). Let

$$\mathbf{v}_t = \left[x_t^{(a)} - x_t^{(b)},\ y_t^{(a)} - y_t^{(b)}\right]^\top$$

be the 2D limb vector defined by keypoints a and b at frame t. For each segment i with start frame $t_i = iL$, we compute the angular displacement between the vectors at the segment endpoints using a dot-product formulation:

$$\Delta\theta_i = \arccos\left(\frac{\mathbf{v}_{t_i} \cdot \mathbf{v}_{t_i+L}}{\|\mathbf{v}_{t_i}\|\,\|\mathbf{v}_{t_i+L}\|}\right).$$

Given N segments, the mean and variance of angular displacement are

$$\Delta\theta_{\text{avg}} = \frac{1}{N}\sum_{i=0}^{N-1}\Delta\theta_i, \qquad \sigma^2 = \frac{1}{N}\sum_{i=0}^{N-1}\left(\Delta\theta_i - \Delta\theta_{\text{avg}}\right)^2.$$

3.1.2. Mean joint-angle features

For joint-angle features, we compute the angle at a joint defined by three keypoints $p_t^{(1)}, p_t^{(2)}, p_t^{(3)}$ at each frame t. For instance, the shoulder angle uses neck, shoulder, and elbow keypoints. The angle is defined as

$$\theta_t = \cos^{-1}\Big(\frac{(p_t^{(1)} - p_t^{(2)}) \cdot (p_t^{(3)} - p_t^{(2)})}{\|p_t^{(1)} - p_t^{(2)}\|\,\|p_t^{(3)} - p_t^{(2)}\|}\Big),$$

and the mean joint angle over T frames is

$$\theta_{\text{mean}} = \frac{1}{T}\sum_{t=0}^{T-1}\theta_t.$$

3.2. Clinical labels and classification task

Pose data were extracted from videos of pediatric patients presenting with at least one movement disorder among ataxia, chorea, myoclonus, dystonia, hyperkinesia, or bradykinesia. For each video, two to three physicians independently scored the presence and severity of these disorders while blinded to each other's assessments. Each movement disorder was rated on a five-point scale: Absent (0), Minimal (1), Mild (2), Moderate (3), and Severe (4).

All clinical ratings were performed on the original video recordings. However, for model development we did not use raw pixel data: instead, each recording was processed with OpenPose/TRACER to obtain 2D joint coordinates, and all features and classifiers were derived exclusively from these pose sequences.

For SLC13A5 patients, only ataxia, chorea, dystonia, and myoclonus scores were observed during physician assessment. For NGLY1 patients, we computed two composite scores: one combining ataxia, chorea, and myoclonus, and another combining dystonia, hypokinesia, and bradykinesia. Although these movement disorders have distinct clinical presentations, their symptoms can overlap, particularly when detailed clinical context is limited, making accurate differentiation challenging for both clinicians and machine-learning models. The limited size of our dataset further complicates robust model development.

To mitigate these challenges, we simplified the classification task into four categories defined by the predominant movement and coordination features. **Normometric** videos were those with absent or minimal movement disorder (all scores 0–1). **Hypometric** videos showed dystonia, hypokinesia, or bradykinesia with a score of at least 2. **Hypermetric** videos showed ataxia, chorea, or myoclonus with a score of at least 2. **Mixed-metric** videos exhibited both at least one hypometric feature (dystonia, hypokinesia, or bradykinesia) and at least one hypermetric feature (ataxia, chorea, or myoclonus). We then trained several machine-learning classifiers, including Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), XGBoost (XGB), Logistic Regression (LR), and K-Nearest Neighbors (KNN), to predict these categories from the handcrafted pose features.

3.3. Transformer-based enhancement

In addition to the classical models, we trained a transformer-based model (Figure 2) to explore what performance could be achieved when leveraging large-scale pretraining. This analysis was not designed as an alternative clinical tool, but rather as an exploratory benchmark.

The input to the transformer is a sequence of t skeletons $\mathbf{X}_{1:t}$ extracted by OpenPose. During pre-training on a public action-recognition dataset, the network jointly learns to predict the next M skeletons $\mathbf{X}_{t+1:T}$ and to classify the action class of each input sequence. In the subsequent fine-tuning stage on our clinical dataset, the model is optimized solely to categorize the movement disorder, while the motion-prediction branch is frozen to reduce overfitting.

Our architecture, adapted from (Martínez-González et al., 2021; Endo et al., 2022), consists of several interconnected modules. A Graph Neural Network (GNN) encoder ϕ maps each input skeleton $\mathbf{x}_t$ to a fixed-dimensional embedding. Positional embeddings are then added to these skeleton embeddings before they pass through L multi-head self-attention layers in the transformer encoder, yielding a latent representation $\mathbf{z}_{1:t}$. A linear classification head processes $\mathbf{z}_{1:t}$ to produce either action class logits (during pre-training) or movement-disorder logits (during fine-tuning).

Simultaneously, the transformer decoder takes as input the encoder outputs $\mathbf{z}_{1:t}$ along with a query sequence $\mathbf{q}_{1:M}$ initialized with the last observed skeleton $\mathbf{X}_t$. We choose $\mathbf{X}_t$ because it is the most recent known state available at inference time and provides a strong, stable anchor that improves continuity at the boundary between observed and

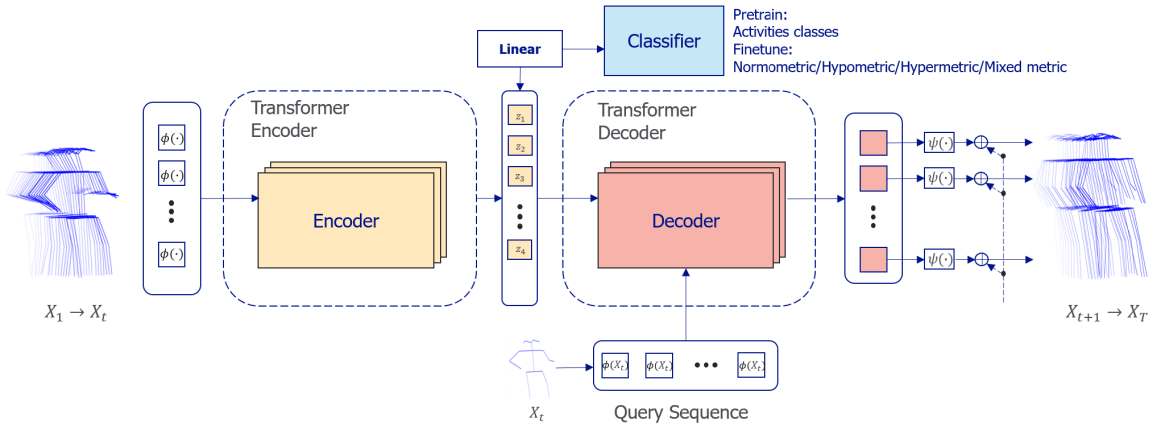

Figure 2: Transformer-based framework for predicting motion and movement disorder.

predicted motion (Martínez-González et al., 2021). After L layers of self-attention, the decoder outputs a sequence of embeddings, which the decoding network ψ transforms into reconstructed future skeletons $\hat{\mathbf{X}}_{t+1:T}$.

The overall loss during pre-training is the sum of a classification term and a motion reconstruction term:

$$L_{\text{total}} = L_{\text{cls}} + L_{\text{motion}}.$$

Here, L_{cls} denotes the cross-entropy loss for action classification. The motion loss L_{motion} is computed by averaging layerwise ℓ_1 reconstruction errors across all decoder layers. If $\hat{\mathbf{y}}_m^l$ is the predicted N-dimensional pose vector at time step m in decoder layer l, and $\mathbf{y}_m^*$ is the ground truth, then each layer's loss is given by

$$L_l = \frac{1}{M \cdot N} \sum_{m=t+1}^{T} \left\| \hat{\mathbf{y}}_m^l - \mathbf{y}_m^* \right\|_1,$$

and $L_{\text{motion}} = \frac{1}{L} \sum_{l=1}^{L} L_l$. In the fine-tuning stage, only the classification head is updated on the clinical labels, and the transformer results are reported primarily as a point of comparison with the more interpretable classical models.

4. Experimental Setup

4.1. Dataset

Our dataset comprises 95 video recordings of 26 pediatric patients with NGLY1 deficiency or SLC13A5 disorder, captured in standing or sitting positions. Repeated videos from the same child are separated by 4 to 24 months and can exhibit noticeable phenotype changes with development and disease progression. Thirteen recordings were excluded because the subjects were too young (less than 2 years old) to stand or sit independently at the time of recording. The raw videos were used only for clinical review and for extracting 2D skeletons.

The mean age of the subjects in the analyzed recordings is 9.86 ± 4.50 years. The mean number of frames per recording is 1259.08 ± 131.52. We trim each recording under

neurologist supervision to remove non-informative lead-in/lead-out segments and retain a fixed-length clip of 1000 frames. After trimming, we apply uniform fixed-stride sampling by keeping every 10th frame, producing a 100-frame pose sequence per recording. This 100-frame representation is used consistently for both handcrafted feature computation (with $L = 10$) and transformer fine-tuning.

After filtering, the dataset contains 7 normometric samples, 11 hypometric samples, 40 hypermetric samples, and 25 mixed-metric samples. Pose data are normalized using the neck (keypoint label 1) as the root: all joint coordinates are translated so that the root is fixed at $(0, 0)$ while the relative distances between joints remain unchanged. To pre-train the transformer, we also used 5,688 videos (standing, sitting, staggering gait, etc.) from the NTU RGB+D 120 dataset (Liu et al., 2019).

4.2. Model training and evaluation

The prepared data were partitioned into an 80% training set and a 20% test set using stratified sampling at the video level to preserve the class distribution. For the classical machine-learning models (RF, MLP, SVM, XGB, LR, and KNN), a feature selection process was conducted on the training set. First, an ensemble selection method ranked the features using the Chi-squared test, mutual information, ANOVA F-test, and recursive feature elimination. Subsequently, the final feature set was determined by a voting system, where features ranked in the top 12 (highlighted in Figure 3) by at least three of the four methods were selected for modeling. All features were standardized using z-scoring, with the mean and standard deviation computed on the training set.

To address class imbalance, class weighting was applied. Training was conducted in two phases. First, all models were evaluated using 5-fold stratified cross-validation on the training set to establish baseline performance based on the weighted F1-score. Next, the top three performing models underwent hyperparameter tuning via grid search, using the same 5-fold stratified cross-validation strategy. Finally, all models, including the optimized versions, were assessed on the held-out test set.

The transformer-based model consists of 4 encoder layers and 4 decoder layers, with an FFN dimension of 2048. It was pre-trained on the NTU RGB+D 120 dataset for 100 epochs with an initial learning rate of 1×10^{-4}, then fine-tuned on our training set. Due to limited patient data, we froze the motion-prediction branch during fine-tuning and trained only the classification head for 50 epochs at the same learning rate. Training was performed on an NVIDIA L40S GPU.

5. Results and Discussion

We first examined the correlation between clinician assessment scores and handcrafted features (Figure 3). Although the correlation coefficients are modest, pose-based features consistently show positive associations with clinical severity ratings, supporting their effectiveness as quantitative proxies for clinician-observed movement abnormalities.

For the classification task, we assessed model performance using accuracy, weighted precision, recall, and F1-score. Balanced metrics and per-class metrics are given in Table 2 and Table 3. The RF classifier achieved the highest accuracy among the classical models at 65%,

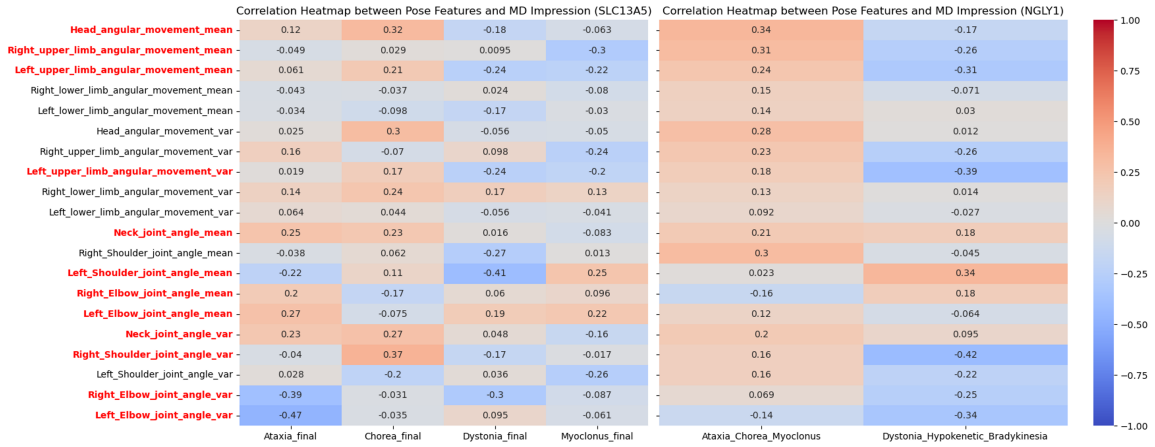

Figure 3: Correlation between clinical severity scores and pose-based features. Features selected for training conventional classifiers are highlighted in red.

while other models such as XGB and LR achieved slightly lower but comparable performance (Table 2). RF also provides greater interpretability compared with MLP, SVM, and other classifiers. The top 6 features used by the RF model include the mean and variance of left upper-limb angular movement, the mean and variance of neck angle, and the mean of head angular movement. Feature importance values of the top 6 features are shown in Figure 4. A differential analysis of the top 6 features using the Mann–Whitney U test was conducted on the full dataset across movement phenotypes and the healthy-reference cohort, correcting for multiple comparisons using the False Discovery Rate procedure, and the results are shown in Figure 5. The upper limbs and neck/head are often considered particularly important by physicians when assessing patients (Schmitz-Hubsch et al., 2006; Kuiper et al., 2016). These features help the model distinguish the hypermetric phenotype from others based on the statistical analysis (Figure 5), which is consistent with the strong RF performance for identifying hypermetric videos (Table 3). Although the healthy-reference cohort from the NTU dataset is generally older than the disease cohort and therefore cannot serve as a matched control group, the differential analysis suggests that the handcrafted features can also help distinguish movement phenotypes in the disease cohort from those in the healthy cohort.

Given the limited patient data, these results are encouraging, especially because RF and related models can provide feature-importance estimates that help clinicians understand which aspects of movement are most discriminative across categories.

The fine-tuned transformer-based model achieved the same accuracy as RF but obtained higher recall and F1-score due to improved performance in predicting the mixed-metric class. In particular, the transformer achieved a recall of 83% for the mixed-metric class, compared with 50% for RF (Table 3), highlighting its ability to identify more complex manifestations of movement disorders. As shown in the confusion matrices in Figure 6, the RF model commonly misclassifies mixed-metric samples as normometric, suggesting a limitation of the conventional classifiers under our feature set. The transformer results illustrate the potential

Model	Precision	Recall	F1-score	Accuracy
RF	**0.72**	0.57	0.62	**0.65**
MLP	0.60	0.50	0.55	0.47
SVM	0.65	0.55	0.60	0.53
XGB	0.70	0.55	0.62	0.59
LR	0.51	0.59	0.50	0.59
KNN	0.60	0.50	0.55	0.47
Transformer	0.71	**0.60**	**0.64**	**0.65**

Table 2: Classification results. The transformer is pre-trained on NTU RGB+D and fine-tuned on patient data; classical models are trained only on clinical features.

Model	Precision				Recall				F1-score			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
RF	0.00	1.00	**0.88**	1.00	0.00	1.00	**0.78**	0.50	0.00	1.00	**0.82**	0.67
MLP	1.00	0.00	0.43	1.00	1.00	0.00	0.33	0.67	1.00	0.00	0.37	0.80
SVM	0.00	1.00	0.67	1.00	0.00	1.00	0.44	0.67	0.00	1.00	0.53	0.80
XGB	0.00	1.00	0.83	1.00	0.00	1.00	0.56	0.67	0.00	1.00	0.67	0.80
LR	1.00	0.00	0.56	0.50	1.00	0.00	0.56	0.67	1.00	0.00	0.56	0.57
KNN	1.00	0.00	0.43	1.00	1.00	0.00	0.33	0.67	1.00	0.00	0.37	0.80
Transformer	1.00	0.00	0.83	1.00	1.00	0.00	0.56	**0.83**	1.00	0.00	0.67	**0.91**

Table 3: Per-class performance metrics. C1: Normometric; C2: Hypometric; C3: Hypermetric; C4: Mixed-metric.

benefit of large-scale pretraining, but this comes at the cost of reduced interpretability and greater complexity, and depends critically on access to extensive non-clinical training data.

6. Conclusion

We presented a quantitative, AI-driven framework for assessing movement disorders in pediatric patients with NGLY1 deficiency and SLC13A5 disorder, focusing on interpretable pose-based features and traditional machine-learning models suitable for small datasets. Our framework deliberately decouples clinical assessment (performed on raw videos) from model training (performed on de-identified pose data), which aligns with privacy constraints in pediatric ultra-rare disease cohorts and increases the feasibility of future cross-center data sharing. The pose-estimator agnostic design of the downstream framework can also ingest keypoints from newer pose estimation systems in the future.

Despite limitations such as the inability to test more recent pose-estimation methods, lack of ablation study for transformer design, limited data for robust patient-level splitting, and possible residual variability in home-recorded videos, our results show that simple, physician-informed features can differentiate broad movement-disorder categories and correlate with clinician severity ratings. An exploratory transformer experiment demonstrates

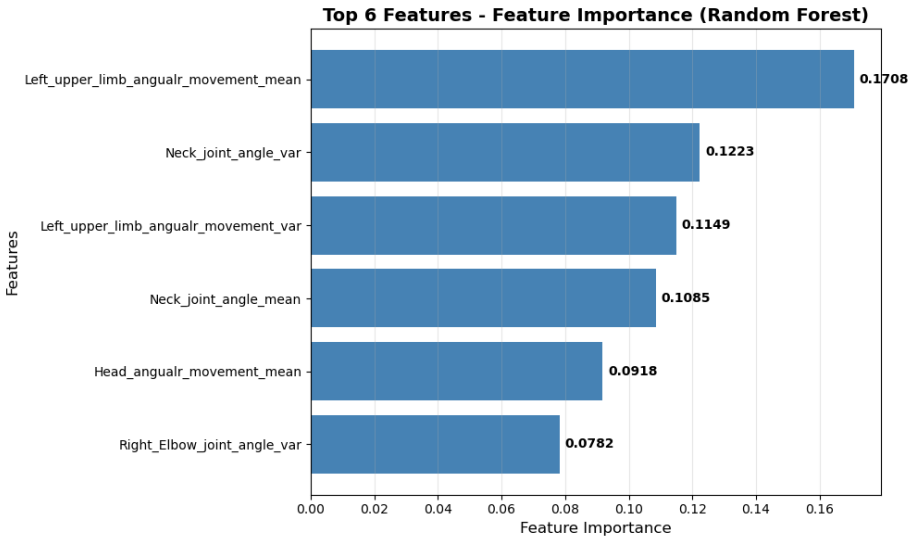

Figure 4: Feature importance values of the top 6 features used by RF.

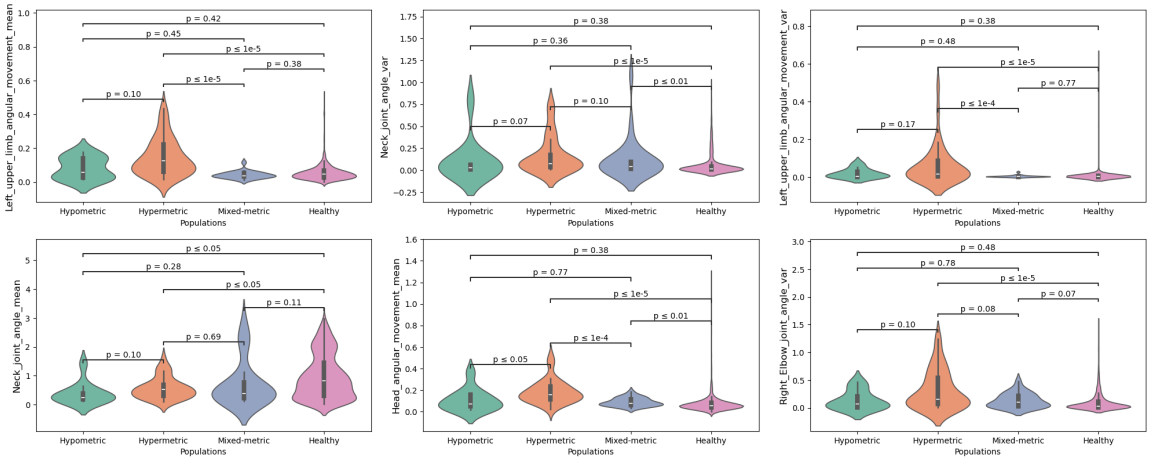

Figure 5: Differential analysis of the top 6 features used by RF for classification.

that comparable or higher predictive performance may be possible when leveraging large-scale pretraining, but such models are less transparent and more difficult to deploy in routine clinical practice.

Future work will focus on refining our feature set, incorporating larger and more diverse clinical cohorts, and explicitly integrating clinician feedback on interpretability and usability. In the longer term, we envisage combining classical models for day-to-day use with more complex deep learning models as research tools to explore subtle patterns in larger, multi-center datasets. Our ultimate goal is to deliver a reliable, objective assessment tool that supports clinicians in monitoring these rare pediatric movement disorders and in assessing treatment outcomes while respecting the practical constraints of ultra-rare disease research.

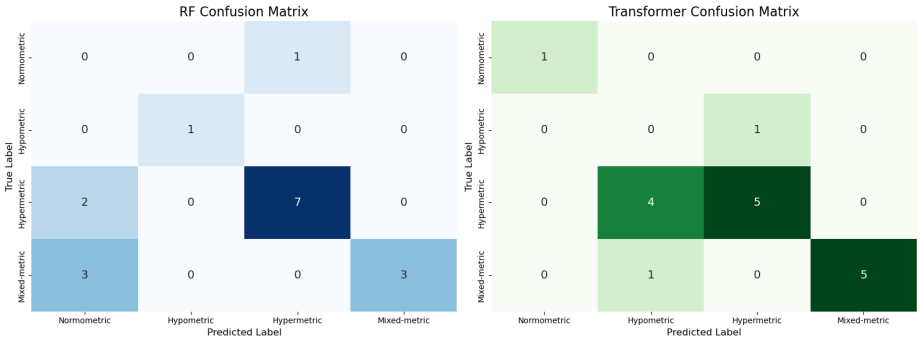

Figure 6: Confusion matrices for RF and transformer results. The RF model commonly misclassifies mixed-metric samples as normometric.

7. Acknowledgment

This research was funded by UCB. Chengliang Dai and Phil Scordis are employees of UCB and may hold shares and/or stock options in UCB. We thank the Tess Research Foundation and Grace Sciences for sponsoring the biomarker discovery trials NCT04681781, NCT06144957, and NCT03834987, during which the videos were collected. We also thank Kasper Claes for his early work in designing and implementing the project's initial architecture.

References

Jeffrey T Anderson, Jan Stenum, Ryan T Roemmich, and Rujuta B Wilson. Validation of markerless video-based gait analysis using pose estimation in toddlers with and without neurodevelopmental disorders. *Frontiers in Digital Health*, 7:1542012, 2025.

Beneufit Inc. Tracer – Beneufit. https://beneufit.com/tracer. Accessed: 2025-06-08.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

Cynthia L. Comella, Sue Leurgans, J. Wuu, Glenn T. Stebbins, and T. Chmura. Rating scales for dystonia: a multicenter assessment. *Movement Disorders*, 18(3):303–312, 2003. doi: 10.1002/mds.10377. The Dystonia Study Group.

Mark Endo, Kathleen L Poston, Edith V Sullivan, Li Fei-Fei, Kilian M Pohl, and Ehsan Adeli. Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–139. Springer, 2022.

Steven J. Frucht, Sue E. Leurgans, Mark Hallett, and Stanley Fahn. The unified myoclonus rating scale. *Advances in Neurology*, 89:361–376, 2002.

Brock Futrell, Christopher Alexander Malaya, David M. Diaz, Carlos Alfaro, Hannah E. Gustafson, S. Chandrasekaran, R. M. Phatak, Bernhard Suter, and Charles S. Layne. Quantifying kinematic tremor in an ngly1-deficient individual: A case study. *Case Reports in Clinical Medicine*, 13(1):25–36, 2024. doi: 10.4236/crcm.2024.131003. URL https://www.scirp.org/journal/paperinformation?paperid=130612.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

Muhammad Hassan Khan, Manuel Schneider, Muhammad Shahid Farid, and Marcin Grzegorzek. Detection of infantile movement disorders in video data using deformable part-based model. *Sensors*, 18(10):3202, 2018. doi: 10.3390/s18103202. URL https://www.mdpi.com/1424-8220/18/10/3202.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

Marieke Johanna Kuiper, Loïs Vrijenhoek, Rick Brandsma, Roelineke J Lunsing, Huibert Burger, Hendriekje Eggink, Kathryn J Peall, Maria Fiorella Contarino, Johannes D Speelman, Marina AJ Tijssen, et al. The burke-fahn-marsden dystonia rating scale is age-dependent in healthy children. *Movement disorders clinical practice*, 3(6):580–586, 2016.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

Tongyi Luo, Jia Xiao, Chuncao Zhang, Siheng Chen, Yuan Tian, Guangjun Yu, Kang Dang, and Xiaowei Ding. Weakly supervised online action detection for infant general movements. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer, 2022. URL https://arxiv.org/abs/2208.03648.

Mohammad Mahdavian, Payam Nikdel, Mahdi TaherAhmadi, and Mo Chen. Stpotr: Simultaneous human trajectory and pose prediction using a non-autoregressive transformer for robot follow-ahead. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9959–9965. IEEE, 2023.

Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2276–2284, 2021. URL https://openaccess.thecvf.com/content/ICCV2021W/SoMoF/html/Martinez-Gonzalez_Pose_Transformers_POTR_Human_Motion_Prediction_With_Non-Autoregressive_Transformers_ICCVW_2021_paper.html.

Avinash Ajit Nargund and Misha Sra. Spotr: Spatio-temporal pose transformers for human motion prediction. *arXiv preprint arXiv:2303.06277*, 2023. URL https://arxiv.org/abs/2303.06277.

Yuyang Quan, Chencheng Zhang, Rui Guo, and Xiaohua Qian. Causality-informed fusion network for automated assessment of parkinsonian body bradykinesia. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer, 2024. URL https://papers.miccai.org/miccai-2024/paper/4133_paper.pdf.

T Schmitz-Hubsch, S Tezenas Du Montcel, L Baliko, J Berciano, S Boesch, Chantal Depondt, P Giunti, C Globas, J Infante, J-S Kang, et al. Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology*, 66(11):1717–1720, 2006.

Jan Stenum, Kendra M Cherry-Allen, Connor O Pyles, Rachel D Reetzke, Michael F Vignos, and Ryan T Roemmich. Applications of pose estimation in human health and performance across the lifespan. *Sensors*, 21(21):7315, 2021.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

Wei Tang, Peter M. A. van Ooijen, Deborah A. Sival, and Natasha M. Maurits. 2d gait skeleton data normalization for quantitative assessment of movement disorders from free-hand single camera video recordings. *Sensors*, 22(11):4245, 2022. doi: 10.3390/s22114245. URL https://www.mdpi.com/1424-8220/22/11/4245.

P. Trouillas, T. Takayanagi, M. Hallett, R. D. Currier, S. H. Subramony, K. Wessel, A. Bryer, H. C. Diener, S. Massaquoi, C. M. Gomez, P. Coutinho, M. Ben Hamida, G. Campanella, A. Filla, L. Schut, D. Timann, J. Honnorat, N. Nighoghossian, and B. Manyam. International cooperative ataxia rating scale for pharmacological assessment of the cerebellar syndrome. the ataxia neuropharmacology committee of the world federation of neurology. *Journal of the Neurological Sciences*, 145(2):205–211, 1997. doi: 10.1016/s0022-510x(96)00231-6.

Edward P Washabaugh, Thanikai Adhithiyan Shanmugam, Rajiv Ranganathan, and Chandramouli Krishnan. Comparing the accuracy of open-source pose estimation methods for measuring gait kinematics. *Gait & posture*, 97:188–195, 2022.

J Xu, M Zhang, EA Turk, et al. Fetal pose estimation in volumetric mri using a 3d convolution neural network. med image comput comput assist interv 11767: 403–410, 2019.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Haozheng Zhang, Edmond S. L. Ho, Xiatian Zhang, and Hubert P. H. Shum. Pose-based tremor classification for parkinson's disease diagnosis from video. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 489–499. Springer, 2022. doi: 10.1007/978-3-031-16440-8_47. URL https://arxiv.org/abs/2207.06828.

Molin Zhang, Junshen Xu, Esra Abaci Turk, P Ellen Grant, Polina Golland, and Elfar Adalsteinsson. Enhanced detection of fetal pose in 3d mri by deep reinforcement learning

with physical structure priors on anatomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 396–405. Springer, 2020.