

RESEARCH ARTICLE OPEN ACCESS

Deep Learning Pose Estimation for Phenotyping of Co-Occurring Hyperkinetic Movement Disorders

Laura Cif^{1,2}  | Diane Demailly^{2,3} | Gabriella A. Horváth⁴ | Juan Dario Ortigoza Escobar^{5,6,7}  | Nathalie Dorison⁸ | Mayté Castro Jiménez¹ | Cécile A. Hubsch¹ | Thomas Wirth^{9,10,11}  | Gun-Marie Hariz¹² | Sophie Huby¹³ | Morgan Dornadic¹³ | Zohra Souei^{2,14}  | Muhammad Mushhood Ur Rehman¹⁵  | Simone Hemm¹⁶ | Mehdi Boulaymen² | Eduardo M. Moraud^{17,18} | Jocelyne Bloch^{17,19} | Xavier Vasques² ✉

¹Service of Neurology, Department of Clinical Neurosciences, Lausanne University Hospital (CHUV), University of Lausanne (UNIL), Lausanne, Switzerland | ²Institut du Neurone, Montferrier sur Lez, France | ³Department of Neurology, Clinique Beau Soleil, Institut Mutualiste Montpelliérain, Montpellier, France | ⁴Department of Pediatrics, British Columbia Children's Hospital, Vancouver, British Columbia, Canada | ⁵Movement Disorders Unit, Pediatric Neurology Department, Institut de Recerca, Hospital Sant Joan de Déu, Barcelona, Spain | ⁶European Reference Network for Rare Neurological Diseases (ERN-RND), Barcelona, Spain | ⁷U-703 Centre for Biomedical Research on Rare Diseases (CIBER-ER), Instituto de Salud Carlos III, Barcelona, Spain | ⁸Pediatric Neurosurgery Department, CCMR Neurogenetique, European Reference Network Brainteam Member, Rothschild Foundation Hospital, Paris, France | ⁹Department of Neurology, University Hospital of Strasbourg, Strasbourg, France | ¹⁰Strasbourg Neuroscience Institute, Strasbourg University, Strasbourg, France | ¹¹Institute of Genetics and Cellular and Molecular Biology, INSERM-U964, CNRS-UMR7104, University of Strasbourg, Illkirch-Graffenstaden, France | ¹²Department of Clinical Neuroscience, Umeå University, Umeå, Sweden | ¹³Department of Neurology, CHU Montpellier, Montpellier, France | ¹⁴Department of Neurosurgery, Military University Hospital of Sfax, Sfax, Tunisia | ¹⁵Edinburgh Medical School, University of Edinburgh, Edinburgh, Scotland, UK | ¹⁶Institute for Medical Engineering and Medical Informatics, School of Life Sciences, University of Applied Sciences and Arts Northwestern Switzerland, Muttens, Switzerland | ¹⁷NeuroRestore, Ecole Polytechnique Fédérale de Lausanne, University Hospital Lausanne, University of Lausanne, Lausanne, Switzerland | ¹⁸Neuro-X Institute, Ecole Polytechnique Fédérale de Lausanne, Geneva, Switzerland | ¹⁹Department of Neurosurgery, Lausanne University Hospital (CHUV), University of Lausanne (UNIL), Lausanne, Switzerland

Correspondence: Laura Cif (lauracif@institutduneurone.fr)

Received: 18 March 2026 | **Revised:** 13 June 2026 | **Accepted:** 22 June 2026

Keywords: co-occurring | deep learning | hyperkinetic movement disorders | phenotyping | pose estimation

ABSTRACT

Objective: To explore whether routine outpatient video combined with deep learning-based pose estimation and clinically interpretable kinematic features can support multi-label phenotyping of co-occurring hyperkinetic movement disorders (HMDs).

Methods: In this exploratory single-centre proof-of-concept study, videos from 21 patients with HMDs and 4 healthy controls were processed with markerless pose estimation (YOLOv8) and 2-dimensional keypoint trajectories transformed into kinematic descriptors spanning statistical, temporal, spectral, and complexity domains. Ten-second windows were aligned to expert annotations for eight hyperkinetic phenomenologies. Conventional supervised classifiers were trained on these tabular descriptors. Window-level predictions were aggregated to the patient level, and label-specific thresholds were tuned on training participants only.

Results: In patient-level multi-label performance reporting, (i) the best single pipeline selected by discrimination (StandardScaler + MLP) achieved a macro-AUPRC of 0.821 ± 0.019 and a macro-receiver operating characteristic area under the curve of 0.830 ± 0.029 . (ii) The best single pipeline selected by Hamming accuracy (MinMaxScaler + SVM) reached 0.764 ± 0.041 . (iii) Under prespecified nested cross-validation with per-label model selection within training folds (primary analysis), macro-AUPRC was 0.717 ± 0.030 , macro-AUROC was 0.767 ± 0.069 , Hamming accuracy was 0.764 ± 0.014 and patient-label agreement was 153/200 (76.5%). (iv) Post hoc per-label selection of the best-performing pipeline defined an exploratory upper bound of 172/200 (86.0%).

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Annals of Clinical and Translational Neurology* published by Wiley Periodicals LLC on behalf of American Neurological Association.

Interpretation: In this exploratory study, a hybrid pipeline combining deep learning pose estimation with feature-engineered supervised classification produced encouraging patient-level multi-label performance for co-occurring HMDs. These findings are proof-of-concept; external, multicentre, prospective validation is required before clinical or trial use.

1 | Introduction

In both adult and pediatric neurology, movement disorders (MDs) are broadly classified from a phenomenological perspective as either hyperkinetic or hypokinetic MDs. Accurate, updated epidemiological data on MDs help to raise awareness, improve recognition, and facilitate timely diagnosis and treatment that remains especially symptomatic [1]. Despite the progress achieved for their clinical characterization and genetic diagnostic [2–4], the recognition of MDs' phenomenology remains challenging during clinical examination with disagreement even between MDs specialists, particularly for the complex presentations of many neurodevelopmental, neurodegenerative and monogenic disorders. The current classification of genetic MDs distinguishes between isolated, combined, and mixed categories [5].

Traditionally, clinicians identify hyperkinetic MDs through direct observation and assess them by using standardized clinical rating scales that usually focus on individual motor phenomena (e.g., scales for dystonia, tremor or tics). Despite their widespread use, they may not capture the full spectrum of motor phenomenology for each patient [6, 7] and have inherent limitations, notably inter-rater variability, limited reproducibility, and the need for specialized training and expertise [8]. Additionally, some scales focus on specific diseases [9] without potential for generalization or application to other conditions. By contrast, other scales [10] may have broader applicability to other developmental disorders. Currently, no available composite assessment tools enable the simultaneous identification and monitoring of MDs over different diseases associating multiple hyperkinetic MDs and through lifespan, in both children and adults. Accurate and objective assessment of hyperkinetic MDs is critical not only for diagnosis, but also for monitoring, facilitating effective patient care, and the development of targeted, symptom-specific interventions [11].

Recent advances in machine learning (ML) [12] enable objective movement analysis from noninvasive clinical videos via markerless pose estimation. This provides anatomically grounded time series that can support scalable phenotyping beyond subjective bedside ratings [8, 13–15]. Research leveraging deep learning (DL) has shown progress in various clinical applications, including gait analysis, rehabilitation assessment [8], tic detection, and tremor quantification [16].

Despite various developments, existing ML approaches for detecting hyperkinetic MDs remain limited in scope, focusing on binary classification of individual hyperkinetic MDs and frequently restricted to short, task-specific video segments [8]. While most existing ML models for MD analysis rely on shallow feature representations, primarily using raw joint coordinates or velocity, they often fail to capture the temporal complexity and variability inherent to a pathological, involuntary movement [17–22]. One unmet need remains the identification and quantification of combined MDs, both in neurology outpatient clinics and through remote

assessments, to support disease monitoring and therapeutic decision-making. The field prospect is to extend DL developments for enhanced accuracy of phenotype classification, the cornerstone of both the diagnostic and treatment process [23].

We therefore developed and evaluated, in a single-centre proof-of-concept cohort, a pose-based hybrid pipeline that couples deep learning-based markerless pose estimation with feature engineering of clinically interpretable kinematic descriptors and conventional supervised classifiers (gradient-boosted decision trees, support-vector machines, logistic regression, random forests, k -nearest neighbors, and a multilayer perceptron). We evaluated, in a hierarchy of analyses with increasing clinical complexity, (i) window-level binary screening of symptom-expressing patient windows against healthy control windows, (ii) patient-level multi-label phenotyping via p90 percentile aggregation of window probabilities with control-aware thresholding, and (iii) decision-level interpretability linking predictions to kinematic families and anatomical regions. Because deep learning is used only in the upstream pose-estimation step and the downstream classification operates on tabular kinematic summaries computed within each 10-s window, the framework is best characterized as feature engineering-based supervised multi-label classification rather than as true sequential or temporal deep learning.

2 | Materials and Methods

2.1 | Participant Recruitment and Video Acquisition

Twenty-one consecutive patients (range 17–75 years, 12 females) diagnosed with isolated or combined MDs including dystonia and associated hyperkinetic MDs (i.e., tremor, myoclonus, chorea, tics, athetosis, stereotypies, ballismus) and four healthy controls were enrolled in the study. The demographics and clinical characteristics are detailed in (Table 1). Video recordings were prospectively collected from routine clinical visits at the movement disorders clinic of the Service of Neurology, Montpellier Beau Soleil Clinic, France. Regulatory approvals were granted for the study (CESSRESS 22075132 Bis and CNIL 2238428). All the participants underwent video recording sessions aimed to be performed under standardized condition, camera positioning, and consistent distance between participants and the camera to minimize variability. The total duration of video recordings varied among patients due to the impact of the MDs severity on task execution and completion time. Videos were captured using smartphone cameras capable of maintaining consistent frame rates (minimum of 30 frames per second, FPS) and resolution (1920 × 1080 pixels). Participants performed predefined, standardized motor tasks known to elicit or exacerbate motor phenomenology, including but not limited to resting, posture-holding tasks, and voluntary movements such as finger tapping, arm flexion/

TABLE 1 | Summary of patient demographics, diagnoses, and video data characteristics.

Patient	Age (years)	Gender	Diagnostic	Motor phenomenology	Number of 10-s windows (from annotations)	Total video duration	Number of frames
1	51	M	Combined dystonia	Cervical dystonia and action tremor	97	16 min 32 s	29,787
2	55	M	Combined dystonia	Cervical dystonia, ataxia, myoclonus	103	17 min 27 s	31,499
3	27	F	Hypoxic cerebral palsy	Generalized dystonia, spasticity, athetosis	131	22 min 00 s	39,636
4	36	M	<i>DIP2B</i> gene related mixed movement disorder	Generalized dystonia, chorea, ataxia and athetosis	150	25 min 35 s	46,095
5	34	M	<i>DIP2B</i> gene related mixed movement disorder	Generalized dystonia, chorea, with ataxia	121	20 min 19 s	36,617
6	69	F	Combined dystonia	Progressive cranial dystonia, mild ataxia, action tremor	107	18 min 09 s	32,695
7	25	M	Hyperkinetic movement disorders	Myoclonus, tics, tremor, dytonia	109	18 min 31 s	33,355
8	60	F	<i>LRRK2</i> gene related Parkinson's disease	Parkinsonism and dystonia	97	16 min 16 s	29,323
9	73	F	Dystonia	Generalized dystonia, mild ataxia	114	19 min 02 s	34,288
10	71	F	Chorea-dystonia	Generalized chorea, focal dystonia	127	21 min 20 s	38,396
11	20	F	<i>ADCY5</i> gene related mixed movement disorder	Generalized dystonia, chorea, myoclonus	95	15 min 51 s	28,556
12	75	F	Chorea and ataxia	Generalized dystonia, chorea, ataxia	115	19 min 16 s	34,702
13	62	F	Dystonia	Generalized dystonia	104	17 min 24 s	31,337
14	22	M	Tremor prominent movement disorder	Action tremor, dystonia	108	18 min 01 s	32,456
15	63	F	Chorea-dystonia	Dystonia, chorea	115	19 min 18 s	34,736
16	69	F	Combined dystonia	Generalized dystonia, mild parkinsonism	105	17 min 36 s	31,699
17	70	F	Dystonia	Generalized dystonia	114	19 min 07 s	34,433
18	37	M	<i>TTPA</i> gene related ataxia with vitamin E deficiency (AVED)	Ataxia, cervical dystonia	119	19 min 58 s	35,965

(Continues)

TABLE 1 | (Continued)

Patient	Age (years)	Gender	Diagnostic	Motor phenomenology	Number of 10-s windows (from annotations)	Total video duration	Number of frames
19	26	M	Neurodevelopmental disorder and tardive dystonia	Dystonia, akathisia	105	17 min 22 s	31,275
20	17	F	Myoclonus dystonia	Dystonia, myoclonus	99	16 min 27 s	29,618
21	23	M	<i>TIMM8</i> gene related combined movement disorder	Generalized dystonia, tremor, spasticity, tics	102	16 min 54 s	30,428

Note: Etiology was confirmed in 33% of cases, an expected finding consistent with rates of etiologic diagnostic confirmation reported in the literature.

extension, target reaching, and writing. The complete content of the conditions and tasks captured in the video recordings is provided as Supporting Information S1. This acquisition setup was intentionally designed to reflect real-world clinic conditions (smartphone, routine exam).

2.2 | Clinical Assessment and Annotation Protocol

Patients were clinically assessed by two physicians with expertise in MDs blinded to each other's assessments. They independently assessed the presence or absence of eight hyperkinetic MDs: dystonia, tremor, myoclonus, chorea, athetosis, tics, ballismus, and stereotypies. Methodological considerations related to phenomenological distinctions are reported in Supporting Information S2.

Following the independent ratings, both clinicians met to review any discrepancies and establish a final consensus label (presence vs. absence) for each hyperkinetic MD, which served as the ground truth for subsequent analyses. Ten-second temporal windows defined based on clinician annotations specifying start and end times were automatically extracted from the complete video recordings of all twenty-one enrolled subjects. Fixed 10-s windows duration has been selected as a practical compromise since long enough to provide stable statistical, spectral, and complexity features for classification, while still short enough to capture the temporal fluctuations typical of hyperkinetic MDs. Clinicians annotated video segments using a predefined rule, marking clear presence ("1"), absence ("0"), or uncertainty ("2") for each hyperkinetic MD. Only segments marked explicitly as present or absent were included in the subsequent analysis; segments containing uncertain annotations were discarded to maintain data integrity. Data cleaning procedures were applied to remove any segments containing missing or non-numeric values. The per-phenotype frequency of uncertain and excluded windows, together with the retained class balance, is reported in Table S1. Table 1 summarizes the video data characteristics for all patients. Each hyperkinetic MD pattern was evaluated across three distinct conditions: at rest, during posture maintenance, and with actions. Healthy control subjects underwent clinical examinations to exclude any MD or other neurological or extra-neurological findings that would preclude their inclusion in the

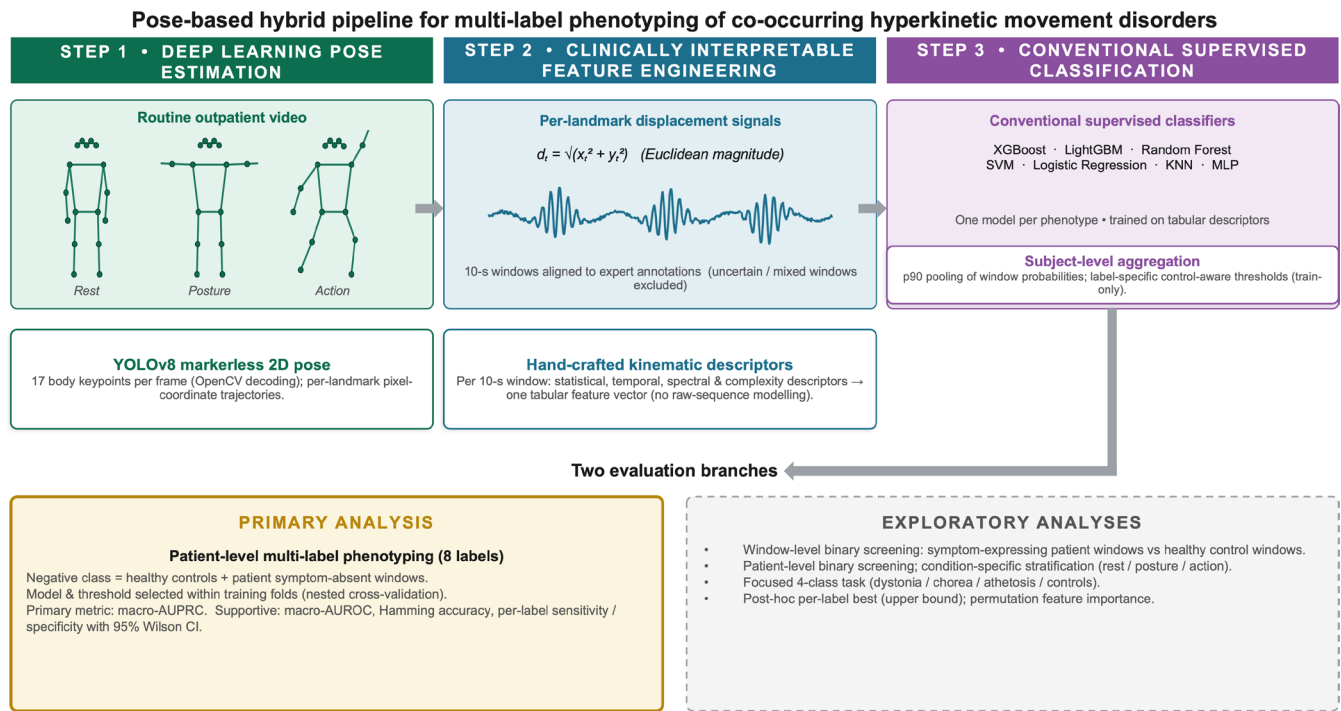
control group. The consensus label (prim_label) served as the operational ground truth for all subsequent analyses. Inter-rater agreement at the window level prior to consensus is reported per phenotype in Table S2 (Cohen's κ).

2.3 | Video Processing and Feature Extraction Pipeline

The video-to-classification pipeline for semi-automated, annotation-guided HMD phenotyping is presented in Figure 1. Each video recording was processed individually, frame by frame, to ensure consistent and reproducible extraction of spatiotemporal features (Figure 1). Frame timestamps (30 frames per second) were derived from the video FPS to align pose time series with clinical annotations at 10-s resolution. Feature extraction was implemented in Python (version 3.10.17), leveraging established computational libraries [24–32] to ensure analytical accuracy and methodological reproducibility (Supporting Information S3). To diagnose hyperkinetic MDs phenomenology from routine clinic videos, we relied on a keypoint-based representation of human pose, which is conceptually close to what clinicians observe at the bedside: the trajectories of recognizable anatomical landmarks over time (e.g., wrists, elbows, shoulders, hips, knees, ankles, and cranial points). Pose estimation was performed using YOLOv8 (*yolov8x-pose-p6.pt*), a state-of-the-art DL architecture for 2D keypoint detection in unconstrained settings. We applied YOLOv8 2D pose estimation to obtain 17 keypoints per frame. Keypoint trajectories were stored with timestamps and used to derive displacement signals for feature computation (Table 2).

2.4 | Derived Additional Time-Series Features From Landmark Trajectories

Beyond raw keypoint coordinates, we derived clinically motivated time-series descriptors from each landmark trajectory to capture movement properties routinely assessed at the bedside, namely the magnitude and variability of movement, the presence of oscillatory pattern, and the degree of irregularity or complexity. For each anatomical landmark tracked in 2D



Semi-automated, annotation-guided pipeline: pose estimation and feature engineering are fully automatic, whereas the 10-s window boundaries and per-window phenotype labels are clinician-provided (consensus of two expert raters). Deep learning is used only for the upstream pose-estimation step; downstream phenotyping is feature-engineering–based supervised classification rather than end-to-end sequence modelling.

FIGURE 1 | Video-to-classification pipeline for semi-automated, annotation-guided HMD phenotyping. Clinical videos are decoded with OpenCV and analyzed with YOLOv8 (the deep learning pose-estimation step) to estimate 17 2D keypoints per frame. Keypoints are converted to pixel coordinates, transformed into displacement signals, and segmented into consecutive 10-s windows aligned to expert window-level annotations. For each window, statistical, temporal, spectral, and complexity descriptors are computed to form a feature vector (the feature-engineering step), which is then classified by conventional supervised models (the classification step); window boundaries and labels are clinician-provided, hence “annotation-guided”. Uncertain segments are excluded. Two datasets are derived: (i) binary screening (symptom-expressing patient windows vs. control windows) and (ii) an 8-label multi-label set with per-window symptom vectors. In the multi-label branch, one model is trained per label; window probabilities are aggregated to the subject level (e.g., p90) and label-specific thresholds are tuned on training subjects under control-aware constraints. Performance is reported with ROC-AUC/AUPRC and multi-label metrics (Hamming, Jaccard).

with pixel coordinates (x_t, y_t) at frame t , we constructed a scalar displacement signal $d_t = \sqrt{x_t^2 + y_t^2}$ (Euclidean magnitude relative to the image origin). Trajectories were segmented into fixed-length windows aligned to expert labels, and within each window we computed a panel of descriptors designed to summarize clinically meaningful aspects of movement. First, to capture amplitude and instability, we extracted distributional summaries of $\{d_t\}$ including central tendency (mean and median), dispersion (standard deviation, variance, interquartile range, and range), and extremes (minimum and maximum). Within each 10-s window, we derived interpretable descriptors intended to match bedside constructs: postural set-point and sustained bias (distributional summaries and short-horizon rolling means), excursions and paroxysmal change (extremes and event-sensitive derivative proxies), rhythmicity (FFT peak frequency/amplitude, excluding the DC component), and irregularity/complexity beyond variance (Higuchi fractal dimension and permutation entropy). Together, these measures separate tonic deviation, oscillatory burden, abrupt jerky movements, and continuously evolving trajectories from the same pose-derived displacement signals. Mathematical definitions, parameter choices, and the complete feature list are provided in the Supporting Information S3. For each window and each landmark displacement signal, this implementation computes 19 primary descriptors (distributional + temporal + spectral +

entropy + complexity) and 3 rolling-mean descriptors, yielding 22 features per landmark per window. With $J = 17$ landmarks, this corresponds to $22 \times 17 = 374$ derived features per window in the multi-label pipeline. Feature names followed a transparent convention $[side]_{[landmark]}_{distance}_{[metric]}$ (e.g., `right_shoulder_distance_max` or `left_knee_distance_perm_entropy`), facilitating direct clinical interpretation and mapping to anatomical regions.

To align explanations with clinical reasoning, we organized features into kinematic families (posture/bias, excursions, variability, rhythmicity, directionality, irregularity/complexity) and anatomical regions (cranial, upper limb, lower limb). We quantified feature importance for decision at the patient level using permutation on held-out outer test folds only, preserving the full inference pipeline (window scores, p90 pooling, and label-specific control-aware thresholds). Importance was defined as the increase in patient-level error $(FP+FN)/N$ induced by permuting one feature and aggregated across folds using a stability criterion. We deliberately chose a generic, phenotype-agnostic kinematic feature set rather than phenotype- or task-specific clinical metrics. The targeted setting is co-occurring multi-label phenotyping across eight hyperkinetic phenomenologies and three examination conditions, for which there is no common validated panel of

TABLE 2 | Overview of extracted features from video pose estimation and time series analysis.

Feature type	#	Examples/names	Methodological role	Clinical interpretation (bedside meaning)
Anatomical landmarks (2D keypoints)	17	Nose, eyes, ears, L/R shoulder, elbow, wrist, hip, knee, ankle	Frame-wise detection of standardized body landmarks	Topography: where movement occurs (distal vs. proximal; cranial, axial vs. appendicular involvement)
Landmark displacement signals (image-plane)	17	$d_t = \sqrt{(x_t^2 + y_t^2)}$ per landmark	Converts 2D trajectories into robust scalar time series per segment	Movement amplitude proxy: how much each body part visibly moves over time
Statistical descriptors	12	Mean, std., median, min/max, range, IQR, variance, skewness, kurtosis, energy, zero-crossings	Summarize distribution and variability of each displacement signal	Severity/variability: sustained large amplitudes versus fluctuating/bursty motion; frequent reversals can suggest oscillatory/intermittent patterns
Temporal/dynamic features	3 (+ rolling means)	Slope, mean absolute acceleration, rolling means ($w = 3, 5, 7$)	Capture trends, abrupt changes, and short-term variability	Abruptness/intermittence: jerk or burst-like events (myoclonus, tics), progressive change during tasks (task exacerbation/fatigability)
Spectral features	2	Dominant frequency, peak amplitude (FFT)	Quantify periodicity/oscillation of displacement signals	Rhythmicity: presence of a dominant oscillatory component consistent with tremor-like patterns or repeated stereotyped movements
Non-linear/complexity features	2	Higuchi fractal dimension, permutation entropy	Quantify irregularity/complexity beyond simple variance	Irregular/chaotic motion: higher complexity often aligns with non-rhythmic hyperkinesia (e.g., chorea-like components) versus more patterned phenomenology
Metadata	4	Timestamp, frame index, video ID, condition (Rest/Posture/Action)	Temporal alignment and stratified analyses	Context of observation: rest versus posture versus action for clinically meaningful stratification

Note: This table summarizes the different types of features extracted from the pose estimation and time series processing pipeline applied to each video recording. For each feature type, the table reports the number of features computed, their specific names or examples, and a brief description of their methodological role in the dataset. This multi-level feature extraction approach enables a comprehensive and physiologically meaningful characterization of movement dynamics for subsequent machine learning analysis.

phenotype-specific clinical metrics. Generic descriptors based on landmark displacement preserve comparability across phenotypes and conditions, are parameter-efficient (essential in the small-cohort regime), and map onto bedside constructs such as movement burden, postural set-point, excursion, rhythmicity and complexity. We anticipate that the generic descriptors used here will be complemented in future work by phenotype-specific clinical features (spatiotemporal gait parameters, tremor-band power and centre frequency for tremor, jerk-spike statistics for myoclonus, segment-level twist for dystonia, finger-tap kinematics for action chorea), most naturally within hierarchical pipelines combining a generic stage with phenotype-dedicated downstream modules.

2.5 | Machine Learning-Based Hyperkinetic MDs Classification

The primary analysis was patient-level multi-label phenotyping with p90 percentile aggregation of window probabilities and label-specific control-aware thresholds, evaluated under nested cross-validation with per-label model selection within training folds, with macro-AUPRC as the primary performance metric. Patient-level sensitivity, specificity (with 95% Wilson score confidence intervals), Hamming accuracy, and macro-AUROC under nested cross-validation were prespecified supportive metrics. The window-level binary screening, patient-level aggregated binary screening, condition-specific stratification, focused four-class analysis, post hoc per-label best (exploratory upper bound), and decision-level permutation feature-importance analyses were exploratory (Table 3). All classifier, preprocessing, and threshold decisions were made within training folds only; outer test folds were used solely for reporting. Because multiple model families and operating points were evaluated, configurations selected as “best” should be regarded as upper-end estimates conditional on this cohort. Throughout the manuscript, the framework is semi-automated and annotation-guided: pose estimation and feature engineering are fully automatic, but the temporal definition of the 10-s windows and the per-window supervision labels are provided by expert clinicians, and a fully end-to-end version would additionally require automatic detection and segmentation of motor-examination tasks within long routine recordings.

2.6 | Binary Classification of Individual Hyperkinetic MD Presence Versus Absence

2.6.1 | Window-Based Binary Classification

As a preliminary benchmarking step, we conducted a systematic evaluation of preprocessing and supervised learning strategies (Supporting Information S3) across the eight hyperkinetic MDs phenomenologies. For each symptom, we formulated a binary window-level classification problem in which a 10-s window was labeled hyperkinetic MD-present if at least one frame within the window was annotated as symptom-positive (label = 1). To construct a conservative negative class in this screening-oriented setting, windows labeled hyperkinetic MD-absent (label = 0) were retained only for healthy controls, whereas windows labeled 0 from patients were not included. This design therefore

evaluates symptom-expressing patient windows against control windows, under routine acquisition conditions. In addition, we derived a subject-level decision by aggregating out-of-fold window predictions within each subject using a majority-vote rule ($\geq 50\%$ positive windows). We report subject-level correct classification rate, sensitivity, and specificity based on this aggregated decision. This aggregation reduces the influence of sporadic window-level errors when multiple windows are available per subject and provides a complementary subject-level summary alongside window-level performance. Since in this window-based step the negative class is derived exclusively from controls, results should be interpreted within this screening-versus-control framework rather than as symptom absence detection within patients.

2.6.2 | Patient-Level Binary Classification Using Aggregated Time Series Features

To complement window-level pre-analyses with a subject-level summary, we implemented a patient-based binary classification framework in which each subject was represented by a single feature vector per symptom. For each recording and target symptom, we constructed a symptom time series by selecting and concatenating windows according to rater annotations. For patients, only windows in which the symptom was annotated as present were retained. The retained windows were concatenated to form a single per-subject time series per anatomical landmark, thereby focusing the representation on segments that are clinically informative for the symptom under consideration. Subjects without any retained windows for a given symptom were excluded from the corresponding symptom-specific analysis. From this concatenated signal, we extracted pose-derived time series spanning the same feature families as the window-based pipeline, yielding one aggregated feature vector per subject and symptom. Subjects were assigned a binary label reflecting group membership in this patient-versus-control setting (controls = 0; patients = 1), and models were trained to discriminate symptom-expressing patient recordings from healthy controls.

Feature rescaling was applied within each training fold. The same panel of supervised classifiers was evaluated with exhaustive grid search over predefined hyperparameter grids. For classifiers with potentially uncalibrated probability estimates, probability calibration was performed using Platt scaling (sigmoid calibration) via a calibrated classifier wrapper with internal cross-validation on the training data. Because feature vectors were derived from symptom-positive windows, this patient-level analysis characterizes a “symptom signature” at the subject level rather than evaluating symptom absence within patients; it therefore complements, rather than replaces, the window-level analysis.

2.6.3 | Condition-Specific Hyperkinetic Individual MD Detection

In addition, to reflect clinical examination practice, we stratified recordings into three predefined conditions commonly used during examination: (i) rest, (ii) posture maintenance, and (iii) voluntary movement (e.g., finger-to-nose maneuver). For each

TABLE 3 | Study-design overview.

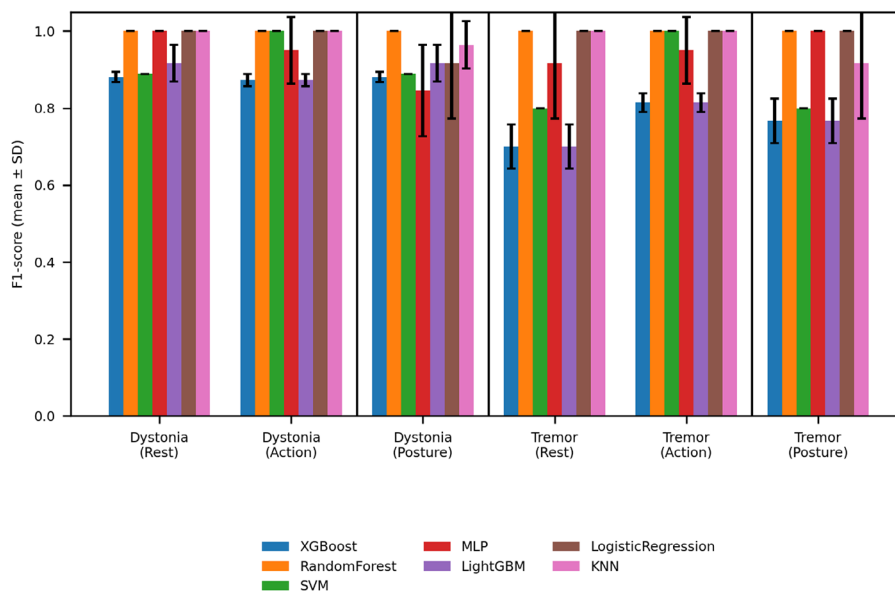
Analysis	Unit of evaluation	Negative-class definition	Aggregation rule	Output/Primary metric
Window-level binary screening (exploratory; Figure 2)	10-s window	Control windows only (patient symptom-absent windows excluded)	None at window level; majority vote $\geq 50\%$ for patient summary	Window $F1_1$, ROC-AUC; patient summary: % correct, sensitivity, specificity
Patient-level binary, symptom-positive windows (exploratory)	Participant	Healthy controls	Concatenation of symptom-positive windows $\rightarrow$ 1 feature vector per participant	Patient accuracy, $F1_1$, ROC-AUC
Condition-specific stratification (rest/posture/action; exploratory)	Participant within condition	Healthy controls within the same condition	Concatenation per condition	$F1$ within condition
PRIMARY: Patient-level multi-label phenotyping (nested cross-validation)	Participant $\times$ 8 labels	Healthy controls + patient symptom-absent windows	p90 percentile of window probabilities; control-aware per-label thresholds tuned on training participants; nested per-label model selection within training folds	Macro-AUPRC (primary); supportive: macro-AUROC, Hamming accuracy, per-label sensitivity/specificity with 95% Wilson CI
Exploratory upper-bound: post hoc per-label best	Participant $\times$ 8 labels	Same as primary	Same as primary, but the best model is selected post hoc per phenotype to minimize errors	Patient-label agreement (172/200); per-label confusion (Table S3a)
Focused 4-class (dystonia/chorea/athetosis/controls; exploratory)	10-s window with unambiguous label	Asymptomatic control windows	Per-label window probability	$F1$, AUROC, AUPRC per class
Permutation feature importance (exploratory)	Participant $\times$ label	—	Permute feature $\rightarrow$ recompute patient-level error using full inference pipeline	Δ error share by kinematic family/anatomical region (Figures 4 and 5)

Note: For each analysis, the unit of evaluation, the definition of the negative class, the aggregation rule used to combine window-level outputs (when applicable), and the output and primary metric are summarized. The PRIMARY analysis is highlighted; all other analyses are explicitly labeled as exploratory.

A

Symptom	Window-level reliability (F1, positive class)	Patient identification (% subjects correct)	Rater consistency (ΔF1; lower is better)
Dystonia	0.86 (R1: 0.84, R2: 0.89)	97% (R1: 94%, R2: 100%)	0.05
Chorea	0.89 (R1: 0.81, R2: 0.96)	100% (R1: 100%, R2: 100%)	0.15
Tics	0.88 (R1: 0.75, R2: 1.00)	100% (R1: 100%, R2: 100%)	0.25
Tremor	0.82 (R1: 0.83, R2: 0.82)	90% (R1: 89%, R2: 92%)	0.01
Athetosis	0.54 (R1: 0.59, R2: 0.49)	86% (R1: 89%, R2: 82%)	0.10
Ballismus	0.18 (R1: 0.14, R2: 0.23)	88% (R1: 100%, R2: 75%)	0.09
Stereotypies	0.54 (R1: 0.54)	88% (R1: 88%)	Single rater
Myoclonus	0.52 (R1: 0.68, R2: 0.36)	78% (R1: 93%, R2: 64%)	0.32

B



C

Symptom	Rater	Scaler	Clf	F1+	F1-	Acc	AUC	Subj-Acc (%)	Patient Sens	Control Spec	Windows (train)	Windows (test)
Dystonia	Rater_1	Robust	RF	0.84 ±0.08	0.73 ±0.04	0.81 ±0.06	0.88 ±0.06	94±24%	0.92	1.00	1992 ±92	473 ±92
Dystonia	Rater_2	Power	RF	0.89 ±0.02	0.84 ±0.06	0.88 ±0.03	0.93 ±0.05	100±0%	1.00	1.00	1606 ±126	402 ±126
Chorea	Rater_1	MinMax	LGBM	0.81 ±0.18	0.93 ±0.08	0.90 ±0.11	0.95 ±0.06	100±0%	1.00	1.00	803 ±109	201 ±109
Chorea	Rater_2	Std	XGB	0.96	0.94	0.95	0.99	100±0%	1.00	1.00	690 ±80	173 ±80
Tics	Rater_1	MinMax	RF	0.75 ±0.25	0.99 ±0.01	0.97 ±0.03	0.96 ±0.04	100±0%	1.00	1.00	387 ±48	97 ±48
Tics	Rater_2	Std	XGB	1.00	1.00	1.00	1.00	100±0%	1.00	1.00	394 ±78	98 ±78
Tremor	Rater_1	Power	RF	0.83 ±0.01	0.84 ±0.05	0.84 ±0.05	0.92 ±0.04	89±31%	0.83	1.00	817 ±92	204 ±92
Tremor	Rater_2	MinMax	RF	0.82 ±0.02	0.87 ±0.01	0.85 ±0.01	0.92 ±0.02	92±27%	0.89	1.00	738 ±117	184 ±117
Athetosis	Rater_1	Std	XGB	0.59 ±0.30	0.92 ±0.02	0.87 ±0.05	0.94 ±0.02	89±31%	0.80	1.00	533 ±87	133 ±87
Athetosis	Rater_2	Std	MLP	0.49 ±0.14	0.86 ±0.05	0.79 ±0.05	0.78 ±0.09	82±39%	0.75	1.00	449 ±63	112 ±63
Ballismus	Rater_1	Power	RF	0.14	0.91	0.84	0.85	100±0%	1.00	1.00	562 ±77	141 ±77
Ballismus	Rater_2	Std	MLP	0.23 ±0.10	0.96 ±0.02	0.92 ±0.04	0.76	75±43%	0.50	1.00	556 ±49	139 ±49
Stereotypies	Rater_1	MinMax	LGBM	0.54 ±0.26	0.96 ±0.02	0.92 ±0.04	0.84 ±0.15	88±33%	0.75	1.00	477 ±84	119 ±84
Myoclonus	Rater_1	MinMax	LGBM	0.68 ±0.14	0.86 ±0.08	0.81 ±0.11	0.88 ±0.09	93±26%	0.90	1.00	534 ±121	134 ±121
Myoclonus	Rater_2	Std	KNN	0.36 ±0.02	0.89 ±0.05	0.81 ±0.07	0.72 ±0.07	64±48%	0.43	1.00	471 ±116	118 ±116

FIGURE 2 | Legend on next page.

FIGURE 2 | Window-based symptom classification performance across rest, posture, and action. (A) Summary of window-based binary symptom detection per phenotype, reporting mean positive-class $F1$ (with rater-specific values), patient-level identification rate (from aggregated window predictions), and inter-rater consistency ($\Delta F1$; lower = higher agreement). Cell colors provide a qualitative cue (green high; orange intermediate; red low; for $\Delta F1$, green low disagreement). (B) Condition-specific performance on 10-s segments (Rest, Posture, Action): Mean positive-class $F1 \pm SD$ across folds for one representative model per classifier family, enabling comparison across model families and motor contexts. (C) For each phenotype and rater (two independent annotation sets), we report the best scaler-classifier pipeline. Performance was evaluated with 5-fold StratifiedGroupKFold cross-validation with subject-level splitting. Window-level metrics ($F1_{1,1}$, $F1_{1,0}$, accuracy, ROC-AUC) are mean $\pm SD$ across folds. Subject-level performance is obtained by majority voting on out-of-fold window predictions ($\geq 50\%$ positive) and reports percent correctly classified subjects, sensitivity in symptom-expressing patients, and specificity in controls. Mean $\pm SD$ train/test window counts are shown. The negative class includes controls only.

condition and symptom, we constructed one condition-specific sample per subject by aggregating the corresponding annotated segments into a single merged time series and extracting the same pose-derived feature set. A subject-condition sample was labeled symptom-positive if the symptom was annotated as present (label=1) in any frame within the merged series. Classifier performance was evaluated using subject-grouped cross-validation within each condition, enabling direct comparison of detection performance across motor tasks.

2.7 | Concurrent Detection of Multi-Label Hyperkinetic MDs

To emulate routine clinical assessment, we evaluated the simultaneous detection of multiple hyperkinetic MDs from the 10-s video windows using an eight-label output space (dystonia, tremor, myoclonus, chorea, athetosis, ballismus, stereotypies, and tics). Each window was assigned an 8-dimensional binary label vector derived from expert annotations (label=1 if the symptom was present at any time within the window; else 0). In this multi-label setting, we explicitly controlled false alarms in asymptomatic recordings by including both the healthy control subjects and symptom-absent windows from the patients, enabling clinically interpretable subject-level readouts under a conservative operating regime.

2.7.1 | Subject-Level Cross-Validation and Fold Balancing

To prevent information leakage, all windows from a given participant were assigned to the same fold. Folds were balanced using patient-level phenotype profiles (presence/absence of each of the 8 phenotypes across that participant's windows) so that individual phenotype prevalence and common co-occurrence patterns were preserved as much as possible across cross-validation splits (MultilabelStratifiedKFold; shuffle with fixed random seed).

2.7.2 | From window probabilities to patient-level multi-label decisions

Each label was modeled as an independent binary task at the window level, yielding per-window probabilities for each symptom. To obtain a clinically actionable patient-level multi-label profile, window probabilities were aggregated within each patient using percentile pooling (90th percentile), producing one

probability per patient and label. Decision thresholds were tuned per label on training patients only using a clinically motivated operating-point policy that minimizes a balanced error criterion under explicit constraints on false positives among control subjects (including label-specific constraints for high-impact phenotypes). Thresholds were then applied unchanged to held-out test patients. In addition, we evaluated an optional nested selection strategy in which the best-performing model configuration could be selected independently for each label within the training data (inner cross-validation), reflecting the heterogeneous learnability and risk profile of the individual phenotypes. Thus, each phenotype had its own threshold, determined from the training data only, to balance sensitivity and specificity while limiting false-positive calls in controls. Similarly, the nested strategy allowed model choice to vary by phenotype within the training data, because phenotypes differ both in detectability and in the clinical consequences of overcalling versus missing a case.

2.7.3 | Focused, clinically relevant scenario: Detection of Dystonia Versus Athetosis Versus Chorea Versus Control

To further assess the pipeline in a clinically relevant scenario of overlapping hyperkinetic MDs observed in dyskinetic "cerebral palsy (one of the most frequently acquired hyperkinetic MDs), we implemented a focused four-class classification task. Ten-second windows were annotated as dystonia, chorea, athetosis, or asymptomatic windows. Only windows with unambiguous labels were retained for analysis.

2.8 | Model Training and Evaluation

Models were evaluated with subject-level cross-validation to prevent leakage. Preprocessing and model hyperparameters were selected within the training data; probability calibration was applied when needed. Full grids and implementation details are provided in Supporting Information S3. Preprocessing consisted of feature scaling (StandardScaler, MinMaxScaler, RobustScaler, and PowerTransformer) when appropriate followed by univariate feature selection using mutual information (SelectKBest; fixed K with minor label-specific adjustments). We evaluated a panel of supervised classifiers commonly used for structured biomedical data: gradient-boosted decision trees (XGBoost, LightGBM), Random Forest, linear models (Logistic Regression, SVM with RBF kernel), instance-based learning (K-Nearest Neighbors), and a multilayer

perceptron (MLP). Hyperparameters were optimized by exhaustive grid search over predefined parameter grids. Class imbalance was handled within each training fold: when the relative difference between class sizes exceeded 20%, we applied random undersampling of the majority class to match the minority class size. For classifiers producing potentially uncalibrated probability estimates, probability calibration was applied using Platt scaling (sigmoid calibration) via a calibrated classifier wrapper with internal cross-validation on the training data. Performance was quantified at the window level using accuracy, class-specific *F1*-scores (reported separately for the positive and negative classes), and ROC-AUC. Specifically, for multilabel evaluation, primary performance was reported at the patient level. Discrimination was quantified using macro-averaged ROC-AUC and macro-averaged average precision (AUPRC), computed label-wise when both classes were present in the test fold and then averaged across eligible labels. For binary predictions, we reported micro- and macro-averaged *F1*-scores and Hamming accuracy for the full eight-label vector. Per-label precision, recall, specificity, and *F1*-score were also reported, and control-only false positives were tracked to verify alignment with the targeted operating regime.

2.9 | Implementation and Computation

The pipelines described in our study were implemented using Python (version 3.10.17). All related Python scripts and results are provided on GitHub: <https://github.com/xaviervasques/CODY>.

3 | Results

3.1 | Binary classification

3.1.1 | Window-Level Detection of Individual Hyperkinetic MDs

The best-performing window-based pipelines for each hyperkinetic MD using 10-s video segments are summarized in Figure 2. Throughout, these window-level binary analyses should be read as a screening contrast between symptom-expressing patient windows (≥ 1 symptom-positive frame within the 10-s window) and healthy control windows; they do not evaluate symptom absence within patient videos, which is addressed only at the patient level in the multi-label analysis. High specificity values therefore quantify how rarely a control window is misclassified as symptomatic, not the model's ability to identify symptom-absent intervals inside complex patients. Across phenotypes and annotation sets, models showed consistently high subject-level specificity after within-subject aggregation (1.00 for all symptoms and raters), indicating that healthy controls were rarely labeled as symptomatic once window-level predictions were integrated at the subject level. Dystonia, chorea, tremor and tics exhibited robust window-level performance and reliable subject identification, consistent with relatively distinctive kinematic signatures: sustained abnormal posturing for dystonia, irregular purposeless excursions for chorea, rhythmic oscillatory structure for tremor, and brief stereotyped bursts for tics. In contrast, athetosis, stereotypies, myoclonus and ballismus were more challenging at the 10-s window scale, consistent with (i)

slower, continuous and often mixed-pattern movements (athetosis), (ii) context-dependent and heterogeneous repetitive behaviors (stereotypies), and (iii) highly intermittent, short-duration, sudden, abrupt events that can be more difficult to time-stamp consistently (myoclonus and ballismus). The lower performance for these latter phenotypes is also consistent with their lower window-level inter-rater agreement (Cohen's $\kappa = 0.116$ for tics, 0.220 for ballismus, 0.259 for athetosis, 0.366 for myoclonus, and 0.000 for stereotypies; Table S2), and is expected from the limitations of the 2D pose representation for fine distal and subtle facial signals as well as from exclusive video-based assessments.

3.1.2 | Symptom-Specific Performance Insights

Dystonia demonstrated strong and consistent performance across annotation sets ($F1_1 = 0.84$ vs. 0.89; patient sensitivity = 0.92 vs. 1.00), supporting the presence of stable pose-derived kinematic signatures that are reproducibly captured despite rater-specific labeling policies. Chorea was also robust ($F1_1 = 0.81$ vs. 0.96) with perfect patient-level sensitivity in both annotation sets; the larger between-annotation performance difference at the window level ($\Delta F1 \approx 0.15$) is consistent with subtle differences in labeling thresholds for borderline choreiform movements. Tremor performance was highly concordant across annotation sets ($F1_1 \approx 0.82$ –0.83; $\Delta F1 \approx 0.01$) with strong patient identification (≈ 0.89 –0.92), in keeping with a relatively rhythmic motor signature. Tics achieved excellent patient-level detection in both annotation sets (patient sensitivity = 1.00), despite greater variability at the window level for Rater 1 ($F1_1 = 0.75 \pm 0.25$). In contrast, athetosis yielded only moderate window-level performance ($F1_1 = 0.59$ vs. 0.49) with reasonable patient sensitivity (0.80 vs. 0.75). Ballismus was the most challenging phenotype ($F1_1 = 0.14$ vs. 0.23), with marked annotation-set dependence in patient sensitivity (1.00 vs. 0.50), compatible with rarity and episodic expression as well as ambiguity in defining segments with ballismus. Stereotypies showed moderate performance in the single available annotation set ($F1_1 = 0.54 \pm 0.26$; patient sensitivity = 0.75), reflecting heterogeneous and context-dependent repetitive patterns. Finally, myoclonus was strongly annotation-set dependent ($F1_1 = 0.68$ vs. 0.36; patient sensitivity = 0.90 vs. 0.43; $\Delta F1 \approx 0.32$).

3.1.3 | Patient-Level Classification Through Window Aggregation

The “percentage of subjects correctly classified” (Figure 2) quantifies subject-level discrimination between symptom-expressing patients and healthy controls obtained by aggregating out-of-fold window predictions within each subject. Each 10-s window is classified independently, and a subject is labeled “symptom-positive” when at least half of his windows are predicted positive ($\geq 50\%$). Clinically, this parallels a physician reviewing multiple short epochs before issuing an overall judgment based on the preponderance of evidence. Subject-level performance was consistently high across most phenotypes and annotation sets, frequently reaching 1.00, indicating that aggregation mitigates sporadic window-level errors, particularly for intermittent phenotypes, while preserving conservative behavior in controls.

3.1.4 | Performance Under Standardized Clinical Conditions

To assess robustness across standardized examination contexts, we evaluated symptom-versus-control classification separately during rest, posture maintenance, and action. As summarized in Figure 2B, dystonia and tremor remained highly discriminable across conditions for several model families, with multiple pipelines achieving near-ceiling performance. Given the smaller condition-specific subsets, these results should be interpreted as supportive evidence of context robustness.

3.2 | Multi-Label Detection of Concurrent Hyperkinetic MDs

We next evaluated whether pose-derived kinematic features can support simultaneous, clinically interpretable detection of multiple hyperkinetic MDs, reflecting the reality that several abnormal movement patterns may co-occur within the same patient. Patient-level multi-label performance was estimated under three model-selection strategies, summarized in Table 4. Under the prespecified primary analysis using nested cross-validation with per-label model selection within training folds, the framework reached macro-AUPRC = 0.717 ± 0.030 , macro-AUROC = 0.767 ± 0.069 , Hamming accuracy = 0.764 ± 0.014 , and patient-label agreement of 153/200 (76.5%). When restricted to a single fixed pipeline selected by discrimination, the best-performing configuration (StandardScaler + MLP) achieved macro-AUPRC = 0.821 ± 0.019 and macro-AUROC = 0.830 ± 0.029 at the patient level (Figure 3); when restricted to a single fixed pipeline selected by Hamming accuracy (MinMaxScaler + SVM), it achieved 0.764 ± 0.041 . As an exploratory upper bound, post hoc selection of the best-performing pipeline for each phenotype (Table 5) yielded patient-label agreement of 172/200 (86.0%). The 9.5-percentage-point gap between this upper bound and the nested cross-validation estimate quantifies the model-selection contribution to apparent performance and supports our framing of these results as exploratory rather than confirmatory. Per-phenotype sensitivity and

specificity, with 95% Wilson score confidence intervals under both the post hoc per-label best and the nested cross-validation strategies, are reported in Tables S3a and S3b.

To contextualize these aggregate metrics, we examined the best per-label models (selected to minimize patient-level errors for each phenotype) and decomposed predictions into true negatives, true positives, false positives and false negatives (Table 5). This exploratory upper-bound analysis comprises 200 patient-label decisions (25 subjects $\times$ 8 labels); under per-label model selection performed post hoc, predictions agreed with clinician consensus labels in 172 decisions (86.0%). Importantly, this 86.0% figure is an upper-bound estimate subject to model-selection optimism, in contrast to the primary nested cross-validation estimate of 153/200 (76.5%) reported above (Table 4). Performance estimates for phenotypes with very few positive participants, notably ballismus (3 positive patients), stereotypies (7) and tics (3), carry wide confidence intervals and should be regarded as exploratory signals. Under nested cross-validation, ballismus sensitivity was 0.00 (95% Wilson CI 0.00–0.56) and tics sensitivity was 0.67 (95% CI 0.21–0.94); under the post hoc per-label best, the same labels reached 0.33 (95% CI 0.06–0.79) and 1.00 (95% CI 0.44–1.00), respectively (Tables S3a/S3b). Bootstrap 95% confidence intervals for macro-AUPRC and macro-AUROC across folds are provided in Table S4). The breadth of these intervals, spanning much of the [0,1] range, illustrates that with denominators of 3 to 7, even one displaced case substantially shifts the point estimate. In this cohort, tics showed apparent perfect agreement under post hoc per-label selection (TP = 3, TN = 22; no discordant decisions), and dystonia demonstrated high sensitivity with only a single missed case (TP = 20, FN = 1) alongside minimal over-calling (FP = 1). Myoclonus and chorea achieved high recall (no false negatives), with remaining discordances dominated by over-calling, an operating profile that, in practice, would be interpreted in clinical context and could be shifted by adopting stricter thresholds for phenotypes where false alarms are particularly costly. In contrast, tremor exhibited the largest discordance burden (FN = 4, FP = 4), consistent with borderline presentations and overlap with dystonia and non-specific movement patterns captured in short windows; athetosis similarly

TABLE 4 | Patient-level multi-label performance across model-selection strategies.

Strategy	Macro-AUPRC	Macro-AUROC	Hamming accuracy	Patient-label agreement
Best single pipeline by macro-AUPRC (StandardScaler + MLP): exploratory	0.821 ± 0.019	0.830 ± 0.029	0.731 ± 0.036	146/200 (73.0%)
Best single pipeline by Hamming accuracy (MinMaxScaler + SVM): exploratory	0.769 ± 0.049	0.749 ± 0.094	0.764 ± 0.041	153/200 (76.5%)
Nested cross-validation, per-label best within training folds: Primary	0.717 ± 0.030	0.767 ± 0.069	0.764 ± 0.014	153/200 (76.5%)
Post hoc per-label best (upper bound)—exploratory	—	—	—	172/200 (86.0%)

Note: Comparison of patient-level multi-label classification performance under four model-selection strategies (3-fold subject-level cross-validation, $n = 25$ participants, 8 phenotypes). The primary analysis (highlighted) is the prespecified nested cross-validation with per-label model selection performed inside the training folds. The other strategies are exploratory and are reported here only to make explicit the contribution of model-selection optimism to apparent performance: the 9.5-percentage-point gap between the post hoc per-label upper bound (172/200 = 86.0%) and the nested cross-validation estimate (153/200 = 76.5%) quantifies this optimism. Values are mean $\pm$ standard deviation across folds; per-phenotype sensitivity and specificity with 95% Wilson confidence intervals are reported in Tables S3a (post hoc) and S3b (nested CV).

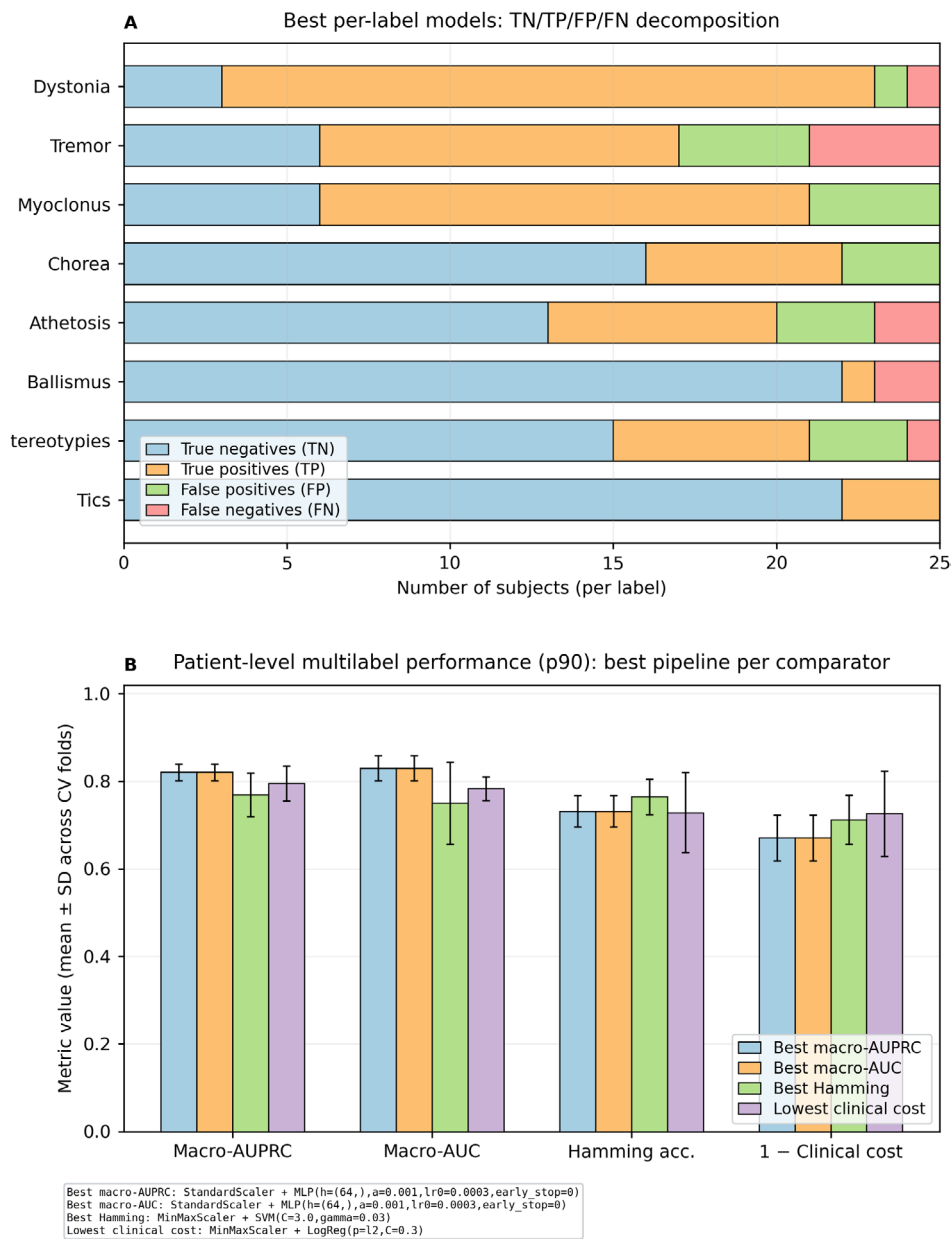


FIGURE 3 | Patient-level multi-label detection of HMDs using p90 aggregation of window probabilities. (A) Subject-level TN/TP/FP/FN counts for the best per-label pipelines (TN+TP+FP+FN per label), highlighting error types across phenotypes in the context of label prevalence. (B) Best overall pipelines under different selection criteria (macro-AUPRC, macro-AUC, Hamming accuracy, 1 – clinical cost), reported as mean $\pm$ SD across folds, illustrating trade-offs between discrimination and error-sensitive objectives under the same p90 aggregation and per-label thresholding.

showed balanced under- and over-calling (FN=2, FP=3), suggesting that subtle, low-frequency fluctuations remain difficult to consolidate into an unambiguous patient-level label. For rare labels, the decomposition was particularly informative. Ballismus, observed in only three positive patients, maintained perfect specificity (FP=0) but lower sensitivity (FN=2 under post hoc; FN=3 under nested CV), illustrating the expected behavior of conservative detection in the low-prevalence regime. Apparent strengths or weaknesses for rare phenotypes (notably ballismus, stereotypies, and tics) cannot be regarded as established phenotype signatures from this cohort alone and are additionally constrained by low inter-rater agreement at the window level ($\kappa=0.220, 0.000$ and 0.116 , respectively; Table S3). The most clinically relative challenging phenotypes reflected familiar sources of ambiguity in routine assessment and therefore

concentrated the residual discordances. The model prioritizes avoiding over-calling a high-impact phenotype at the expense of missed cases. Overall, patient-level multi-label inference achieved high agreement.

Collectively, these focused results suggest only partial separability of dystonia, chorea, and athetosis based on pose-derived kinematic summaries from short 10-s windows. Dystonia (prevalence 0.84) showed near-ceiling sensitivity with minimal false alarms, whereas chorea and athetosis remained more challenging, consistent with low prevalence (6/25 and 9/25 positive participants, respectively) and substantial phenotypic overlap in routine outpatient videos. The residual chorea–athetosis confusion is consistent with the known phenomenological overlap between these two phenotypes in routine clinical assessment

TABLE 5 | Best per-label patient-level classifiers and confusion-matrix breakdown (multi-label, p90 aggregation).

Model	Label	P	N	Prevalence	TP	FN	FP	TN	FNR	FPR	Errors	Errors per patient	Recall	Specificity
MinMaxScaler+SVM ($C=3.0$, $\text{gamma}=0.03$)	Dystonia	21	4	0.84	20	1	1	3	0.05	0.25	2	0.08	0.95	0.75
StandardScaler+SVM ($C=3.0$, $\text{gamma}=\text{scale}$)	Tremor	15	10	0.6	11	4	4	6	0.27	0.40	8	0.32	0.73	0.60
MinMaxScaler+KNN ($k=9$, $w=\text{uniform}$, $p=2$)	Myoclonus	15	10	0.6	15	0	4	6	0	0.40	4	0.16	1	0.60
MinMaxScaler+KNN ($k=9$, $w=\text{uniform}$, $p=2$)	Chorea	6	19	0.24	6	0	3	16	0	0.16	3	0.12	1	0.84
MinMaxScaler+KNN ($k=9$, $w=\text{uniform}$, $p=2$)	Athetosis	9	16	0.36	7	2	3	13	0.22	0.19	5	0.20	0.78	0.81
StandardScaler+SVM ($C=3.0$, $\text{gamma}=\text{scale}$)	Ballismus	3	22	0.12	1	2	0	22	0.67	0	2	0.08	0.33	1
StandardScaler+SVM ($C=10.0$, $\text{gamma}=0.02$)	Stereotypies	7	18	0.28	6	1	3	15	0.14	0.17	4	0.16	0.86	0.83
MinMaxScaler+SVM ($C=1.0$, $\text{gamma}=\text{scale}$)	Tics	3	22	0.12	3	0	0	22	0	0	0	0	1	1.00

Note: For each HMD phenotype, we report the pipeline with the fewest patient-level errors. We provide P/N counts and prevalence, patient-level TP/FN/FP/TN, FNR and FPR, total errors (FP+FN), errors per subject, and recall/specificity.

and with the moderate-to-fair inter-rater agreement at the window level ($\kappa=0.600$ for chorea, 0.259 for athetosis; Table S4,S5) sT2 and should not be interpreted as definitive discrimination. Larger, multicentre cohorts with adjudicated labels are needed before any clinical claim of chorea-versus-athetosis discrimination can be made. When applying the same multi-model approach to the full symptom set instead of the focused, three-label assessment, the error profile shifts in a phenotype-specific way. Chorea improves in the full panel of hyperkinetic MD labels (errors decrease from 6 $\rightarrow$ 3), suggesting that additional phenotypes and richer negative structure can offer better contrast and help isolate chorea-specific kinematic signatures. Athetosis appears broadly stable, indicating that its separability is less sensitive to whether training is restricted or expanded. In contrast, dystonia performs slightly better in the reduced panel (difference of one additional error).

3.3 | Clinically Interpretable Feature Importance at the Patient Level

To identify which kinematic descriptors most strongly influenced clinically meaningful classification decisions, we quantified patient-level permutation importance on the held-out outer test folds only, thereby preventing information leakage.

For interpretability, we summarized importance within clinically grounded kinematic families (baseline posture, sustained bias, excursions, variability, rhythmicity, and directionality) (Figure 4A) and across coarse anatomical regions (Figure 4B) (cranial, upper limb/proximal, lower limb/distal). Across phenotypes, importance concentrated in descriptors capturing baseline/sustained displacement and extreme excursions, whereas spectral and higher-order descriptors contributed more selectively (Figure 5). Here, “higher-order” features mainly comprised distributional and complexity measures (e.g., skewness/kurtosis, entropy-based measures including permutation entropy, and fractal-like complexity via Higuchi fractal dimension). These features generally showed secondary, context-dependent contributions, improving discrimination in selected cases rather than acting as the primary determinants of the patient-level decision.

Anatomically, influential descriptors frequently involved cranial and proximal upper-limb landmarks, consistent with their robust visibility in standard clinical videos and their sensitivity to axial/postural adjustments. Lower-limb descriptors contributed more variably: they were prominent for phenotypes with stereotyped limb pattern, and they may also become important in patients with generalized involvement (e.g., generalized dystonia), although their contribution is inherently dependent on task, camera framing, and the captured field of view. Despite shared global trends, each symptom exhibited a distinct importance profile aligned with clinical phenomenology. Dystonia was predominantly driven by baseline and sustained postural descriptors, consistent with a persistent abnormal postural set-points; excursion-type descriptors were secondary, compatible with superimposed phasic dystonic movements. The dominant topographic contribution arose from cranial/cervical and proximal upper-limb landmarks, reflecting clinically salient axial components, frequent involvement of axial and upper body

segments in adulthood dystonia and compensatory upper-body adjustments. Tremor relied primarily on baseline/sustained displacement and overall movement burden rather than on a narrow spectral signature. While rhythmicity-related descriptors (FFT peak frequency/amplitude) were present, decision-level attribution suggested that tremor detection in this dataset was largely mediated through posture-linked oscillatory load and sustained variability, which is clinically plausible in a heterogeneous tremor population (including dystonic tremor/tremulous dystonia), where frequency can vary across individuals, tasks, and recording conditions and where short windows can yield unstable single-peak spectral estimates. Myoclonus showed an event-like signature dominated by paroxysmal excursions (extreme minima/maxima and related excursion proxies), consistent with brief jerks producing disproportionate extremes in displacement time series; baseline descriptors contributed secondarily. Athetosis was distinguished by a comparatively stronger contribution from directionality-related descriptors (e.g., repeated reversals captured by derivative sign changes and trajectory evolution proxies), consistent with slow, writhing movement characterized by continuously evolving trajectories rather than periodic oscillation or isolated events. Stereotypies exhibited a predominantly segmental signature with comparatively greater limb (including lower-limb) involvement than most other phenotypes, consistent with repeated patterned actions that preserve a characteristic set-point while producing reproducible excursions. For rarer phenotypes, we observed plausible but less stable signatures. Chorea, ballismus, and tics tended to be driven by excursion- and variability-related descriptors, consistent with irregular, abrupt and/or large-amplitude deviations. Aggregating decision-level feature importance into clinically grounded kinematic families revealed structured overlaps between phenotypes that mirror bedside phenomenology and highlight plausible sources of diagnostic ambiguity. First, dystonia and tremor shared a strong reliance on baseline and sustained postural signals, consistent with the clinical observation that tremor often occurs on top of an abnormal postural set-point and that axial/proximal adjustments can dominate the video signature. In this context, tremor was differentiated primarily by a greater relative contribution of movement-burden descriptors (including segmental upper-limb contributors), whereas dystonia remained more strongly anchored in sustained postural bias. Second, we observed a partial overlap between myoclonus and tremor, driven by shared reliance on cranial/proximal descriptors capturing overall displacement burden. Clinically, this overlap is expected because both phenotypes can increase positional variability and intermittently produce excursions in standard video recordings; however, the relative weighting differed, with myoclonus exhibiting a more event-like profile dominated by extreme excursions (brief jerks) and tremor relying more on sustained oscillatory load (a more continuous movement pattern). Third, event-dominated phenotypes (tics, ballismus, and to some extent chorea) showed convergent importance patterns characterized by excursion- and variability-related evidence, consistent with abrupt, irregular and/or large-amplitude deviations, albeit with the above prevalence-related caveats. In contrast, athetosis emerged as the phenotype most enriched for directionality-related evidence, supporting its differentiation from both oscillatory and purely paroxysmal phenomena.

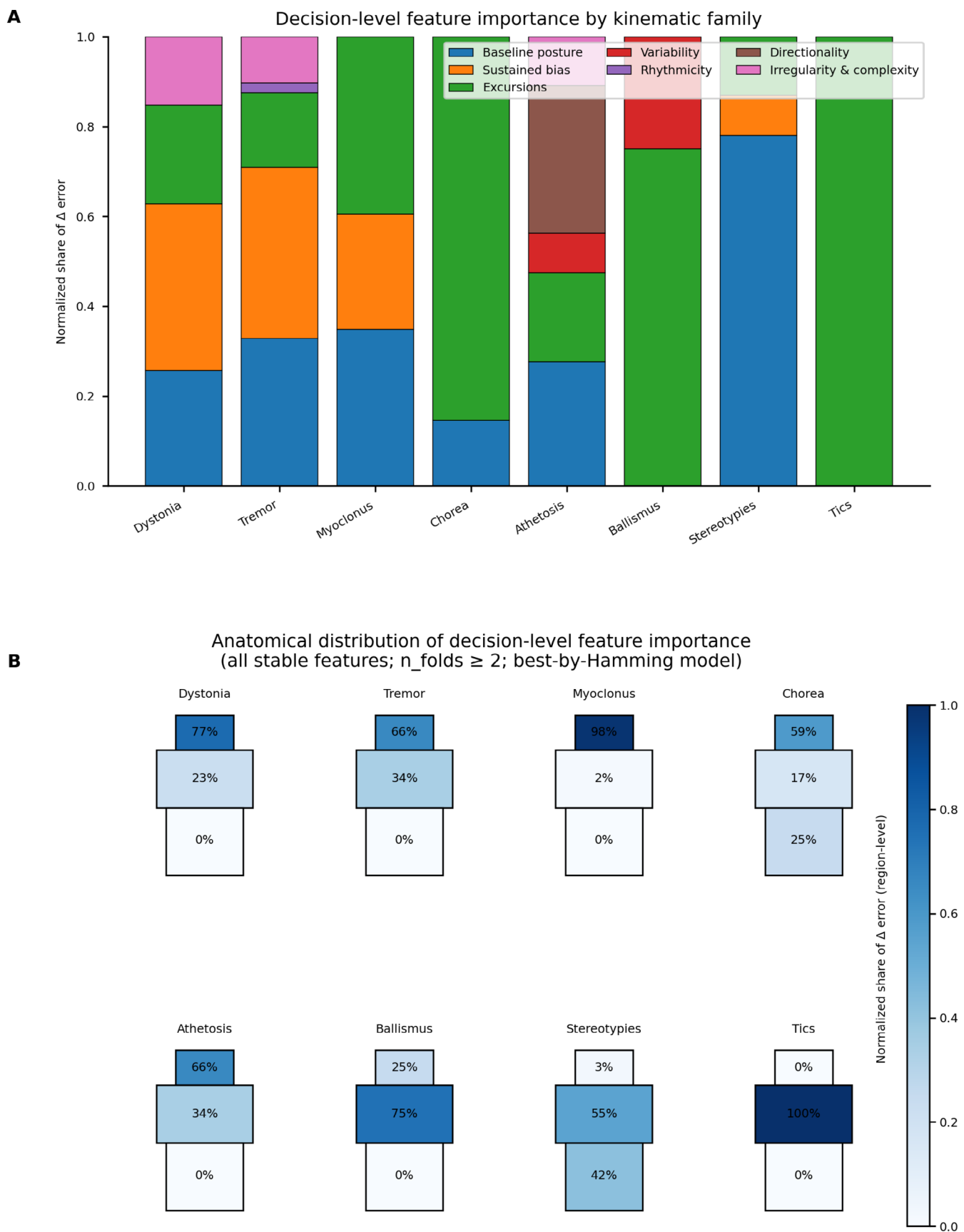


FIGURE 4 | Legend on next page.

FIGURE 4 | Patient-level decision-level feature importance. (A) Permutation importance by kinematic family for the best-by-Hamming model (MinMaxScaler+SVM; $C = 3.0$, $\gamma = 0.03$), computed on held-out outer folds ($n_{\text{folds}} \geq 2$). Importance is the increase in patient-level error ($FP + FN$)/ N after permuting each feature, preserving the full pipeline (window scores $\rightarrow$ p90 aggregation $\rightarrow$ label-specific thresholds). Bars show the normalized Δ error share per kinematic family within each symptom. Importance is dominated by postural set-point and excursion features, with more selective contributions from rhythmicity and irregularity/complexity. (B) The same attributions aggregated by landmark region (head/face, upper limb, lower limb) as normalized Δ error share per symptom, highlighting predominant cranial and proximal upper-limb evidence, with phenotype-dependent lower-limb contributions.

4 | Discussion

We developed a pose-based DL framework for automatic detection of concurrent hyperkinetic MDs from outpatient videos. Previous applications explored detection, classification [33], and quantification of various specific movement disorders, including Parkinson's disease [34], gait disorders [33], ataxia [35], and tics [19]. Several performance measures are typically reported, yet they are not always intuitive for clinicians [36]. We report ranking-based and threshold-based patient-level metrics because they address different clinical questions: whether the model correctly prioritizes patients with likely symptom expression, and whether it makes appropriately cautious final present/absent calls when used for patient-level phenotyping. In our window-based pre-analyses, the negative class was deliberately defined conservatively as healthy control windows rather than symptom-absent windows from patients, so performance should be interpreted as discrimination between clinically labeled symptom expression and normative movement patterns, not as “symptom absence within patients” detection. Within this screening-oriented definition, patient-level aggregation yielded consistently high control specificity, indicating that repeated short observations rarely led to symptomatic labeling in healthy controls. This behavior is clinically desirable and suggests that pose-derived features capture stable normative movement signatures that are not misclassified as hyperkinetic activity when evidence is aggregated.

Window-level discrimination varied by phenotype, mirroring differences in phenomenology and the ease of defining symptom boundaries on routine videos. Dystonia, chorea, tremor, and tics showed robust window-level reliability (positive-class $F1$ values in the ~ 0.8 – 0.9 range across raters) and strong subject identification after aggregation, consistent with relatively distinctive kinematic signatures: sustained abnormal posturing for dystonia, irregular purposeless excursions for chorea, oscillatory structure for tremor, and brief, variable, or stereotyped bursts for tics. In contrast, athetosis, stereotypies, myoclonus, and ballismus were more challenging at the 10-s scale, consistent with (i) slower, continuous, and often mixed-pattern movements (athetosis), (ii) context-dependent repetitive behaviors (stereotypies), and (iii) highly intermittent, short-duration events that can be difficult to time-stamp consistently (myoclonus and ballismus). This pattern is reinforced by rater-dependence: tremor was highly concordant between annotation sets ($\Delta F1 \approx 0.01$), while myoclonus showed marked annotation-set dependence ($\Delta F1 \approx 0.32$), underscoring the clinical difficulty of time-stamping brief shock-like events and the sensitivity of learned signatures to labeling policy. Importantly, subject-level identification generally exceeded window-level reliability, supporting

the clinical utility of aggregating repeated short observations: majority-vote consolidation reduces the influence of sporadic window-level errors and recapitulates the clinical process of forming a judgment after reviewing multiple epochs rather than relying on a single short clip.

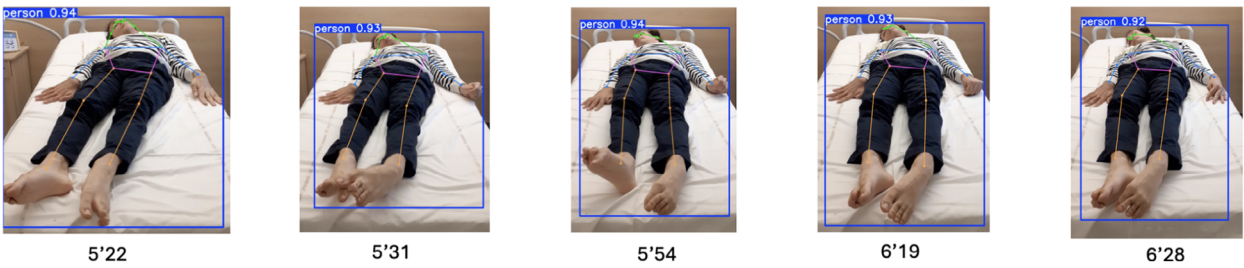
Under standardized examination conditions (rest, posture maintenance, action), symptom-versus-control separation for dystonia and tremor remained strong across several model families, with multiple pipelines reaching near-ceiling performance across conditions.

Because clinical assessment frequently involves concurrent hyperkinetic phenotypes, we next asked whether pose-derived kinematic features can support simultaneous and clinically interpretable multi-label inference at the patient level. Discrimination was strong when probabilities were interpreted as ranked evidence rather than hard calls, indicating robust separability of symptomatic and asymptomatic subjects across cross-validation folds. Importantly, optimizing for different clinical priorities produced the expected trade-offs: maximizing overall correctness across all patient–phenotype decisions (best Hamming accuracy) yielded 0.764 ± 0.041 , whereas choosing thresholds to minimize clinically meaningful misclassification led to more cautious symptom assignment, with fewer positive calls unless the evidence was stronger. This pattern is consistent with how thresholds may be adapted to different clinical contexts, from initial screening to more confirmation-oriented use. This “phenotype-specific operating landscape” is a critical clinical message: a probabilistic backbone can be tuned to match the intended context of use rather than forcing a single operating point across all phenotypes.

To contextualize aggregate metrics, we examined best per-label models (selected to minimize patient-level errors) and decomposed predictions into true negatives, true positives, false positives and false negatives. Across 200 patient–label decisions (25 subjects $\times$ 8 labels), per-label selection agreed with clinician labels in 172/200 decisions (86.0%). Models were more in agreement for phenotypes with distinctive and/or consistently annotated patterns: tics were perfectly identified in this cohort, and dystonia showed high sensitivity with only a single missed case and minimal overcalling. Myoclonus and chorea also showed high recall in this cohort (no false negatives), with the remaining errors mainly consisting of false positives. In practice, such false-positive calls would often be resolved by clinical assessment and may be reduced by applying stricter thresholds when a more cautious classification is preferred. Tremor carried relative disagreement, consistent with borderline presentations, task dependence, and overlap with non-specific movement patterns

A

Video Frames with Pose Estimation Overlay



B

Representative temporal signatures of keypoint-derived kinematics across phenotypes

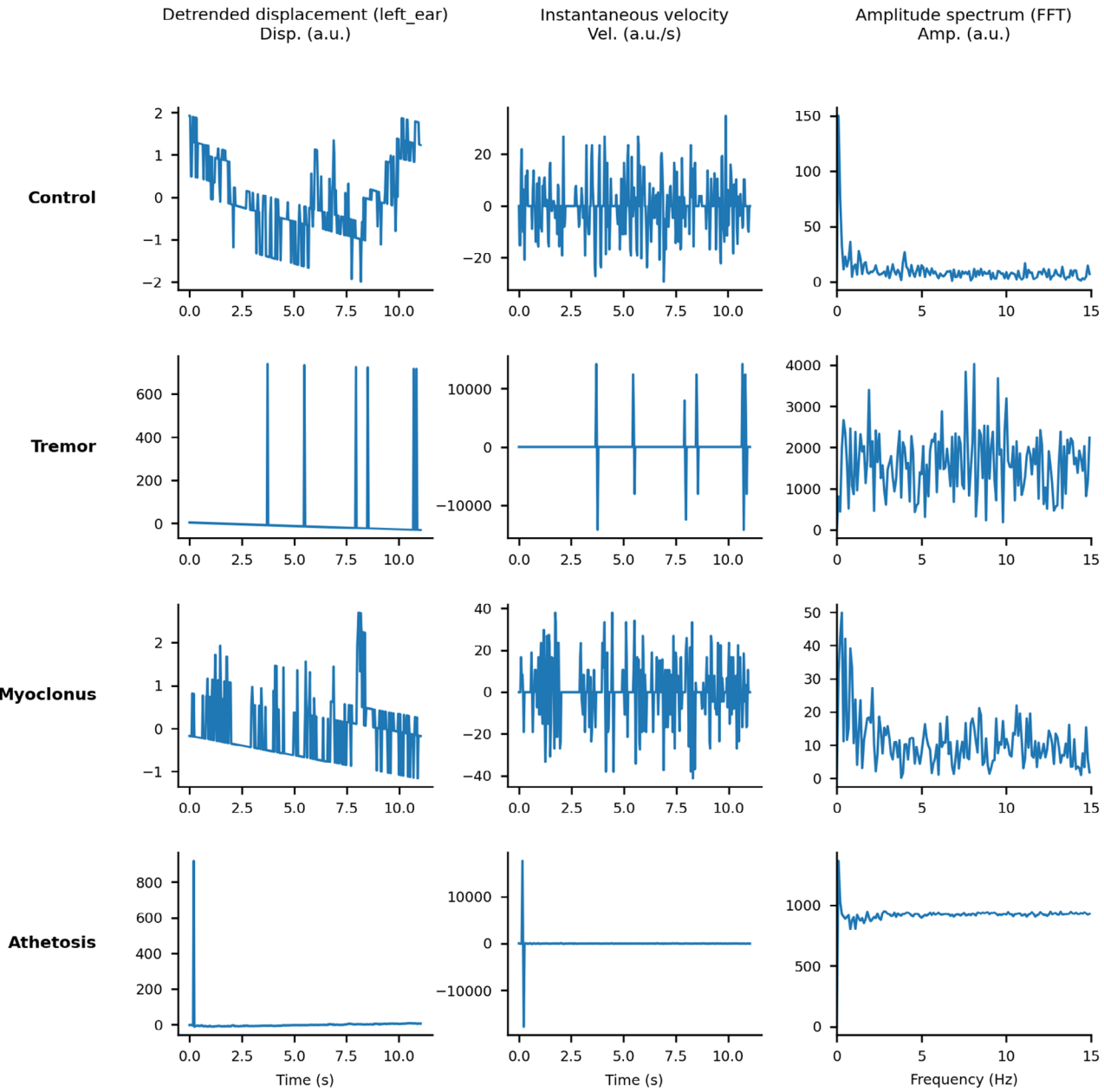


FIGURE 5 | Legend on next page.

FIGURE 5 | Pose estimation and kinematic signatures across phenotypes. (A) Example video frames with YOLOv8 pose-estimation overlay. (B) Representative keypoint-derived kinematic signatures extracted from the selected landmark (here, the left ear) for one control and three phenotypes. Three complementary signals are shown for each example: Detrended displacement (Disp., arbitrary units [a.u.]), defined as the landmark displacement relative to its median position after linear detrending; instantaneous velocity (Vel., a.u./s), computed as the first derivative of the detrended displacement; and the amplitude spectrum (Amp., a.u.) obtained by fast Fourier transform (FFT). Controls show low-amplitude fluctuations without a dominant spectral peak. Tremor shows sustained oscillatory burden with repeated velocity alternations and enhanced spectral power. Myoclonus shows abrupt excursions and sharp velocity spikes, consistent with brief, jerky movements. Athetosis shows slower evolving fluctuations with frequent directional changes, consistent with trajectory-evolution features.

captured in short windows; athetosis similarly exhibited few false negatives and false positives, suggesting that subtle low-frequency fluctuations remain difficult to consolidate into an unambiguous patient-level call from brief windows alone. For rare labels, the decomposition was particularly informative: ballismus, observed in only three positive subjects, maintained perfect specificity but relative lower sensitivity, illustrating cautious detection in the low-prevalence regime, avoiding overcalling a high-impact diagnosis at the expense of missed cases.

A core methodological and clinical challenge in this domain is that no absolute ground truth exists for hyperkinetic MDs classification. As highlighted in the literature, rated tasks are vulnerable to uncertainty and inter-rater variability [37]. Consequently, some apparent model “errors” may reflect ambiguous or borderline phenomenology rather than purely algorithmic failure; in practice, a discordant prediction can sometimes flag a segment where the reference label itself is contestable, particularly for brief, intermittent events and for phenotypes with overlapping kinematic signatures. Majority voting can provide a pragmatic pseudo-ground truth at the patient level, but it has caveats, especially when disagreement reflects true ambiguity rather than noise. Models trained on assessments from multiple assessors could potentially outperform any individual assessors in specific settings [33], but realizing this advantage will require explicit uncertainty-aware labeling (probabilistic labels, adjudication, or structured disagreement modeling) rather than excluding ambiguous cases. This issue becomes even more critical when moving from classification to quantification, where disagreement typically increases and clinical consequences of misestimation may be higher. Most current DL-based quantification work focuses on individual Parkinson’s disease symptoms such as levodopa-induced dyskinesia (LID) [38], tremor, rising from a chair [39], and in the hyperkinetic domain has addressed eye tics [19], action myoclonus [40] and focal dystonia [41]; quantification of multiple hyperkinetic MDs remains a key next step for the field and is the focus of our ongoing work.

Across phenotypes, decision-level importance concentrated in kinematic families capturing baseline posture and sustained displacement as well as extreme excursions and variability, whereas rhythmicity and higher-order irregularity/complexity descriptors contributed more selectively. This pattern supports a clinically intuitive interpretation: the models largely rely on descriptors that correspond to what clinicians observe at the bedside, movement burden, dispersion around a postural set-point, and episodic departures, while using spectral and complexity cues to refine decisions when physiology predicts them. Importances also revealed meaningful phenotype-specific profiles: dystonia and tremor were strongly shaped by baseline and

sustained postural descriptors, consistent with tonic deviation and oscillatory load on top of a postural set-point; chorea and tics were dominated by excursion-type evidence, consistent with irregular, abrupt deviations; myoclonus showed an event-like profile driven by excursion and baseline burden consistent with brief jerks that disproportionately influence extremes; and athetosis showed relatively greater contributions from directionality-related evidence, consistent with slow writhing trajectories characterized by continuous evolution and repeated reversals rather than periodic oscillation or isolated paroxysms. At the anatomical level, influential evidence frequently arose from cranial and proximal upper-limb landmarks, which are robustly visible in routine clinic videos and sensitive to axial and upper-body adjustments; lower-limb contributions were more phenotype-dependent, becoming prominent for stereotypies and for phenotypes where limb patterning is salient within the field of view.

Several considerations should be kept in mind when interpreting these results. First, the cohort is modest: 21 patients and 4 healthy controls from a single centre, with imbalance between phenotypes ($n = 21, 15, 15, 9, 7, 6, 3, 3$ positive participants for dystonia, tremor, myoclonus, athetosis, stereotypies, chorea, ballismus and tics, respectively). Because all 10-s windows from a given participant share a common pathophysiology, examination context and label vector, the effective independent sample size for every analysis is the number of participants, not the number of windows. For phenotypes with single-digit numbers of positive participants (notably ballismus, stereotypies and tics), per-label estimates carry wide confidence intervals (e.g., 95% Wilson CI for ballismus sensitivity = 0.00–0.56 under nested cross-validation; Table S3b) and are best interpreted as exploratory signals. Although subject-level cross-validation prevents window-level leakage, within-patient windows share physiology, anatomy, examination context and label structure, and patient-level aggregation reduces noise without changing this constraint; the reported performance therefore likely overestimates what would be observed in an external, multicentre prospective setting. Second, data originate from a single centre. Third, the fixed 10-s window is a deliberate compromise that enables stable statistical and spectral descriptors but can dilute very transient events and may broaden narrowband signatures. Fourth, within each 10-s window the trajectory is summarized by interpretable, hand-crafted descriptors rather than modeled as a full temporal sequence. We chose this design deliberately: it is interpretable, parameter-efficient and well suited to the present small-cohort regime, and it captures the principal kinematic properties clinicians assess at the bedside. Because hyperkinetic movement disorders also involve onset–offset behavior, intermittency and rhythmicity, this window-summary representation

may underrepresent very brief or rapidly modulated patterns, which end-to-end temporal models could capture in future work. Fifth, inter-rater differences remain a central source of label noise. Inter-rater agreement at the window level, computed between the two independent raters prior to consensus, was substantial for dystonia (Cohen's $\kappa=0.762$) and tremor ($\kappa=0.724$), and lower for chorea ($\kappa=0.600$), myoclonus ($\kappa=0.366$), athetosis ($\kappa=0.259$) and the rare phenotypes ballismus, tics and stereotypies ($\kappa=0.220$, 0.116 and 0.000 , respectively) (Table S3). For the rarest phenotypes, extreme prevalence imbalance depresses κ even when raw agreement is high, and the two independent raters differed most in how often they annotated stereotypies before consensus, a reminder that shared operational criteria are still maturing for these phenomenologies and the reason we adopted a consensus label as the operational ground truth. Because uncertain or contested windows were excluded before model fitting, the present task is cleaner than routine clinical use, so reported metrics are likely optimistic relative to everyday clinical ambiguity; future work adopting uncertainty-aware labeling, probabilistic targets, adjudication panels, or structured disagreement modeling, would better reflect the absence of an absolute ground truth. Sixth, our representation is 2D and is based on 17 body keypoints, omitting axial rotations, depth motion, fine distal hand movements, lip and tongue movements, and subtle facial expressions, all of which carry phenotype-specific clinical information. Phenotypes whose clinically meaningful signal depends most on these finer or out-of-plane components, axial or three-dimensional displacement (e.g., torsion dystonia), fine distal kinematics (e.g., low-amplitude action tremor, subtle distal myoclonus, choreiform finger movements), or facial activity (e.g., facial tics), would therefore benefit most from future extensions to 3D pose and to dedicated hand and face tracking; the lower nested-CV performance observed for ballismus, myoclonus and stereotypies in our cohort is consistent with this expectation. Seventh, broader and more diverse control cohorts are needed to strengthen specificity estimates and calibrate thresholds under wider real-world variability. Eighth, although subject-level fold assignment prevents window-level leakage, the breadth of the model panel and the per-label model selection introduce a residual risk of selection optimism that cannot be eliminated in a single-centre cohort of this size; configurations selected as "best" should be regarded as upper-end estimates conditional on this cohort. Finally, the pipeline is more accurately described as semi-automated and annotation-guided rather than fully end-to-end automatic, because the temporal definition of the 10-s windows and the per-window supervision labels are provided by expert clinicians. Establishing clinical utility was beyond the scope of this proof-of-concept study and defines our roadmap for subsequent work: external and multicentre validation, prospective evaluation, test-retest reliability assessment, and comparison against independent expert panels, followed by studies assessing whether the pipeline improves diagnosis, monitoring or trial-eligibility assessment. We therefore present these findings as a foundation for those next steps rather than as evidence of clinical readiness.

We identify five concrete methodological priorities for future work. (i) External, multicentre, prospective validation with diverse cameras, distances, lighting conditions, age ranges (including pediatric and genetically defined cohorts) and adjudicated

labels, to quantify domain shift and recalibrate thresholds. (ii) Window-length ablation studies (e.g., 5s, 10s, 20s, sliding overlap) and direct comparison of alternative temporal aggregation strategies (mean, median, percentiles, attention pooling) to identify the optimal trade-off between event capture and statistical stability. (iii) Extension to 3D pose, dedicated hand and face trackers, and multimodal inputs (synchronized audio for vocal tics, accelerometry for tremor) to address phenotypes under-served by 2D body keypoints. (iv) True temporal sequence modeling (temporal convolutional, recurrent or transformer architectures) operating on per-frame keypoint trajectories and on cross-window dependencies, to capture onset-offset behavior, intermittency and amplitude modulation, contingent on substantially larger and multicentre cohorts. (v) Uncertainty-aware labeling and quantification (probabilistic targets, adjudication panels, structured disagreement modeling), and downstream test-retest reliability, comparison against independent expert panels, and prospective evaluation of clinical added value. In addition, symptom-specific modeling, one model per phenotype rather than a single monolithic model, appears aligned with both feature signatures and clinical reasoning, and may better accommodate phenotype-specific prevalence, ambiguity, and risk profiles, most naturally within hierarchical pipelines that combine the present generic-descriptor stage with phenotype-dedicated downstream modules incorporating task-specific clinical features (e.g., spatiotemporal gait parameters, tremor-band power and centre frequency, jerk-spike statistics for myoclonus, segment-level torsion for dystonia, finger-tap kinematics for action chorea). Workflow integration and governance will also be essential, including privacy-preserving on-device processing, real-time feedback during structured exams, and human-in-the-loop review with versioned model cards and bias audits.

5 | Conclusions

In summary, this single-centre, proof-of-concept study suggests that pose-based hybrid analysis of routine clinic video may capture clinically interpretable signatures of co-occurring hyperkinetic movement disorders. The modest cohort size, the imbalanced phenotype prevalence, the single-centre design, and the absence of external or prospective validation mean that conclusions about clinical utility, generalizability, and readiness for deployment must await further validation. Translation to clinical practice will require external multicentre validation, test-retest assessment, prospective evaluation against independent expert panels, and demonstration of added value for diagnosis, monitoring, or trial eligibility.

Author Contributions

Conceptualization: L.C., D.D., G.A.H., G.-M.H., J.B., X.V. Methodology: L.C., M.B., E.M.M., X.V. Investigation: L.C., D.D., G.A.H., J.D.O.E., N.D., M.C.J., C.A.H., T.W., G.-M.H., S.He, M.B., X.V. Assessment: L.C., D.D., G.A.H., J.D.O.E., N.D., M.C.J., C.A.H., T.W., S.Hu, M.D., Z.S., M.M.U.R., S.He. Data Collection: L.C., D.D., G.A.H., S.Hu, M.D. Data Curation: L.C., M.B., X.V. Validation: L.C., X.V. Writing – Review and Editing: L.C., D.D., G.A.H., J.D.O.E., N.D., M.C.J., C.A.H., T.W., G.-M.H., S.Hu, M.D., Z.S., M.M.U.R., S.He, M.B., E.M.M., J.B., X.V. Supervision: L.C., X.V.

Acknowledgments

The study was supported by a research grant from Dystonia Medical Research Foundation Canada (Registration Number 126616598 RR0001). Open access publishing facilitated by Universite de Lausanne, as part of the Wiley - Universite de Lausanne agreement via the Consortium Of Swiss Academic Libraries.

Funding

This work was supported by Dystonia Medical Research Foundation, Canada (Grant 126616598 RR0001).

Ethics Statement

This study was conducted in accordance with applicable regulations for research involving human participants and with respect for participants' privacy rights. Regulatory approvals were obtained (CESSRESS 22075132 Bis; CNIL 2238428). All participants provided written informed consent for video recording and research use. Data were handled under approved governance procedures, and analyses were performed on de-identified keypoint time series and derived kinematic features.

Conflicts of Interest

D.D. received honoraria for expert opinion from TEVA. J.B. is a shareholder of ONWARD Medical B.V., a company developing products for stimulation of the spinal cord, not related to this research. All other authors declare no competing interests.

Data Availability Statement

The online version contains [Supporting Information](https://github.com/xaviervasques/CODY) available at <https://github.com/xaviervasques/CODY> (data and code).

References

1. S. A. O'Shea and L. C. Shih, "Global Epidemiology of Movement Disorders: Rare or Underdiagnosed?," *Seminars in Neurology* 43, no. 1 (2023): 4–16.
2. B. Pérez-Dueñas, K. Gorman, A. Marcé-Grau, et al., "The Genetic Landscape of Complex Childhood-Onset Hyperkinetic Movement Disorders," *Movement Disorders* 37, no. 11 (2022): 2197–2209.
3. R. Brandsma, M. E. Van Egmond, M. A. J. Tijssen, and the Groningen Movement Disorder Expertise Centre, "Diagnostic Approach to Paediatric Movement Disorders: A Clinical Practice Guide," *Developmental Medicine and Child Neurology* 63, no. 3 (2021): 252–258.
4. A. Di Fonzo, E. Monfrini, and R. Erro, "Genetics of Movement Disorders and the Practicing Clinician: Who and What to Test for?," *Current Neurology and Neuroscience Reports* 18, no. 7 (2018), <https://doi.org/10.1007/s11910-018-0847-1>.
5. L. M. Lange, P. Gonzalez-Latapi, R. Rajalingam, et al., "Nomenclature of Genetic Movement Disorders: Recommendations of the International Parkinson and Movement Disorder Society Task Force—An Update," *Movement Disorders* 37, no. 5 (2022): 905–935.
6. R. J. Elble and W. Ondo, "Tremor Rating Scales and Laboratory Tools for Assessing Tremor," *Journal of the Neurological Sciences* 435 (2022): 120202.
7. A. Tarakad, "Clinical Rating Scales and Quantitative Assessments of Movement Disorders," *Neurologic Clinics* 38, no. 2 (2020): 231–254.
8. W. Tang, P. M. A. Van Ooijen, D. A. Sival, and N. M. Maurits, "Automatic Two-Dimensional & Three-Dimensional Video Analysis With Deep Learning for Movement Disorders: A Systematic Review," *Artificial Intelligence in Medicine* 156 (2024): 102952.
9. A. Darling, C. Tello, M. J. Martí, et al., "Clinical Rating Scale for Panthothenate Kinase-Associated Neurodegeneration: A Pilot Study," *Movement Disorders* 32, no. 11 (2017): 1620–1630.
10. J. Domínguez-Carral, W. G. Ludlam, M. Junyent Segarra, et al., "Severity of *GNAO1*-Related Disorder Correlates With Changes in G-Protein Function," *Annals of Neurology* 94, no. 5 (2023): 987–1004.
11. H. Haberfehlner, S. S. Van De Ven, S. A. Van Der Burg, et al., "Towards Automated Video-Based Assessment of Dystonia in Dyskinetic Cerebral Palsy: A Novel Approach Using Markerless Motion Tracking and Machine Learning," *Frontiers in Robotics and AI* 10 (2023), <https://doi.org/10.3389/frobt.2023.1108114/full>.
12. X. Vasques, H. Paik, and L. Cif, "Application of Quantum Machine Learning Using Quantum Kernel Algorithms on Multiclass Neuron M-Type Classification," *Scientific Reports* 13, no. 1 (2023): 11541.
13. M. U. Friedrich, S. Relton, D. Wong, and J. Alty, "Computer Vision in Clinical Neurology: A Review," *JAMA Neurology* 82, no. 4 (2025): 407.
14. R. Martínez-García-Peña, L. H. Koens, G. Azzopardi, and M. A. J. Tijssen, "Video-Based Data-Driven Models for Diagnosing Movement Disorders: Review and Future Directions," *Movement Disorders* 40, no. 10 (2025): 2046–2066.
15. W. Gong, X. Zhang, J. González, et al., "Human Pose Estimation From Monocular Images: A Comprehensive Survey," *Sensors* 16, no. 12 (2016): 1966.
16. A. L. Silva De Lima, L. J. W. Evers, T. Hahn, et al., "Freezing of Gait and Fall Detection in Parkinson's Disease Using Wearable Sensors: A Systematic Review," *Journal of Neurology* 264, no. 8 (2017): 1642–1654.
17. C. Bandt and B. Pompe, "Permutation Entropy: A Natural Complexity Measure for Time Series," *Physical Review Letters* 88, no. 17 (2002): 174102.
18. M. Delrobaei, S. Tran, G. Gilmore, K. McIsaac, and M. Jog, "Characterization of Multi-Joint Upper Limb Movements in a Single Task to Assess Bradykinesia," *Journal of the Neurological Sciences* 368 (2016): 337–342.
19. C. Conelea, H. Liang, M. DuBois, et al., "Automated Quantification of Eye Tics Using Computer Vision and Deep Learning Techniques," *Movement Disorders* 39, no. 1 (2024): 183–191.
20. M. M. Ali and S. I. Mohamed, "A Pose Estimation for Motion Tracking of Infants Cerebral Palsy," *Multimedia Tools and Applications* 84, no. 10 (2024): 8261–8286.
21. M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Vision-Based Assessment of Parkinsonism and Levodopa-Induced Dyskinesia With Pose Estimation," *Journal of Neuroengineering and Rehabilitation* 15, no. 1 (2018): 97.
22. L. He, F. He, Y. Li, X. Xiong, and J. Zhang, "A Robust Movement Quantification Algorithm of Hyperactivity Detection for ADHD Children Based on 3D Depth Images," *IEEE Transactions on Image Processing* 31 (2022): 5025–5037.
23. A. M. M. Van Der Stouwe, I. Tuitert, I. Giotis, et al., "Next Move in Movement Disorders (NEMO): Developing a Computer-Aided Classification Tool for Hyperkinetic Movement Disorders," *BMJ Open* 11, no. 10 (2021): e055068.
24. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA (IEEE, 2016), 779–788, accessed July 28, 2025, <http://ieeexplore.ieee.org/document/7780460/>.
25. G. Bradski, "The OpenCV Library," *Dr. Dobbs Journal: Software Tools* 25, no. 11 (2000): 120–123.
26. Z. Cao, G. Hidalgo, T. Simon, et al., *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields* (2018), accessed July 30, 2025, <https://arxiv.org/abs/1812.08008>.

27. H. H. Pham, L. Khoudour, A. Crouzil, et al., *Video-Based Human Action Recognition Using Deep Learning: A Review* (2022), accessed July 30, 2025, <https://arxiv.org/abs/2208.03775>.
28. G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLOv8* (2023), Retrieved from, <https://github.com/ultralytics/ultralytics>.
29. A. Mathis, P. Mamidanna, K. M. Cury, et al., "DeepLabCut: Markerless Pose Estimation of User-Defined Body Parts With Deep Learning," *Nature Neuroscience* 21, no. 9 (2018): 1281–1289.
30. C. R. Harris, K. J. Millman, S. J. Van Der Walt, et al., "Array Programming With NumPy," *Nature* 585, no. 7825 (2020): 357–362.
31. P. Virtanen, R. Gommers, T. E. Oliphant, et al., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods* 17, no. 3 (2020): 261–272.
32. W. McKinney, *Data Structures for Statistical Computing in Python* (2010), 56–61, <https://doi.org/10.25080/Majora-92bf1922-00a>.
33. S. Rupprechter, G. Morinan, Y. Peng, et al., "A Clinically Interpretable Computer-Vision Based Method for Quantifying Gait in Parkinson's Disease," *Sensors* 21, no. 16 (2021): 5437.
34. M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Automated Vision-Based Analysis of Levodopa-Induced Dyskinesia With Deep Learning," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Seogwipo (IEEE, 2017), 3377–3380, accessed September 12, 2025, <https://ieeexplore.ieee.org/document/8037580/>.
35. A. S. Nunes, N. Kozhemiako, C. D. Stephen, J. D. Schmahmann, S. Khan, and A. S. Gupta, "Automatic Classification and Severity Estimation of Ataxia From Finger Tapping Videos," *Frontiers in Neurology* 12 (2022): 795258.
36. S. A. Hicks, I. Strümke, V. Thambawita, et al., "On Evaluation Metrics for Medical Applications of Artificial Intelligence," *Scientific Reports* 12, no. 1 (2022): 5979.
37. M. Lu, K. Poston, A. Pfefferbaum, et al., "Vision-Based Estimation of MDS-UPDRS Gait Scores for Assessing Parkinson's Disease Motor Severity," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*, ed. A. L. Martel, P. Abolmaesumi, D. Stoyanov, et al. (Springer International Publishing, 2020), 637–647, https://doi.org/10.1007/978-3-030-59716-0_61.
38. M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Automated Assessment of Levodopa-Induced Dyskinesia: Evaluating the Responsiveness of Video-Based Features," *Parkinsonism & Related Disorders* 53 (2018): 42–45.
39. G. Morinan, Y. Peng, S. Rupprechter, et al., "Computer-Vision Based Method for Quantifying Rising From Chair in Parkinson's Disease Patients," *Intelligence-Based Medicine* 6 (2022): 100046.
40. J. Hyppönen, A. Hakala, K. Annala, et al., "Automatic Assessment of the Myoclonus Severity From Videos Recorded According to Standardized Unified Myoclonus Rating Scale Protocol and Using Human Pose and Body Movement Analysis," *Seizure* 76 (2020): 72–78.
41. Z. Zhang, E. Cisneros, H. Y. Lee, et al., "Hold That Pose: Capturing Cervical Dystonia's Head Deviation Severity From Video," *Annals of Clinical Translational Neurology* 9, no. 5 (2022): 684–694.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Supporting Information S1:** CODY-SAMP video recording protocol and instructions for raters. **Supporting Information S2:** Considerations related to phenomenological distinctions. **Supporting Information S3:** Computational methods. **Table S1:** Per-phenotype window inventory and per-fold class balance. **Table S2:** Inter-rater agreement (Cohen's κ) at the window level prior to consensus. **Table S3a:** Patient-level sensitivity and specificity under post hoc per-label best (exploratory upper bound). **Table S3b:**

Patient-level sensitivity and specificity under nested cross-validation (PRIMARY analysis). **Table S4:** Discrimination metrics with 95% bootstrap (BCa) confidence intervals across folds.