

Video-Based Activity Recognition for Automated Motor Assessment of Parkinson's Disease

Grzegorz Sarapata , Yuriy Dushin , Gareth Morinan , Joshua Ong , Sanjay Budhdeo , Bernhard Kainz , Senior Member, IEEE, and Jonathan O'Keeffe 

Abstract—Over the last decade, video-enabled mobile devices have become ubiquitous, while advances in markerless pose estimation allow an individual's body position to be tracked accurately and efficiently across the frames of a video. Previous work by this and other groups has shown that pose-extracted kinematic features can be used to reliably measure motor impairment in Parkinson's disease (PD). This presents the prospect of developing an asynchronous and scalable, video-based assessment of motor dysfunction. Crucial to this endeavour is the ability to automatically recognise the class of an action being performed, without which manual labelling is required. Representing the evolution of body joint locations as a spatio-temporal graph, we implement a deep-learning model for video and frame-level classification of activities performed according to part 3 of the Movement Disorder Society Unified PD Rating Scale (MDS-UPDRS). We train and validate this system using a dataset of $n = 7310$ video clips, recorded at 5 independent sites. This approach reaches human-level performance in detecting and classifying periods of activity within monocular video clips. Our framework could support clinical workflows and patient care at scale through applications such as quality monitoring of clinical data collection, automated labelling of video streams, or a module within a remote self-assessment system.

Index Terms—Activity recognition, computer vision, graph neural networks, Parkinson's disease, telemedicine.

Manuscript received 23 November 2022; revised 31 March 2023 and 10 July 2023; accepted 17 July 2023. Date of publication 25 July 2023; date of current version 5 October 2023. This work was supported by Machine Medicine Technologies. (*Corresponding authors:* Grzegorz Sarapata; Jonathan O'Keeffe.)

This work involved human subjects or animals in its research. Ethical approval for the capture of the Kelvin data was obtained from the National Hospital for Neurology and Neurosurgery Research Ethics Committee Application No. 19/YH/0421.

Grzegorz Sarapata, Yuriy Dushin, Gareth Morinan, Joshua Ong, and Jonathan O'Keeffe are with the Machine Medicine Technologies, SE16 4DG London, U.K. (e-mail: greg@machinemedicine.com; yuriy@machinemedicine.com; gareth@machinemedicine.com; joshua@machinemedicine.com; jonathan@machinemedicine.com).

Sanjay Budhdeo is with the Department of Clinical and Movement Neurosciences, Institute of Neurology, University College London, Queen Square, WC1N 3BG London, U.K. (e-mail: s.budhdeo@ucl.ac.uk).

Bernhard Kainz is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K., and also with the FAU Erlangen-Nürnberg, DE 91054 Erlangen, Germany (e-mail: b.kainz@imperial.ac.uk).

Digital Object Identifier 10.1109/JBHI.2023.3298530

I. INTRODUCTION

A. Background

HUMAN Activity Recognition (HAR) is the process of detecting and classifying human actions through sensor data [1]. The widespread availability of easy-to-use video recording devices increases the appeal of computer-vision-based HAR, operating on a sequence of captured images. An automated activity recognition framework would facilitate many practical applications such as behavioural analysis [2], automatic security surveillance [3], and human-computer interactive systems [4].

In the context of healthcare, HAR has found use as an assistive tool for patient monitoring, remote medicine services or eldercare [5], [6]. Mobile health technology has been shown to effectively complement traditional public health services, and the rapid growth of its popularity further motivates the development of machine-driven HAR [7], [8]. However, HAR remains challenging, due in part to the complexity of scenes and human movements [9].

B. Motivation

Parkinson's disease (PD) severity is often assessed using the Movement Disorder Society Unified PD Rating Scale (MDS-UPDRS) [13]. Part 3 of this four-part assessment comprises 14 motor activities performed by the patient and rated by an examining clinician. The in-person character and duration of the assessment impose constraints on the frequency of patient monitoring efforts, as patients typically travel to each assessment site and clinician time is finite. These limitations may be offset by a remote, asynchronous and self-administered system capable of collecting and analysing data with a high frequency and at scale. To automatically guide the patient through an assessment and validate that each motor task has been completed, an assessment-focused activity recognition framework needs to be developed. In addition to enhancing patient care, such a system could be used to automatically annotate large quantities of motor assessment data in clinical trials, greatly increasing the statistical power of future PD research.

C. Previous Work

In recent years deep learning (DL) techniques have increasingly been applied to HAR. Facilitated by large datasets and

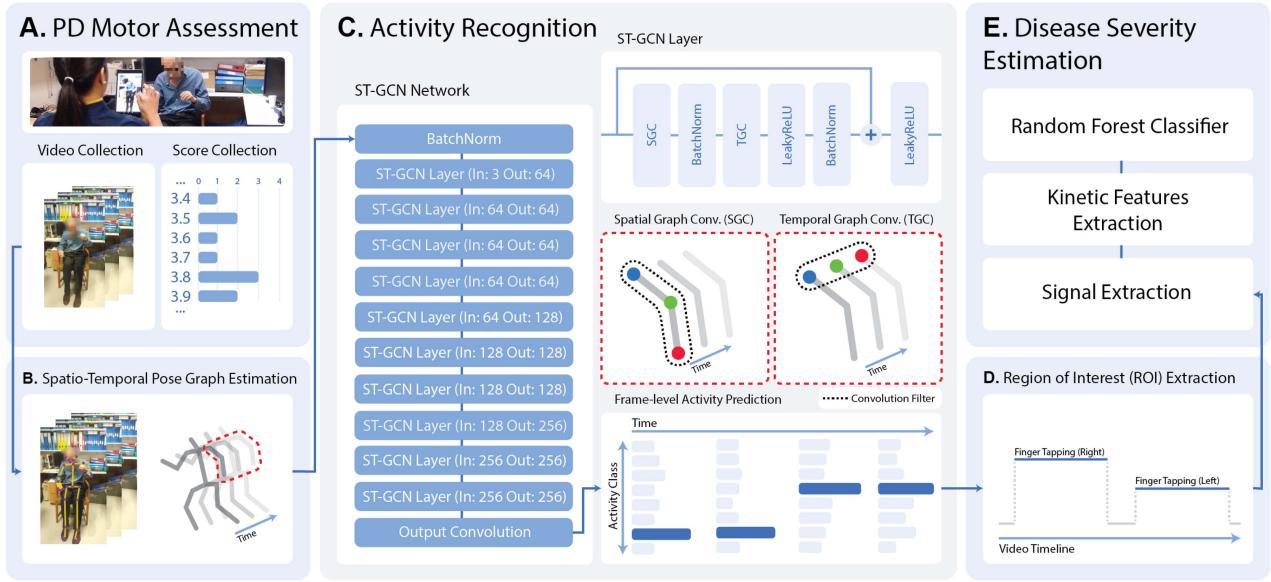


Fig. 1. Automated motor assessment pipeline. (A) The MDS-UPDRS motor assessment is recorded, for example, using a mobile device such as a tablet and a collection of videos and corresponding severity scores is saved. (B) A pose estimation library *OpenPose* is used to estimate the patient's joint locations in each video frame and a spatio-temporal graph representation of the patient's movement is created. (C) The pose graph is processed using a deep-learning model which outputs a vector of predicted activity confidence scores for each frame of the input. (D) Regions of patient activity are identified based on the frame-level prediction output. (E) Activity-specific joint-based signal is extracted using the predicted region of activity and forms an input to a disease severity estimation model [10], [11], [12].

faster computing, DL frequently yields superior performance compared to more traditional “shallow” techniques. An added benefit of many DL methods is the end-to-end learning and prediction setup, removing the need for feature engineering and validation [14], [15].

Previous HAR work broadly falls into two categories according to input data, accelerometer data [16] and image data. Within vision-based HAR, two further subdivisions can be delineated:

- 1) Pixel-based activity recognition from a video stream [17], [18]. These methods process the video input captured by a camera directly which allows the model to extract features describing both the human subjects and their surroundings. However, the typically large input dimensionality makes these methods vulnerable to over-fitting in the absence of vast quantities of training data, which may prevent such systems from generalising well [19].
- 2) A two-step framework whereby human pose estimation (i.e. annotation of body joint locations) is performed first, followed by activity recognition based on the estimated key points [20], [21]. In this approach, the prediction errors from both pose estimation and activity recognition modules may compound, but the great dimensionality reduction implicit in the first stage appears to more than compensate, resulting in a good performance on tractable data sets.

In the context of Parkinson's disease, HAR methods have been applied to recognise a set of universal activities such as walking or standing [22], [23], but its application to the clinically relevant MDS-UPDRS motor assessment remained, to our knowledge, unexplored. Additionally, the performance of

deep learning methods in medical settings is often limited by the amount of available training data, preventing in-depth evaluation of typically data-hungry deep learning frameworks [24].

D. Our Approach

In this work, we present an application of vision-based HAR method for the classification of motor assessment tasks performed by Parkinson's patients. We train a deep learning model to classify patient's activity based on estimated body joint locations obtained using an open-source pose estimation library *OpenPose* [25] and evaluate the approach using a large multi-site dataset of MDS-UPDRS part 3 assessment video recordings. We extend the framework beyond video-level activity classification to perform the prediction on a frame-by-frame basis, allowing the model to recognise multiple activities in any order within a single video clip of arbitrary length and specify the beginning and the exact duration of the activity. This flexibility predisposes the system to a greater number of real-world applications.

II. METHODS

Fig. 1 illustrates the pipeline for video collection, pose estimation, and activity recognition. The introduced activity recognition module can facilitate the automatic localisation of MDS-UPDRS motor tasks within videos, accommodating further analysis such as model-based disease severity estimation. Clinical assessments of Parkinson's patients were conducted across multiple centres in the U.K. and USA, in the course of clinical care. Video recordings of these assessments were

TABLE I
VIDEO AND SEVERITY RATING COUNTS BY MDS-UPDRS ITEM CLASS

MDS-UPDRS item	Dataset video count			Activity class	Severity rating				
	Cross-validation	Held-out	Total		0	1	2	3	4
Finger Tapping (FT)	1022	56	1078	left hand FT	152	437	295	157	37
				right hand FT	177	471	285	123	22
Hand Movement (HM)	1085	55	1140	left hand HM	235	493	271	123	18
				right hand HM	263	565	246	58	8
Pronation Supination (PS)	844	51	895	left hand PS	194	329	237	111	24
				right hand PS	250	383	186	70	6
Toe Tapping (TT)	1000	42	1042	left leg TT	191	391	255	171	34
				right leg TT	240	477	213	90	22
Leg Agility (LA)	895	35	930	left leg LA	267	375	192	70	26
				right leg LA	327	389	157	41	16
Arising From Chair (AFC)	921	36	957	AFC	648	235	39	24	11
Gait	867	49	916	away from camera Gait towards camera Gait	213	473	195	31	4
Postural Tremor of Hands (PT)	327	25	352	PT	216	103	17	10	6
				-	225	92	20	13	2
N/A	-	-	-	No Action	-	-	-	-	-
Total	6961	349	7310	15 classes	3598	5213	2608	1092	236

For some items, two activities can be distinguished, which is reflected in the activity class labels. The cumulative distribution of clinical MDS-UPDRS severity ratings exhibits a strong positive skew with a mode rating of 1.

collected through KELVIN™ [26], a video-based motor assessment platform developed by Machine Medicine Technologies (MMT) [27].

A. Data

1) Activity Classes: The MDS-UPDRS instructions specify a set of items to be performed by a patient and evaluated by an expert clinician in terms of PD symptom severity. In our dataset, every assessment has been recorded and saved in the form of multiple video clips, one video per item. For training of the activity recognition model, we selected a subset of items that can be performed without assistance, taking into consideration overall informativeness about a patient's holistic disease state [28]. Some MDS-UPDRS items correspond to two distinct activities, such as *left* and *right hand Finger Tapping*. This has been reflected in the activity class labels, shown in Table I, as the model makes the distinction between laterals. An additional *No Action* class has been included to indicate when patients are not performing any of the target activities.

2) Multi-Site Patient Population: The data include 7310 videos recorded as a part of 1170 PD motor assessments completed at five sites: DCMN (Department of Clinical and Movement Neurosciences, Institute of Neurology, University College London), NRC (Neuroscience Research Centre, Molecular and Clinical Sciences Research Institute, St. George's, University of London), DRC (Dementia Research Centre, Institute of Neurology, University College London), PDMDC (Parkinson's Disease and Movement Disorders Centre, Baylor College of Medicine), TSL (The Starr Lab, University of California San Francisco).

To evaluate the cross-patient performance of the model we utilise a 5-fold cross-validation split of the data on an assessment level, stratified by the patient's disease severity. This way, all videos from a single assessment appear in only one of the cross-validation folds, ensuring that the model is not trained and tested using the same patient data.

To test the model's ability to generalise to unseen assessment environments, data from a single site, namely TSL, have been held out of the cross-validation setup for cross-site performance comparison. The distribution of assessments across sites is shown in Table II. Furthermore, data on patient characteristics have been available for 620 out of 1170 assessments. A summary of the information is presented in Table III.

3) Region of Interest: Along with the MDS-UPDRS item label, the dataset contains a human-annotated region of interest (ROI) for each activity in the video. The ROI specifies the interval of frames within which an item-specific activity took place. Combining the ROI annotation with the activity label enables frame-level prediction training, where each frame within the specified range is considered as belonging to the target activity class. Frames outside the annotated region are assigned the *No Action* label.

4) Data Collection: The video assessments used for this work were collected using consumer-grade handheld devices with an installation of KELVIN™, a clinical tool that has been adopted in line with the respective procedures of each institution. No specific requirements regarding the distance of the camera to the subject were imposed during data collection. Informed consent was obtained from all subjects and the agreements formed with each institution allow for the data to be used in

TABLE II
SITE DISTRIBUTION

Site	Assessment count	MDS-UPDRS part 3		Hoehn & Yahr stage							Dataset
		Mean (SEM)	Q1-Q3	0	1	2	3	4	5	Missing	
DCMN	657	30.4 (0.7)	18-40	16	20	564	30	20	5	2	Cross-validation
NRC	221	35.4 (1.2)	22-46	1	2	159	44	14	1	-	
DRC	134	18.3 (1.1)	9-25	38	32	53	9	2	0	-	
PDMDC	101	36.7 (1.3)	27-46	0	2	88	7	2	2	-	
TSL	57	35.2 (2.6)	22-43	0	0	44	13	0	0	-	Held-out
Total	1170	30.7 (0.5)	18-41	55	56	908	103	38	8	2	

Model cross-validation has been conducted using data from 4 assessment sites. The remaining site, TSL, has been held out of the cross-validation setup to test the model's ability to generalise to unseen environments. In total, the dataset comprises 1170 MDS-UPDRS part-3 motor assessments.

TABLE III
PATIENT CHARACTERISTICS

Site	Age in years		Disease duration in years		Sex		Medication		DBS	
	Mean (SEM)	Q1-Q3	Mean (SEM)	Q1-Q3	Female	Male	Off	On	Yes	No
DCMN	59 (0.5)	54-66	7 (0.4)	3-9	86	209	179	116	261	34
NRC	61 (0.6)	57-66	12 (0.5)	9-14	40	116	42	114	39	117
DRC	66 (0.9)	59-72	5 (0.3)	3-7	41	45	8	78	85	1
PDMDC	63 (1.7)	58-71	9 (0.6)	6-10	15	22	30	7	37	0
TSL	52 (2.3)	40-63	10 (0.6)	7-13	9	37	10	36	0	46

Additional subject information has been available for 620 out of 1170 assessments. We present summary statistics (mean, SEM, lower- and upper-quantiles) of patients' age and disease duration broken down by assessment site, and show the split by sex, medication state and presence of a deep brain stimulation device.

this context. A subset of the data has been collected as part of clinical trials: Exenatide for Parkinson's Disease (EXENATIDE-PD, NCT01971242), Antidepressants Trial in Parkinson's Disease (ADept-PD, NCT03652870) and The Motor Network in Parkinson's Disease and Dystonia: Mechanisms of Therapy (NCT03582891).

5) Pose Estimation Processing: Prior to activity classification, 2D coordinates of estimated body joint locations were extracted from every frame in each video using a markerless deep-learning pose estimation library OpenPose [25]. The framework is integrated as part of the KELVIN™ platform. Pixel coordinates and estimation confidence scores for 25 body key points and 10 hand key points (fingertips) have been selected to capture poses and movements characteristic to the predicted tasks.

B. Model

1) Spatio-Temporal Pose Graph: Instead of treating each key-point independently, the model views the input as a spatio-temporal graph representing the evolution of the estimated body pose. Key-point locations correspond to the graph vertices and are connected by two sets of edges: 1) spatial and 2) temporal. The spatial edges connect joints within each frame, according to the anatomical construction of the human skeleton. The temporal edges connect corresponding joints in neighbouring frames. A simplified representation of a spatiotemporal pose graph has been depicted in Fig. 1(B).

2) Network Architecture: The model architecture is based on the ST-GCN model introduced in [14]. It consists of 10

graph convolution layers operating in both spatial and temporal dimensions, followed by a final prediction convolution layer. Compared to common convolution neural networks (CNNs) where the convolution window slides over a well-defined grid-structured input, graph convolution networks (GCNs) perform the calculation on a graph-based input instead.

Given a series of poses, the original model outputs a single action prediction vector of size (C), where C is the number of target classes. It makes use of strided temporal convolutions and a global averaging layer. Increasing the stride of temporal convolutions increases the receptive field of successive layers and allows the model to capture features of predicted activity spanning a broader time window. The global averaging layer allows video-level action prediction for image sequences of arbitrary length.

The model structure has been augmented to predict the performed action on a frame level, outputting a prediction vector of shape (T, C) . To preserve the size of the input's time dimension T , the stride values in all temporal convolution layers have been set to 1, and the time averaging step has been omitted.

C. Training Setup

1) Data Sampling: Each input video captures a single MDS-UPDRS item and may contain up to two ROI annotations (one for each lateral activity). To prevent the model from learning order dependencies between action classes (for example, right-hand movements are generally followed by left-hand movements), the input for double-labelled items is split between the two ROI annotations during training.

Manual ROI labellers were instructed to label only one example of an activity class in each video. To avoid presenting the model with a *No Action* label for non-annotated regions of true activity, for training, we cut the input to a maximum of 45 frames outside either end of the annotated ROI. To offset the loss of true *No Action* regions further away from the ROI, we reverse the remaining 45-frame segments of *No Action* and append them to the ends of the input. For validation and testing no cutting has been applied.

The duration of PD assessment can vary considerably within and between activity classes. Considering the frame-level prediction setup we leverage the fully convolutional model architecture and view each input as a set of T examples and randomly sample a fixed number of consecutive frames (sampling window) from the processed videos to accommodate input length differences for batch training. We have chosen a sampling window of 150 frames and a batch size of 16. Model validation and testing have been performed without subsampling.

2) Data Augmentation: During training, the sampled pose sequences were randomly augmented using four transformations: shift, rotation, scaling and horizontal flip; designed to simulate varying camera placements. Moreover, flipping the pose horizontally allowed for left-lateral movements to be considered as right-lateral activity examples and vice versa. Additionally, out-of-frame masking was randomly applied to chosen classes, removing pose information from lower or upper parts of the video, for upper- and lower-body activity classes respectively.

3) Training: The models have been trained using the common cross-entropy loss L_{CE} . In the frame-level setup, it took the form:

$$L_{CE} = - \sum_{n=1}^N \left(\sum_{l=1}^L \left(\sum_{c=1}^C \log(x_{n,l,c}) q_{n,l,c} \right) \frac{1}{L} \right) \frac{1}{N} \quad (1)$$

where N is the size of a sample batch, L is the length of a sampling window and C is the number of predicted classes. $x_{n,l,c}$ is the predicted probability for activity class c in the l th frame of the n th sample. $q_{n,l,c}$ is the target activity label for the corresponding frame. Target labels have been augmented according to a smoothing strategy introduced in [29]. To prevent the model from becoming too confident in a single prediction and overfitting to the training data, the true binary frame label y is transformed according to:

$$q(y) = (1 - \epsilon)y + \frac{\epsilon}{C} \quad (2)$$

The smoothing parameter ϵ has been set to 0.1 as per the original implementation.

The training gradient descent has been performed using the Adam [30] optimiser with an initial learning rate of 0.01. A maximum number of training epochs has been set to 100. The adjustment of the learning rate during training has been guided by a linear scheduler, which monitored a validation set loss and decreased the learning rate by a factor of 3 if no validation score improvement has been recorded for 5 consecutive epochs. Model training has been distributed using AWS batch service over GPU-enabled (NVIDIA V100) compute instances.

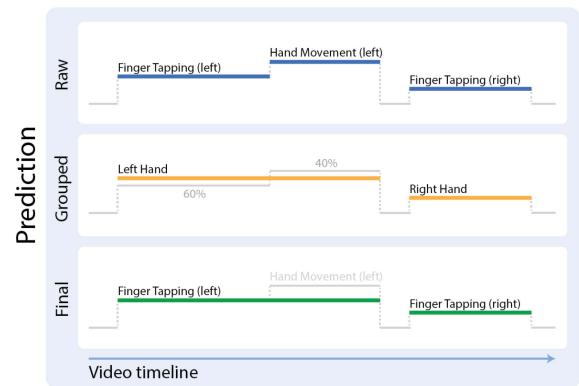


Fig. 2. Model-predicted ROI extraction strategy. Raw frame-level output of the ST-GCN network (top). Connected regions of similar activities are grouped together (middle). Final prediction output after frame majority voting over grouped regions (bottom).

D. Post-Processing and ROI Prediction

One of the objectives of the framework is to accurately annotate the region of a patient's activity. As the activity prediction is performed on a frame-by-frame basis, the connectivity of outputted single-class regions may be affected by factors such as noise present in pose estimation data. To counter confusion between classes sharing similar poses and movement patterns, for example *left hand FT* and *left hand HM*, the frame-level output is processed by first finding connected regions of limb-grouped activities, and then assigning the final class label to each region according to the majority vote of the constituting frames. The process is visualised in Fig. 2. Since the human-annotated ROIs specify only one region of distinct activity per video, the longest connected region of each predicted activity class is considered the final model annotation.

E. Evaluation

1) Video-Level Classification: In the video-level setup, given a series of poses from a single video, the model has to identify the corresponding MDS-UPDRS task being performed by the patient.

Model performance is evaluated using two metrics: balanced accuracy: prediction accuracy averaged across MDS-UPDRS items; and top-3 accuracy, where the target class is among the top 3 most confident predictions.

We compare the ST-GCN's performance with two "shallow" machine learning models: KUPDRS [10], [11], [12], originally developed for PD severity estimation, extracts MDS-UPDRS item-specific signals and features and performs the prediction using a random forest classifier. The model has been adopted to perform activity recognition instead of severity estimation; Cov3DJ [31], constructs a key-point covariance-based descriptor to characterise pose movements throughout the video. The obtained representation is classified using a linear SVM model.

2) ROI Prediction: Predicting the activity on a frame-by-frame basis allows the model to localise and classify regions of patient activity within an assessment video. The processing

of the raw model prediction into ROIs has been described in II-D. We choose a popular segmentation accuracy measure, the intersection over union (IOU), as the primary metric to evaluate the correspondence of human- and model-provided ROI annotations. The IOU score specifies the quality of overlap between two activity region annotations. It has been calculated for all labelled regions of patient activity.

3) Annotation Agreement Comparison: To analyse the level of agreement between different human annotations and compare them to the model-predicted ROI, 80 videos (10 videos per predicted MDS-UPDRS item), not found in the main dataset and without an existing ROI annotation, have been selected and independently annotated by three human labellers.

For each such video, we construct a “robust” label by combining two human annotations. Given 3 possible pairings of human labellers, we obtain 3 such annotations. Each “robust” annotation consists of 3 regions: 1) the overlap of two ROI annotations, 2) segments covered by only one of the ROI annotations, and 3) frames covered by neither ROI annotation. Regions 1 and 3 are considered “robust” annotations of activity and *No Action* regions respectively and form the video ground-truth label. Region 2 is regarded as an area of human disagreement and is excluded from evaluation.

We then compute the IOU score for all “robust” labels, once using the model prediction, and once based on the remaining human annotation (one that has not been used to construct the “robust” label). We compare the scores to see whether the model- and human-predicted activity regions differ significantly in the presence of inter-labeller variability.

4) Disease Severity Estimation: For every video in the dataset, there exists a set of corresponding PD severity scores assigned by expert clinicians according to the MDS-UPDRS manual. Previous work has shown that machine-learning models can analyse PD assessment videos within manually annotated ROIs of patient activity through the extraction of kinetic signals and features, and yield estimates of the motor impairment severity [10], [11], [12], [32]. To test the adequacy of using model-predicted ROIs for PD severity inference, we cross-validate the severity-estimation models trained with human-annotated ROI data and compare test-fold prediction accuracy based on both human- and model-annotated ROIs.

III. RESULTS

A. Video-Level Classification

The confusion matrix for video-level activity classification is shown in Fig. 3. The trained model achieves a mean cross-validation accuracy score of 96.51% balanced across predicted MDS-UPDRS items. Noticeable confusion occurs between items sharing similar movement patterns and poses: hand-items including *FT*, *HM* and *PS*; as well as leg-items *TT* and *LA*.

Cross-validation and held-out site performance comparison of the deep learning framework with two “shallow” machine learning models for video-level activity classification is shown in Table IV. Although the traditional machine learning models have indicated the ability to discriminate between the target classes in most cases, the deep learning model significantly outperforms both alternatives.

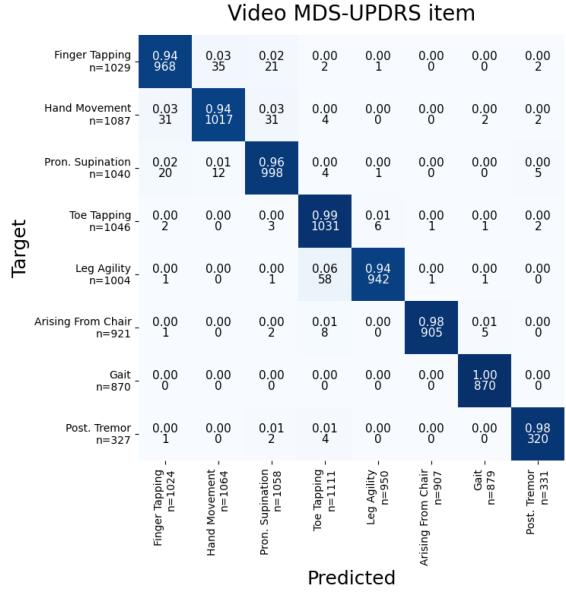


Fig. 3. Row-normalised confusion matrix for video-level classification of MDS-UPDRS related activities. The ST-GCN achieves almost perfect prediction accuracy regardless of the target item. Most prominent confusion arises within groups of hand- and leg-related items, which could be explained by the high similarity of poses involved.

TABLE IV
VIDEO-LEVEL ACTIVITY RECOGNITION

Model	Balanced Accuracy		Top-3 Accuracy	
	Cross-val. mean [min-max]	Held-out (N)	Cross-val. mean [min-max]	Held-out (N)
KUPDRS	79.02% [77.2-80.6%]	76.98% (329)	96.86% [96.1-97.4%]	93.01% (329)
Cov3DJ	82.61% [81.3-83.4%]	72.75% (386)	98.74% [98.5-99.0%]	96.37% (386)
ST-GCN	96.51% [96.0-97.2%]	91.82% (386)	99.69% [99.6-99.9%]	98.96% (386)

Comparison of video-level activity classification performance between ST-GCN and two traditional machine learning models. The 5-fold cross-validation average test performance is shown alongside the held-out site score. The ST-GCN significantly outperforms both “shallow” models across all presented metrics.

The held-out test site performance decreases compared to the cross-validation result for all models, suggesting the methods might not generalise as well to unseen assessment settings. Further inspection of the results revealed the KUPDRS model can be sensitive to patients not being fully in-frame, which leads to the omission of samples. Recalculation of ST-GCN held-out site scores excluding samples discarded by the KUPDRS model yields 95.19% balanced accuracy and 99.39% top-3 accuracy scores, almost matching the cross-validation data performance.

B. ROI Prediction

Compared to human annotators, the model has no prior knowledge of the motor assessment item being captured in the video and can misclassify the identified activity ROI. Misclassification of a correctly localised region of activity will result in a zero IOU score. Similarly, if the longest-connected region of the target

TABLE V
REGION OF INTEREST PREDICTION

Activity class	Cross-validation sites		Held-out site		p-value	
	mean IOU (non-zero)	N (non-zero)	mean IOU (non-zero)	N (non-zero)	all	non-zero
Finger Tapping - Left hand	0.821 (0.891)	1022 (942)	0.890 (0.890)	56 (56)	0.995	0.957
Finger Tapping - Right hand	0.791 (0.860)	1022 (941)	0.805 (0.835)	56 (54)	0.918	0.824
Hand Movement - Left hand	0.837 (0.913)	1085 (995)	0.819 (0.920)	55 (49)	0.370	0.497
Hand Movement - Right hand	0.856 (0.917)	1085 (1013)	0.824 (0.889)	55 (51)	0.515	0.552
Pron. Supination - Left hand	0.848 (0.889)	844 (805)	0.745 (0.810)	50 (46)	0.031	0.055
Pron. Supination - Right hand	0.863 (0.897)	844 (812)	0.687 (0.778)	51 (45)	<0.001	<0.001
Toe Tapping - Left leg	0.765 (0.795)	1000 (963)	0.600 (0.630)	42 (40)	<0.001	<0.001
Toe Tapping - Right leg	0.786 (0.816)	1000 (963)	0.610 (0.641)	42 (40)	<0.001	<0.001
Leg Agility - Left leg	0.834 (0.879)	894 (848)	0.636 (0.655)	35 (34)	0.017	0.006
Leg Agility - Right leg	0.869 (0.900)	895 (864)	0.670 (0.689)	35 (34)	0.004	0.002
Arising From Chair	0.610 (0.643)	920 (873)	0.443 (0.514)	36 (31)	<0.001	0.001
Gait - From camera	0.730 (0.857)	867 (738)	0.628 (0.832)	49 (37)	0.061	0.288
Gait - Towards camera	0.735 (0.827)	867 (771)	0.593 (0.830)	49 (35)	0.042	0.674
Postural Tremor of Hands	0.467 (0.807)	327 (189)	0.782 (0.888)	25 (22)	0.998	0.728
Mean	0.772 (0.849)		0.695 (0.771)			

ROI prediction performance measured by the IOU score for cross-validation and held-out datasets. Zero values of the metric can occur when an incorrect activity class has been matched to the human-labelled region of activity or a correct, but unlabelled, region has been chosen by the model. Metric values for samples where the model has correctly localised the target activity region have been reported in brackets. The difference between cross-validation and held-out site results has been tested using the (one-sided) Mann Whitney U-test [33]. Corresponding p-values have been reported and significance at 5% marked with the Bonferroni correction applied.

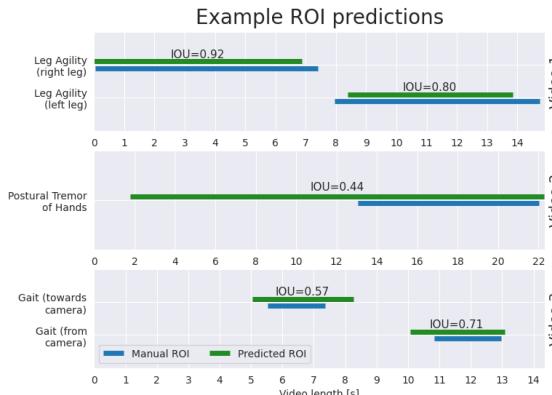


Fig. 4. Examples of human- and model-annotated regions of activity. Video 1) The model has been able to identify the activity in the very first frame of the video despite the absence of prior context. Video 2) The predicted region of *Postural Tremor of Hands* extends significantly beyond the human annotation. The way *PT* activity is performed can vary between assessment sites, leading to inconsistency in labelling. Video 3) Regions of correct activity are accurately localised within the video. Despite the relatively low IOU score for *Gait Towards Camera* class, the automatically annotated region exceeds the human label by less than a second on either end.

activity (chosen as the final model ROI prediction according to II-D) has been unlabelled, the metric score will again be 0. Mean IOU scores, both including and excluding misclassified or mislocalised model ROIs have been reported in Table V.

For correctly localised and classified regions of activity (the predicted region overlaps with the labelled one), the mean IOU score across most of the activity classes exceeds 0.8. Examples of model-annotated regions of activity together with corresponding IOU scores are visualised in Fig. 4. The cross-validation

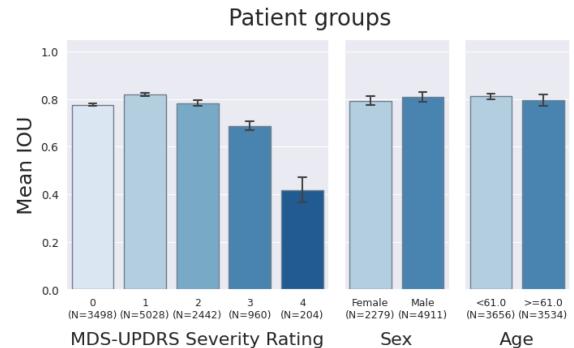


Fig. 5. ROI prediction performance for different patient groups. The performance decreases for patients with higher PD severity. The severity score of 4 is often assigned to patients who are unable to perform the desired activity. A lack of kinetic information can lead to model failure, as it cannot see the subject's intention to perform the task. The model displays no significant difference within sex and age groups. Age grouping has been based on the median age. Standard deviations of 5-fold cross-validation results were used to produce the presented error bars.

results broken down by disease severity, sex and age are shown in Fig. 5.

The misclassification rate for hand and leg-related items is around 5%, which corresponds to the video-level confusion for these tasks. For *Gait*, patients often perform the task twice, repeating the walk towards and away from the camera. Choice of the unlabelled *Gait* ROI contributes to its mislocalisation rate. Although *Postural Tremor of Hands* has been well-identified on the video level, many of the predicted regions do not overlap with the human annotation. *PT* assessment can vary between sites, which can cause divergence in human labelling and negatively

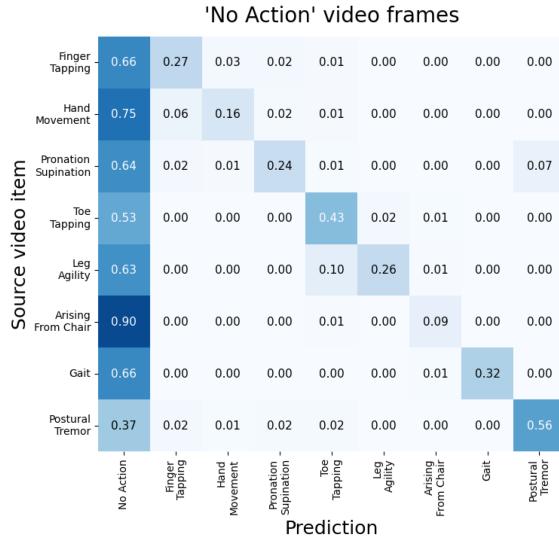


Fig. 6. Confusion between *No Action* frames and predicted classes, grouped by MDS-UPDRS items. The *No Action* frames are mainly confused with the video item label, which suggests that regions of true activity are over-predicted by the model, or, more likely in our view, have not been annotated.

impact model training and evaluation. For overlapping predictions, however, the IOU score matches that of other activity classes.

Although most of the items have shown no significant decrease in performance for the held-out site, notable differences can be seen for leg-related activities and *Arising from Chair*. Substantially different camera placements or failure to capture the entire patient in the frame can negatively affect the connectivity of the predicted activity region. The result highlights the importance of data capture standardisation.

The confusion for *No Action* frames grouped by MDS-UPDRS items is shown in Fig. 6. The regions of *No Action* are mostly confused with the corresponding video item label, despite the similarity of *No Action* poses between classes such as sitting down for all hand- and leg-items. This may suggest that the ground-truth ROIs occasionally cover only a subsection of the intervals during which the relevant activity was being performed, or the video contains an additional, unlabelled region of the target activity.

C. Multi-Item Video Stream

One of the benefits of a frame-by-frame activity prediction setup is the ability to process video inputs of arbitrary length and automatically annotate and classify multiple regions of patient activity. To validate the model's capability of continuous, multi-item activity recognition, a single video of a subject performing MDS-UPDRS specified activities in a randomised order has been recorded and processed by the proposed framework. As shown by Fig. 9, the model has identified all regions of human activity and correctly classified 91.8% of all frames labelled by a human annotator.

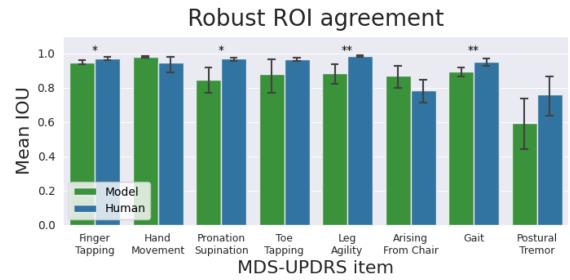


Fig. 7. Comparison of human and model performance in identifying "robust" regions of patient activity. For half of the item classes, the model-human agreement level matches that of human-human. 95% confidence intervals presented in the figure are based on 1000 bootstrap iterations. Annotation of significant differences is based on a paired t-test with Bonferroni correction applied to the resulting p-values.

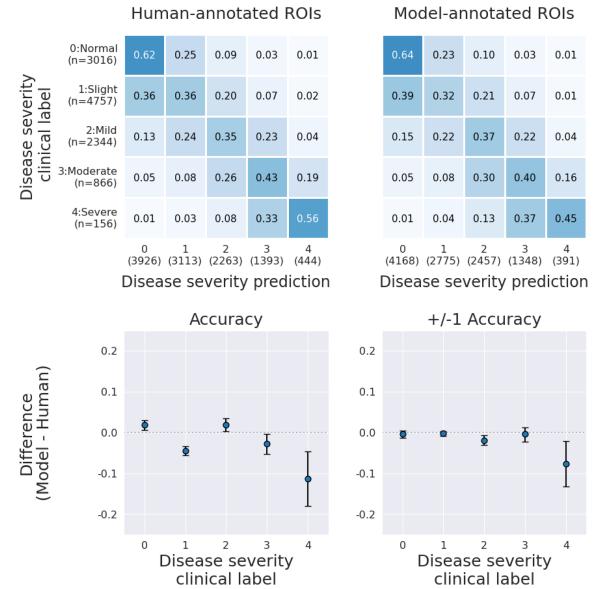


Fig. 8. Row-normalised confusion matrices of disease severity estimates based on human-annotated ROIs (Top-left) and model-annotated ROIs (Top-right). The difference in severity score estimation accuracy based on human- and model-annotated regions of activity with 95% bootstrap confidence intervals (Bottom). In summary, for most clinical scores the severity estimation models display no systematic differences in performance when model-annotated ROIs and associated kinetic features are used to infer the disease severity.

D. Annotation Agreement Comparison

Comparison of mean IOU for each predicted MDS-UPDRS item under robust ROI labels, for human annotations and model predictions, has been shown in Fig. 7. The model, in general, performs on par with human annotators. Despite the statistically significant differences in mean IOUs between human and model video annotations, the magnitudes may be considered negligible. For *Pronation Supination* and *Leg Agility*, the significant difference in reported TPR arises due to the model identifying the wrong class, while still accurately identifying the activity region. This happens in 1-2 videos of these items. For *Gait*, the decrease

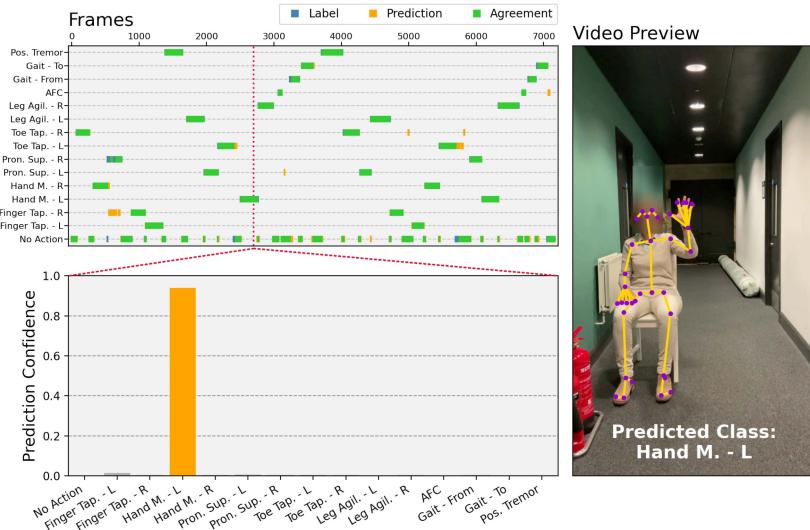


Fig. 9. Example implementation of continuous activity recognition with raw (no post-processing) frame-level prediction. The model has been able to identify multiple activity classes in a single video, with overall frame prediction accuracy of 91.8%. Video available at <https://youtu.be/7dqgAHdJank>.

in performance may be caused, in part at least, by noise present in pose estimates when patients are far from the camera. The difficulty of consistent *PT* and *AFC* activity annotation is shown by relatively low human-to-human agreement, which coincides with previous results.

E. Disease Severity Estimation

Severity-estimation model results based on human and model ROI annotations are presented in Fig. 8. Model annotated ROIs cause no systematic decrease in SE model performance for severity scores lower or equal to 3. In particular, acceptable accuracy, or $+/-1$ accuracy (i.e. prediction error within 1 point on the MDS-UPDRS scale), shows no clinically nor statistically significant change. As previously shown in Fig. 5, the ROI prediction accuracy decreases for the highest-severity patients. The model might fail to recognise when the patient intends but is unable to perform the activity, which consequently would bias the severity estimate towards lower scores, as we limit the input to “active” regions only.

IV. DISCUSSION

A. Summary of Result

We developed a vision-based HAR framework capable of dynamically classifying motor assessment tasks performed by Parkinson’s disease patients. A deep learning model, utilising the estimated body joint locations of the patient as input, was trained to distinguish between 8 MDS-UPDRS items and corresponding 15 activity classes. The model achieves 96.51% balanced accuracy in video-level activity classification and outputs highly accurate annotations at a frame-by-frame temporal resolution.

We have additionally shown how our method generalises to new patient data, recorded at settings not seen during training, displaying comparable performance on data collected at a held-out site for most of the predicted activities.

By comparing mean inter-human and human-model differences in the annotation of patient activity regions, we demonstrated a close correspondence between human and model activity labelling. Finally, using model-annotated regions of interest for algorithmic disease severity estimation yielded similar performance, potentially obviating the need for (previously required) human labelling.

B. Limitations and Future Work

While in most cases the model is able to provide a continuous region of correct activity class prediction, areas of prominent confusion have been identified. Further work will be undertaken to improve the activity recognition accuracy for classes sharing similar movement patterns and body poses. Exploration of extensions and alternatives to the presented framework inspired by advances in general HAR modelling will be a part of future research endeavours. This may entail the addition of higher-order pose information including inter-frame key-point velocity or acceleration.

Considering the effect of inconsistent ROI labelling on frame-level-based activity prediction training, a model-assisted relabelling effort may be justified. Additionally, techniques such as self-distillation may be used to offset the negative impact of noisy labels.

The presented activity recognition pipeline has been developed using cloud servers. However, many practical applications will require real-time local computation that is deployable on commercially available mobile device hardware. Future development efforts may therefore focus on mobile device implementation.

C. Practical Applications

Combined with disease severity estimation models [10], [11], [12], this HAR framework is a crucial element in achieving remote, asynchronous PD motor assessment at scale. Potential

applications include, but are not limited to, massive remote clinical trials, closed-loop and patient-driven programming of both invasive and non-invasive neuromodulation devices, and the general medical management of patients.

ACKNOWLEDGMENT

Data were collected at; Department of Clinical and Movement Neurosciences, Institute of Neurology, University College London; Neuroscience Research Centre, Molecular and Clinical Sciences Research Institute, St. George's, University of London; Dementia Research Center, Institute of Neurology, University College London; Parkinson's Disease and Movement Disorders Center, Baylor College of Medicine; The Starr Lab, University of California San Francisco. We thank all the staff involved in the data collection.

We thank the employees at Machine Medicine Technologies for labelling the regions of interest in all videos.

Competing interests: Machine Medicine Technologies have funded this work and are the owner of the developed technology. The Authors declare no other Competing Financial or Non-Financial Interests.

REFERENCES

- [1] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern discovery," *IEEE Pervasive Comput.*, vol. 9, no. 1, pp. 48–53, Jan.–Mar. 2010.
- [2] M. Popa, A. K. Koc, L. J. M. Rothkrantz, C. Shan, and P. Wiggers, "Kinect sensing of shopping related actions," in *Proc. Int. Joint Conf. Ambient Intell.*, 2012, pp. 91–100.
- [3] R. Nar, A. Singal, and P. Kumar, "Abnormal activity detection for bank ATM surveillance," in *Proc. IEEE Int. Conf. Adv. Comput. Commun. Inform.*, 2016, pp. 2042–2046.
- [4] A. Almeida and A. Alves, "Activity recognition for movement-based interaction in mobile games," in *Proc. 19th Int. Conf. Human-Comput. Interact. Mobile Dev. Serv.*, New York, NY, USA, 2017, Art. no. 55, doi: [10.1145/3098279.3125443](https://doi.org/10.1145/3098279.3125443).
- [5] T. T. Zin et al., "Real-time action recognition system for elderly people using stereo depth camera," *Sensors*, vol. 21, no. 17, 2021, Art. no. 5895. [Online]. Available: <https://www.mdpi.com/1424-8220/21/17/5895>
- [6] S. L. Lau, I. König, K. David, B. Parandian, C. Carius-Düssel, and M. Schultz, "Supporting patient monitoring using activity recognition with a smartphone," in *Proc. 7th Int. Symp. Wireless Commun. Syst.*, 2010, pp. 810–814.
- [7] G. D. Giebel and C. Gissel, "Accuracy of mhealth devices for atrial fibrillation screening: Systematic review," *JMIR Mhealth Uhealth*, vol. 7, no. 6, Jun. 2019, Art. no. e13641. [Online]. Available: <https://doi.org/10.2196/13641>
- [8] S. N. Gajjarawala and J. N. Pelkowski, "Telehealth benefits and barriers," *J. Nurse Practitioners*, vol. 17, no. 2, pp. 218–221, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1555415520305158>
- [9] B. Romaissa, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: A survey," *Multimedia Tools Appl.*, vol. 79, pp. 30509–30555, 2020.
- [10] S. Rupprechter et al., "A clinically interpretable computer-vision based method for quantifying gait in Parkinson's disease," *Sensors*, vol. 21, no. 16, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/16/5437>
- [11] G. Morinan et al., "Computer-vision based method for quantifying rising from chair in Parkinson's disease patients," *Intell.-Based Med.*, vol. 6, 2022, Art. no. 100046. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666521221000223>
- [12] G. Morinan et al., "Computer vision quantification of whole-body Parkinsonian bradykinesia using a large multi-site population," *npj Parkinson's Dis.*, vol. 9, no. 1, Jan. 2023, Art. no. 10. [Online]. Available: <https://doi.org/10.1038/s41531-023-00454-8>
- [13] C. G. Goetz et al., "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (MDS-UPDRs): Scale presentation and clinimetric testing results," *Movement Disord.: Official J. Movement Disorder Soc.*, vol. 23, no. 15, pp. 2129–2170, 2008.
- [14] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc AAAI Conf. Artif. Intell.*, 2018, vol. 32.
- [15] W. Yang, J. Zhang, J. Cai, and Z. Xu, "Shallow graph convolutional network for skeleton-based action recognition," *Sensors*, vol. 21, no. 2, 2021, Art. no. 452. [Online]. Available: <https://www.mdpi.com/1424-8220/21/2/452>
- [16] N. Dua, S. Singh, and V. Semwal, "Multi-input CNN-GRU based human activity recognition using wearable sensors," *Computing*, vol. 103, pp. 1–18, 2021.
- [17] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
- [18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.
- [19] L. Chen, *Curse of Dimensionality*. Berlin, Germany: Springer, 2009, doi: [10.1007/978-0-387-39940-9_133](https://doi.org/10.1007/978-0-387-39940-9_133).
- [20] F. M. Noori, B. Wallace, M. Z. Uddin, and J. Torresen, "A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network," in *Proc. Image Anal. 21st Scand. Conf.*, 2019, pp. 299–310, doi: [10.1007/978-3-030-20205-7_25](https://doi.org/10.1007/978-3-030-20205-7_25).
- [21] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12026–12035.
- [22] M. Kazemimoghadam and N. P. Fey, "An activity recognition framework for continuous monitoring of non-steady-state locomotion of individuals with parkinson's disease," *Appl. Sci.*, vol. 12, no. 9, 2022, Art. no. 4682, doi: [10.3390/12094682](https://doi.org/10.3390/12094682).
- [23] B. Rezaei et al., "Target-specific action classification for automated assessment of human motor behavior from video," *Sensors*, vol. 19, no. 19, 2019, Art. no. 4266. [Online]. Available: <https://www.mdpi.com/1424-8220/19/19/4266>
- [24] D. Chen et al., "Deep learning and alternative learning strategies for retrospective real-world clinical data," *Npj Digit. Med.*, vol. 2, no. 1, May 2019, Art. no. 43, doi: [10.1038/s41746-019-0122-0](https://doi.org/10.1038/s41746-019-0122-0).
- [25] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, p. 1.
- [26] Machine medicine technologies limited, "Kelvin software webpage," 2021. [Online]. Available: <https://machinemedicine.com/getting-started/>
- [27] Machine Medicine Technologies Limited, "The company's website," 2021. [Online]. Available: <https://machinemedicine.com/>
- [28] G. Morinan, R. A. Hauser, A. Schrag, J. Tang, J. O'Keeffe, and M.-N. S. D. S. Group, "Abbreviated MDS-UPDRs for remote monitoring in PD identified using exhaustive computational search," *Parkinson's Dis.*, vol. 2022, Jun. 2022, Art. no. 2920255, doi: [10.1155/2022/2920255](https://doi.org/10.1155/2022/2920255).
- [29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
- [30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2014.
- [31] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.
- [32] Y. Chen, H. Ma, J. Wang, J. Wu, X. Wu, and X. Xie, "PD-Net: Quantitative motor function evaluation for parkinson's disease via automated hand gesture analysis," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 2683–2691, doi: [10.1145/3447548.3467130](https://doi.org/10.1145/3447548.3467130).
- [33] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, pp. 50–60, 1947.