Artificial Intelligence In Medicine 110 (2020) 101966

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Supervised classification of bradykinesia in Parkinson's disease from smartphone videos

Stefan Williams [a,b], Samuel D. Relton [a], Hui Fang [c], Jane Alty [g], Rami Qahwaji [d], Christopher D. Graham [a,e], David C. Wong [f,*]

[a] Leeds Institute of Health Sciences, Univ. of Leeds, UK
[b] Leeds Teaching Hospital NHS Trust, UK
[c] Dept. of Computer Science, Loughborough University, UK
[d] School of Electronic Engineering and Computer Science, Univ. of Bradford, UK
[e] School of Psychology, Queen's University Belfast, UK
[f] Centre for Health Informatics, Univ. of Manchester, UK
[g] Wicking Dementia Research and Education Centre, University of Tasmania, Australia

ARTICLE INFO

Keywords:
Classification
Parkinson's
Bradykinesia
Video
Computer vision
Diagnosis
Support vector machine

ABSTRACT

Background: Slowness of movement, known as bradykinesia, is the core clinical sign of Parkinson's and fundamental to its diagnosis. Clinicians commonly assess bradykinesia by making a visual judgement of the patient tapping finger and thumb together repetitively. However, inter-rater agreement of expert assessments has been shown to be only moderate, at best.
Aim: We propose a low-cost, contactless system using smartphone videos to automatically determine the presence of bradykinesia.
Methods: We collected 70 videos of finger-tap assessments in a clinical setting (40 Parkinson's hands, 30 control hands). Two clinical experts in Parkinson's, blinded to the diagnosis, evaluated the videos to give a grade of bradykinesia severity between 0 and 4 using the Unified Pakinson's Disease Rating Scale (UPDRS). We developed a computer vision approach that identifies regions related to hand motion and extracts clinically-relevant features. Dimensionality reduction was undertaken using principal component analysis before input to classification models (Naïve Bayes, Logistic Regression, Support Vector Machine) to predict no/slight bradykinesia (UPDRS = 0–1) or mild/moderate/severe bradykinesia (UPDRS = 2–4), and presence or absence of Parkinson's diagnosis.
Results: A Support Vector Machine with radial basis function kernels predicted presence of mild/moderate/severe bradykinesia with an estimated test accuracy of 0.8. A Naïve Bayes model predicted the presence of Parkinson's disease with estimated test accuracy 0.67.
Conclusion: The method described here presents an approach for predicting bradykinesia from videos of finger-tapping tests. The method is robust to lighting conditions and camera positioning. On a set of pilot data, accuracy of bradykinesia prediction is comparable to that recorded by blinded human experts.

1. Introduction

Parkinson's disease is a neurodegenerative disorder that affects approximately 1 in 500 adults [1]. The diagnosis is a clinical one, based on the clinician detecting the presence of a slowness of movement termed bradykinesia, together with at least one of rigidity, rest tremor or postural instability (United Kingdom Parkinson's Disease Society Brain Bank Criteria) [2–4].

Clinician assessment of the presence and severity of bradykinesia is visual, and almost always includes an observation of finger tapping. In this test, a patient is asked to repetitively tap their forefinger against their thumb as wide and quickly as possible. The clinician will typically observe ten finger taps whilst looking for impairment of speed, amplitude or rhythm, often including a progressive 'decrement' seen over the duration of the test [4,5].

However, this visual clinical judgment is inherently subjective, and

* Corresponding author.
E-mail address: david.wong@manchester.ac.uk (D.C. Wong).

https://doi.org/10.1016/j.artmed.2020.101966
Received 17 January 2020; Received in revised form 3 September 2020; Accepted 2 October 2020
Available online 6 October 2020

there is no objective measure of bradykinesia in routine clinical use. Given both the imprecise definition of the term, and the difficulty for human observers to quantify small differences in movement, it is little surprise that inter-rater agreement of assessment of bradykinesia is moderate at best [4,5]. Current evidence suggests that human observers prioritise changes in movement amplitude over changes in tapping frequency or rhythm [4].

Given the fundamental importance of bradykinesia to diagnose and monitor Parkinson's, and the relatively small group of neurologists trained to assess it, an automatic and objective method of determining the level of bradykinesia has the potential to improve early diagnosis and to standardise follow-up assessment, including home monitoring.

Other approaches have previously been suggested for objective bradykinesia assessment [6–9]. However, all require either sensors that may not be readily available, or patient interaction with a specific computer program or smartphone app. To our knowledge, only one previous report used standard video to measure finger tapping bradykinesia, but featured only participants with advanced stage Parkinson's and required video recording of the face [10]. Here we propose a solution that uses the ubiquitous smartphone video camera to capture the relevant data during standard clinical assessment of finger tapping.

Our primary aim is to provide proof-of-concept that the assessment of bradykinesia can be automated using simple camera input, negating the impact of inter-rater variability and providing easily accessible clinical decision support. We also investigate the potential to predict diagnosis of Parkinson's itself. We describe how the video signal is processed and how pertinent features may be extracted to predict both bradykinesia and the presence of a Parkinson's diagnosis. Finally, we present initial results from a case-control pilot study.[1]

2. Existing work

The standard clinical method to assess bradykinesia is a visual judgment of finger tapping made by an experienced clinician. The two main validated clinical rating scales for finger tapping are Item 3.4 of the Unified Parkinson's Disease Rating Scale (UPDRS) [11], and the Modified Bradykinesia Rating Scale (MBRS) [4]. The UPDRS amalgamates the judgment of finger tapping speed, amplitude, and rhythm into a single score, such that those three elements can contribute to the score as 'and/or' definitions (Table 1). The score ranges from 0 (normal) to 4 (severe). In contrast, the MBRS is comprised of three separate scores for speed, amplitude, and rhythm.

Table 1
Summary of the Movement Disorder Society revision to UPDRS Item 3.4 (Finger Tapping) rating scale [11].

Score	Description
0 – Normal	No problems.
1 – Slight	Any of the following: (a) regular rhythm broken with 1–2 interruptions, (b) slight slowing, (c) amplitude decreases towards end.
2 – Mild	Any of the following: (a) 3–5 interruptions, (b) mild slowing, (c) amplitude decreases midway.
3 – Moderate	Any of the following: (a) 6+ interruptions or long freeze in movement, (b) moderate slowing, (c) amplitude decreases from start.
4 – Severe	Cannot perform the task due to slowing, interruptions, or decrements.

A variety of devices have been studied as methods to objectively measure bradykinesia during finger tapping. These include: contact sensors (e.g. MIDI keyboards or smartphone screens) [13–16]; accelerometers or gyroscopes attached to the index finger [4,17,18]; electromagnetic systems with magnetic generation and detection coils placed on finger and thumb [19,9,7]; infrared cameras with passive or active markers on the hand [20].

Example measures of finger tapping derived from such devices include opening velocity (speed) [4,20], excursion angle (amplitude) [4,18,21], and coefficient of variation (rhythm) [4,13,17]. Most metrics used show significantly different mean values in Parkinson's compared with control groups across several studies, albeit with considerable overlap of the group scores.

Multiple reports show that tapping measures correlate with clinical rating scales. For example, good correlation has been shown for gyroscope angular velocity with UPDRS (Spearman correlation coefficient: − 0.78) [18], variation in duration of keyboard taps with UPDRS (Pearson: − 0.61) [13], and gyroscope excursion angle with the amplitude component of MBRS (Pearson: − 0.81) [4]. Several studies of finger tapping measurement show AUROC for patient/control discimination in the range of 0.7–0.9. For example, 0.88 for dwelling time with smartphone tapping [14], 0.75 for inter-peak interval using accelerometer [7], 0.81 for opening velocity and 0.87 for amplitude decrement using infrared [20]. In a 'clinician v.s. machine' trial, a gyroscope system showed better intraclass correlation and minimal detectable change compared with clinician (MBRS) ratings during adjustment of deep brain stimulation treatment strength [22].

There is variation across studies in terms of which specific aspect of tapping measurement (speed/amplitude/rhythm) shows the largest group differences or is most strongly correlated with clinical categories. There is no clear pattern of results or methods to explain this variation, except that finger tap frequency alone is often not predictive [17,23,24] and protocols in which patients are temporarily 'off' medication likely make it easier to find differences [4,21], but are less relevant to clinical practice.

Multiple tests using non-camera sensors in smartphones can be combined [25–27], and previous reports suggest that application of machine learning techniques to such data can discriminate patients from controls (96% sensitivity with random forests [25]) while the combined data correlates strongly with clinical ratings [26]. However, all such approaches require patients to independently interact with the app, usually for a prolonged period of time, more than once per day (e.g. minimum of twice per day in reference [26]). In our view, the vast majority of patients lack sufficient motivation for this, which could possibly explain why no such apps have entered routine clinical practice. In contrast, camera-based computer vision can simply observe existing clinical examination, and augment or assist clinical judgement, without a requirement for patient motivation to regularly use an app.

To our knowledge, only one previous study used computer vision with simple video to detect bradykinesia on finger tapping, by tracking finger motion [10]. A feature of tapping rhythm, 'cross-correlation between the normalised peaks', showed a strong Guttman correlation of − 0.8 with UPDRS, and a support vector machine with multiple tapping features distinguished between patients and controls with an accuracy of 95%. However, only 13 participants were recorded and all were described as having "advanced" Parkinson's: a disease stage at which diagnosis is rarely an issue. Furthermore, they required video of the patient's face (to approximate hand length) which could be considered intrusive in practice.

3. Method

3.1. Data collection (video recording and clinician rating)

The study was approved by the UK Health Research Authority (IRAS

[1] This work is an extended version of the conference paper presented at IEEE CBMS 2019 [12].

no. 224848). Patients with Parkinson's disease, previously diagnosed by a consultant neurologist at Leeds Teaching Hospitals NHS Trust, were invited to attend a research clinic appointment. All patients were in the 'on' motor state, by which we mean that: (i) patients reported that they felt 'on' – a widely accepted and understood term that patient's use to describe an overall sense that they feel their medications are working and they have reduced symptoms of Parkinson's [2], (ii) the neurologist reported the patient looked 'on' – a clinically accepted term for recognising a response to medications, and (iii) no medication had been withheld prior to recording.

Control participants were invited from the companions of participants, or from hospital staff. Control participants did not have any neurological diagnosis or take any medication that could cause Parkinsonism, tremor, bradykinesia or other movement impairment.

Each hand was filmed tapping forefinger and thumb 'as quick and as big as possible' for 15 s. This convenience sample comprised 40 patient hands and 30 controls hands (20 patient participants and 15 control participants).

The recordings were made using an integrated smartphone camera (iPhone SE), set to 60 frames per second, 1920 × 1080 pixels, and placed on a tripod, with only ambient lighting. Participants were asked to rest their elbow on a chair arm during the finger tapping and only the hand/forearm was filmed (no identifiable patient details were filmed). The distance from camera to hand was not tightly defined; in practice the camera was positioned at approximately 1m from the participant. The lateral (thumb) surface of the hand faced the camera. There were no specific instructions for the position of digits 3–5.

The degree of bradykinesia in each video was independently rated by two consultant neurologists with a special interest in Parkinson's, according to the section 3.4 of the UPDRS scale (UPDRS-FT) (Table 1) [11]. The raters were blinded to patient/control group.

For both groups, the correlation between UPDRS-FT scores from the right and left hand for an individual participant was very low (Patients $k = 0.17$, 95%CI: -0.18 to 0.47, Controls $k = 0.18$, 95%CI: -0.07 to 0.41). Consequently, we treated videos from each hand as independent samples.

3.2. *Data analysis*

3.2.1. *Data processing*

A schematic of the data processing framework is presented in Fig. 1.

Initially, the video frames were segmented to pixels corresponding to a participant's hand. Traditional skin color methods were unsuitable, given the uncontrolled lighting conditions used. Instead, the hand regions of interest were first detected using a convolutional neural network, originally proposed by Bambach et al. [29]. The detector is based on a MobileNet-V2 mode architecture and the single shot multi-box approach using the TensorFlow Object Detection API [30,31]. This architecture uses depth-wise separable convolutions to reduce computer overhead for mobile devices. We trained our model using manual annotation of 500 randomly selected frames from our dataset.

The output of the model was refined using a secondary pixel-level segmentation to remove erroneous background pixels. We used the grabcut method [32], which iteratively updates two Gaussian Mixture Models representing the background and foreground. We set two mixture components to model the foreground colors and 3 mixture components for the background colors.

The segmented frames were then converted into an optical flow field [33]. In such a field, each position corresponds to the vector pixel movement of a point object between two sequential frames. The magnitude of the vector thus represents the instantaneous speed of a point (in pixels/frame). We sum the magnitude at each point in the region of interest to obtain a metric of overall hand movement.

Optical flow magnitude is affected by camera distance and hand size (as well as actual movement), so to convert optical flow magnitude into true hand velocity, we scale the magnitude by the number of pixels in the hand region of interest, such that our metric M_t is:

$$M_t = \frac{\sum_j^H \sum_i^W b_{\mathrm{ij}} \sqrt{u_{\mathrm{ij}}^2 + v_{\mathrm{ij}}^2}}{\sum_j^H \sum_i^W b_{\mathrm{ij}}}, \tag{1}$$

where H and W are the height and width of the optical flow field, u and v are the horizontal and vertical components of the flow, and b is the pixel mask obtained from the image segmentation. By evaluating M_t over a sequence of video frames we produce a 1D signal over time. Examples of the signal are shown in Fig. 2.

3.2.2. *Feature extraction*

Candidate features were derived from the 1D signal via clinical knowledge and visual inspection. In particular, we derived a set of features that described the frequency, amplitude, and tap-to-tap variability, to reflect the UPDRS assessment criteria as follows.

Frequency: **Tapping frequency** was estimated as the frequency corresponding to the maximal amplitude peak in the fast Fourier transform (FFT) spectrum. This assumes that the finger tapping motion corresponds to the greatest movement (and thus energy) between frames and that other movements, such as finger tremor, have smaller magnitude.

Amplitude: **Energy spectral density** was calculated as the squared integral of the FFT spectrum, a measure that would be expected to increase with the amplitude of tapping. In addition, we assumed that bradykinesia movement is distinctive in some frequency bands. Therefore the energy spectral density is separated into six non-overlapping equal frequency bands ranging from 0 to 18.36 Hz with bandwidth interval 3.06 Hz. The upper frequency threshold was selected heuristically to avoid having multiple uninformative zero-energy frequency bins. The threshold represents the frequency up to which, on average, 99% of the signal energy is contained.

Variability: Two variability features were derived using the peaks of

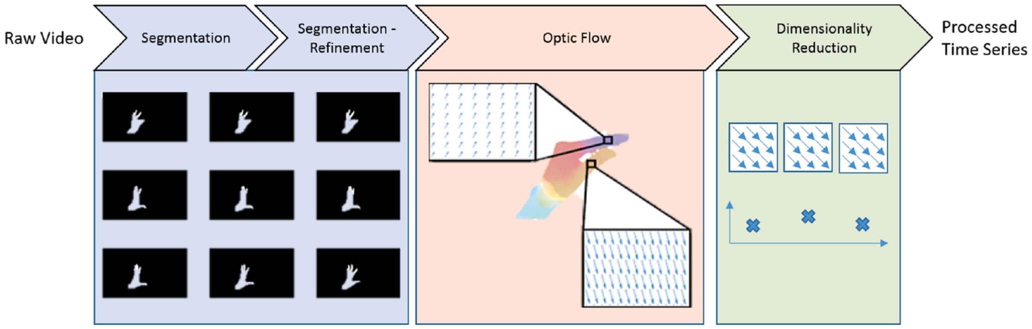

Fig. 1. Illustration of the data processing in which raw video is converted to an anonymous 1D time series. Raw video is first segmented using a convolutional neural network. The segmentation is refined using the grabcut method. Frame-by-frame movement of the hand is extracted using optical flow. The optical flow field is then reduced so that the magnitude of movement between two frames is summarised by a single value.

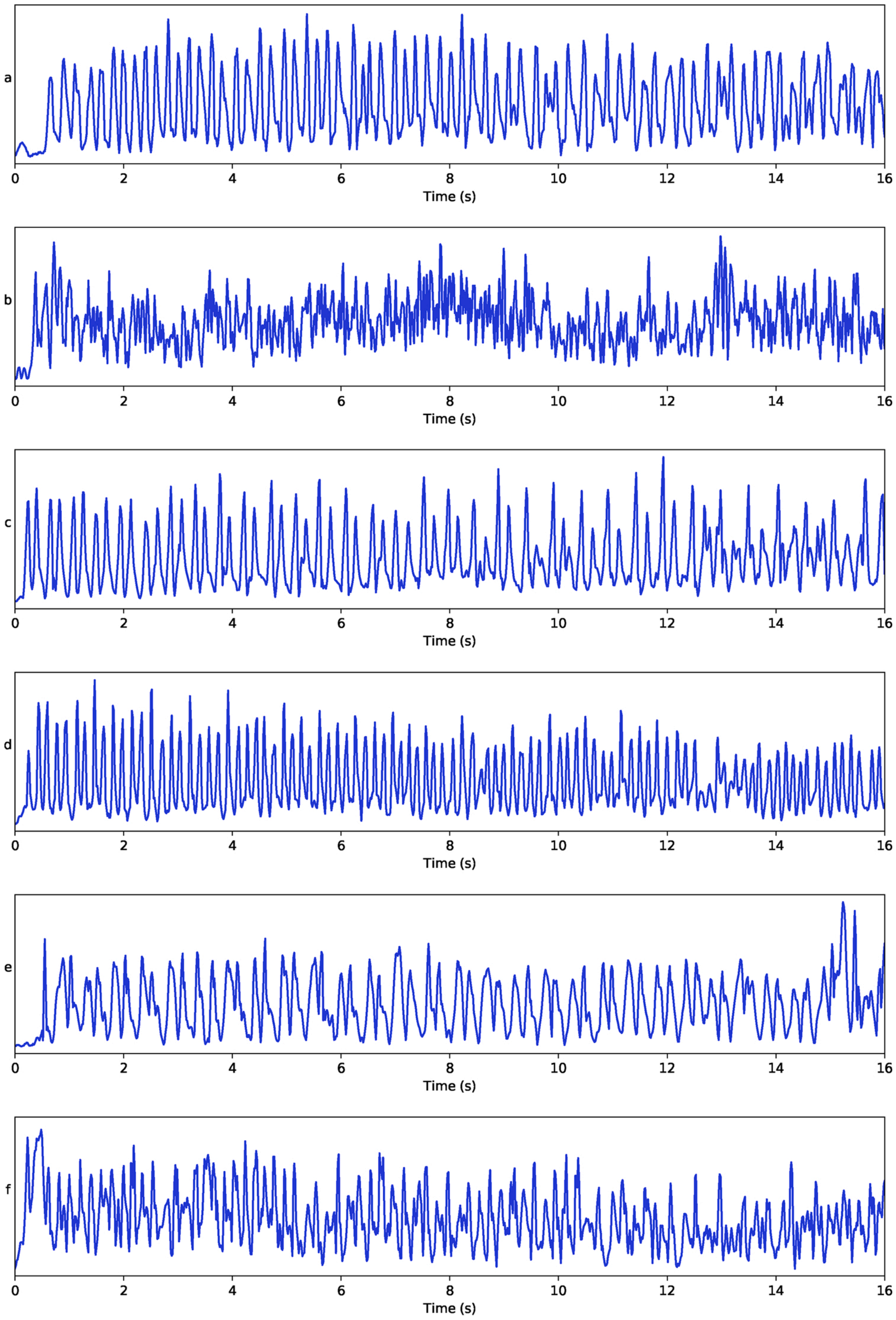

Fig. 2. Examples of the optical flow magnitude time series, plots (c)-(f) are discussed in Section 4.4. (a) – no bradykinesia (UPDRS-FT = 0). (b) – severe bradykinesia (UPDRS-FT = 4). (c) – UPDRS-FT = 0–1 misclassified as UPDRS-FT = 2–4, close to decision boundary. (d) – UPDRS-FT = 2–4 misclassified as UPDRS-FT = 0–1, close to decision boundary. (e) – UPDRS-FT = 0–1 misclassified as UPDRS-FT = 2–4, far from decision boundary. (f) – UPDRS-FT = 2–4 misclassified as UPDRS-FT = 0–1, far from decision boundary.

the optical flow waveform. Peaks were calculated via the MATLAB function *findpeaks* with zero minimum peak prominence. Peaks were then classified as *maxima* or *minima* by fitting a 1D Gaussian mixture model with two clusters to the peak amplitude values. We then defined:

Jitter: We hypothesise that there are differences between the hand closing and hand opening motions. From visual inspection, we observed differences in higher frequency movement between the signal *maxima* and *minima* – troughs in the signal appeared more jittery than the peaks. To quantify the jitter we include the ratio of number of *maxima* to number of *minima* over the entire time series as a predictor.

Peak-to-peak variability: was calculated as the standard deviation of the time between *maxima* peaks. This feature models variation in tapping frequency across the time series and may be considered analogous to the standard deviation of RR intervals (SDRR) for ECG signals [34].

3.3. Classification

We performed binary classification using Naïve Bayes (NB), logistic regression (LR), and both linear and RBF-based support vector machines (SVM-L and SVM-R, respectively) [35] to predict two outcomes: (1) a UPDRS-FT score > 1, and (2) clinical diagnosis of Parkinson's disease (previous clinical diagnosis by a consultant neurologist). Where there was disagreement in rater UPDRS-FT scores, the higher score was selected for training of the models.

Given the relatively small number of samples in the dataset we begin by reducing the feature space into two dimensions using principal component analysis. Indeed, preliminary work fitting models with all 10 features led to significant overfitting. We then explore the effect of analyzing up to 5 principal components, to look for any additional gain in accuracy.

The NB model was chosen as a simple baseline classifier providing a sensible lower bound for performance.

LR provides a linear separation of the data points and this simplicity may lead to lower generalisation error. We incorporated ridge ($\mathcal{L}_2$) regularisation with strength determined via a grid search of 100 log-spaced values in the interval [1e − 4, 1e+4] to minimise 10-fold cross-validation accuracy loss.

The SVM-L model optimises a different cost function than the LR model and therefore gives a different linear separation of the classes. Meanwhile, the SVM-R model has the ability to model nonlinear decision boundaries. The slack and (for SVM-R) kernel scaling hyperparameters were again estimated using a grid search to minimise 10-fold cross-validation accuracy loss. The grid search consisted of 100 log-spaced values in the intervals [1e+0, 1e+3] and [1e+0, 1e+5], respectively.

We report the training accuracy and AUC score for each model with two principal components, and for 3–5 components. We used permutation tests ($\alpha = 0.05$) on the varient obtaining highest accuracy to assess whether classifiers had meaningful predictive ability [36].

Due to the relatively small size of our pilot data we estimate the out-of-sample test accuracy, sensitivity, and specificity of each model by reporting the mean value of leave-one-out cross-validation (LOO-CV). Hyperparameters were preset according to outputs of the 10-fold cross-validation procedure described above.

We also investigate the contribution of each feature to the principal component analysis, to investigate the most discriminative features of the timeseries and compare with other research on this topic.

Finally, a visual inspection of the raw videos underlying the timeseries that were misclassified by the model with highest LOOCV accuracy was performed by two neurology clinicians (SW, JA). Analyses were performed using MATLAB 2017b and the scikit-learn and TensorFlow packages for Python 3 [37,38].

4. Results

A total of 70 videos were collected from 35 participants (left and right hands), Characteristics of the participants are presented in Table 2. 40 videos corresponded to the hands of participants with an established clinical diagnosis of Parkinson's. UPDRS-FT scores from 0 to 4 were assigned by two expert clinicians and then categorised into our binary outcome: UPDRS-FT ≤ 1 (no/slight bradykinesia) and UPDRS-FT > 1 (mild/moderate/severe bradykinesia). Their assessment matched in 73% of cases (κ = 0.46). In Fig. 2 we show an example of UPDRS-FT = 0 and UPDRS-FT = 4 for comparison.

4.1. Two principal components

The performance of each model for the prediction of UPDRS-FT category is shown in Table 3. The SVM-R model achieved the highest scores in all of our metrics. The other three models perform quite similarly, reflecting the fact that their decision boundaries are close to one another (see Fig. 3). The test accuracy (estimated using LOO-CV) drops to 0.8 for the SVM-R model, with the other models similarly dropping a few points of accuracy.

In Fig. 3 we show each time series plotted in feature-space after dimensionality reduction, marked according to category. We also show the decision boundaries of each method: an unbroken line for NB, dashed for SVM-R, dash-dotted for SVM-L, and dotted for LR.

Table 2
Study participant characteristics split by Parkinson's patients and control hands. The modified Hoehn and Yahr (H&Y) is a brief overall clinical rating to describe the stage of symptom progression in Parkinson's (higher number represents more advanced disease). UPDRS-FT refers to the Unified Parkinson's Disease Rating Scale Item 3.4 (Finger Tapping). Where raters disagreed the highest of the two UPDRS-FT scores was used.

	Patients	Controls
Age (Std. Dev.) yrs	67 (10.1)	66 (12.2)
Male/female	26 / 14	12 / 18
Median years since diagnosis	4	–
Median H&Y [IQR]	2 [1, 2.5]	–
H&Y = 1	9	–
H&Y = 1.5	0	–
H&Y = 2	5	–
H&Y = 2.5	1	–
H&Y = 3	4	–
H&Y = 4	1	–
H&Y = 5	0	–
Median UPDRS-FT [IQR]	2 [1, 3]	1 [0, 1]
UPDRS-FT = 0	2	8
UPDRS-FT = 1	11	13
UPDRS-FT = 2	17	7
UPDRS-FT = 3	7	2
UPDRS-FT = 4	3	0

Table 3
Results for each model when predicting whether UPDRS-FT > 1 using two principal components. *Accuracy* and *AUC* are estimated from the training 10-fold cross validation and may be considered as upper-bounds. The test accuracy, sensitivity and specificity are estimated using LOO-CV. The emboldened values are the best result for each metric.

Method	Accuracy	AUC	Test Acc	Test Sens	Test Spec
NB	0.74	0.74	0.70	0.67	0.70
LR	0.73	0.73	0.69	0.72	0.65
SVM-L	0.71	0.71	0.71	0.72	0.71
SVM-R	**0.84**	**0.84**	**0.80**	**0.86**	**0.74**

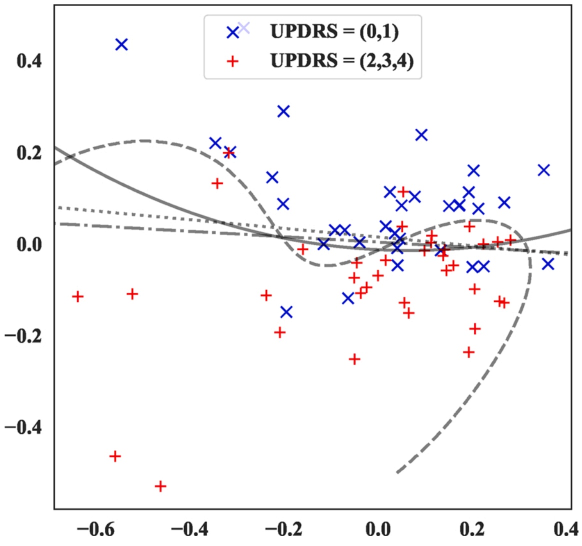

Fig. 3. Decision boundaries for prediction of UPDRS-FT > 1 using two principal components. The unbroken line is for NB, dashed for SVM-R, dash-dotted for SVM-L, and dotted for LR.

Our second task was the prediction of Parkinson's disease diagnosis itself based upon these features. The performance of each model for this task is shown in Table 4. Both the NB and SVM-R methods had very similar performance in terms of accuracy and AUC – with NB having

Table 4
Results for each model when predicting Parkinson's diagnosis using two principal components. *Accuracy* and *AUC* are estimated from the training 10-fold cross validation may be considered as upper-bounds. The test accuracy, sensitivity and specificity are estimated using LOO-CV. The emboldened values highlight the best result for each metric.

Method	Accuracy	AUC	Test Acc	Test Sens	Test Spec
NB	**0.69**	**0.70**	**0.64**	0.58	**0.73**
LR	0.61	0.59	0.61	**0.78**	0.40
SVML-L	0.63	0.60	0.60	**0.78**	0.40
SVM-R	**0.69**	0.68	0.63	0.68	0.57

better specificity but SVM-R having better sensitivity. Neither LR or SVM-L were competitive for this task unless high sensitivity is desired.

A plot of the time series in feature-space, colored by category, and the decision boundary of each method is displayed in Fig. 4.

4.2. Additional principal components

In addition, we experimented with adding additional principal components into the models, training them using the same cross-validation procedure as above.

Tables 5 and 6 show the resulting accuracy and AUC scores when additional principal components are added, with the test accuracy estimated using LOO-CV.

The NB model shows improvement in both accuracy and AUC as more components are included. This did not translate into improved test accuracy, probably indicating model overfitting. The LR and SVM-L models showed minor improvements which again fail to translate into improved test accuracy. The non-monotonic gains in accuracy may be due to the effect of the bias-variance trade-off in this dataset. Overall the SVM-R model with two principal components performs had the highest metrics; additional components degrade its accuracy due to the bias-variance trade-off.

The resulting accuracy and AUC when predicting a Parkinson's diagnosis with additional principal components is shown in Tables 7 and 8.

All models except for LR benefited from additional components in terms of in-sample performance but none of these gains translate into improvements in estimated test accuracy. Non-monotonicity in performance as the number of components grows implies there may be some

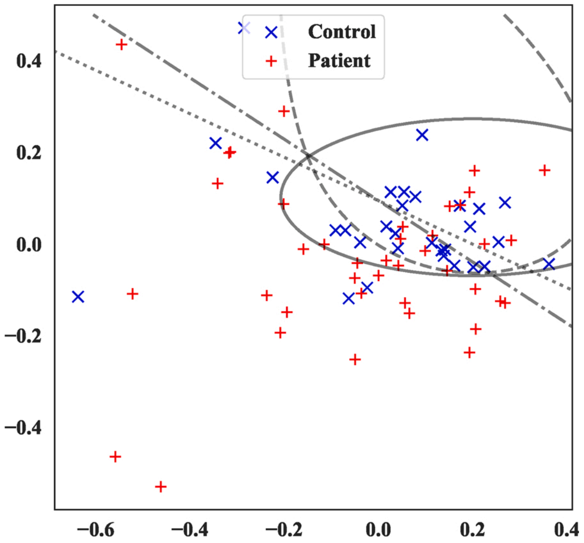

Fig. 4. Decision boundaries for prediction of Parkinson's diagnosis using two principal components. The unbroken line is for NB, dashed for SVM-R, dash-dotted for SVM-L, and dotted for LR.

Table 5
Resulting accuracy when predicting UPDRS-FT > 1 using 2–5 principal components. The emboldened values highlight the model with highest accuracy in each case The best performing number of principal components was used to estimate the test accuracy with LOO-CV and the permutation test *P*-value.

Method	PCA-2	PCA-3	PCA-4	PCA-5	Test Acc.	P
NB	0.74	0.76	0.79	**0.81**	0.73	0.02
LR	0.73	0.74	**0.76**	0.74	0.73	0.01
SVM-L	0.71	**0.79**	0.76	0.79	0.71	0.01
SVM-R	**0.84**	0.84	0.83	0.81	0.8	0.01

Table 6
Resulting AUC when predicting UPDRS-FT > 1 using 2–5 principal components. The emboldened values highlight the model with highest AUC in each case.

Method	PCA-2	PCA-3	PCA-4	PCA-5
NB	0.74	0.76	0.79	**0.81**
LR	0.73	0.74	**0.76**	0.74
SVM-L	0.71	**0.79**	0.76	0.79
SVM-R	**0.84**	0.84	0.83	0.81

Table 7
Resulting accuracy when predicting Parkinson's diagnosis using 2–5 principal components. The emboldened values highlight the model with highest AUC in each case. The best performing number of principal components was used to estimate the test accuracy with LOO-CV and the permutation test *P*-value.

Method	PCA-2	PCA-3	PCA-4	PCA-5	Test Acc.	P
NB	0.69	0.63	0.67	**0.74**	0.67	0.46
LR	**0.61**	0.57	0.6	0.57	0.57	0.57
SVM-L	0.63	0.66	0.67	**0.7**	0.63	0.97
SVM-R	0.69	0.8	**0.8**	0.76	0.66	0.49

Table 8
Resulting AUC when predicting Parkinson's diagnosis using 2–5 principal components. The emboldened values highlight the model with highest AUC in each case.

Method	PCA-2	PCA-3	PCA-4	PCA-5
NB	0.69	0.64	0.69	**0.74**
LR	**0.59**	0.5	0.57	0.57
SVM-L	0.6	0.65	0.67	**0.68**
SVM-R	0.68	**0.81**	0.79	0.75

Table 9
Contribution of each feature to the first 5 principal components in percentages. The column names SV-n denote contributions to the nth singular vector. ESD is short-hand for Energy Spectral Density.

Feature	SV-1	SV-2	SV-3	SV-4	SV-5
Max peak (Hz)	9.5	14.8	1.3	12.1	9.7
Total ESD	3	16.5	4.5	18.7	19.6
ESD (0–3.06 Hz)	8.3	10.7	23.2	6	4
ESD (3.06–6.12 Hz)	8.3	4	27.8	1.8	8.9
ESD (6.12–9.18 Hz)	12.7	10.8	11.3	8.6	4.6
ESD (9.18–12.24 Hz)	13.2	8	7	10.3	2.9
ESD (12.24–15.3 Hz)	11.8	10.5	4.7	4.1	12.2
ESD (15.3–18.36 Hz)	12	11.6	9.8	9.6	5.9
Maxima-minima ratio	6.4	13	8.6	20.9	20.2
Peak-to-peak std. dev.	14.9	0.1	1.8	8	12.1
Variance explained	37.5	24.1	15.3	7.6	5.5
Cumulative variance	37.5	61.6	76.9	84.5	90

effect from the bias-variance trade-off.

4.3. Feature contribution to PCA

Table 9 lists the percentage contribution of all of derived features to the first 5 principal components, along with the variance explained by each components.

Our first component explaining 37.5% of the overall variance is comprised primarily of the peak-to-peak standard deviation – measuring variability in rhythm throughout the timeseries – and the energy spectral density (ESD) in higher frequency bands, which measure jittery movement.

The second component included strong influence from the frequency of the maximal peak (measuring rhythm), the total power in the signal (corresponding to average amplitude across the time series), and the maxima-minima ratio (corresponding to jitter in hand motions).

4.4. Misclassified UPDRS-FT categories

We investigated the misclassified examples when predicting UPDRS-FT category using our the SVM-R with two principal components model, to glean insight into where our models may be improved.

This model misclassified 11 examples. 7 were misclassified as mild/moderate/severe bradykinesia (UPDRS-FT > 1) (5 controls, 2 patients). Meanwhile, 4 were misclassified as no/slight bradykinesia (UPDRS-FT 0–1) (1 control, 3 patients). The misclassified examples were close to the decision boundary in 4 cases; for these cases there was expert rater disagreement. All misclassified videos had a UPDRS-FT grade of either 1 or 2, i.e. no large misclassifications occurred.

The time series of two of the examples closest to the decision boundary (one patient and control) are shown in Fig. 2. The two misclassified cases furthest from the decision boundary (one patient and control) are also shown in Fig. 2.

Re-examination of the original videos and optical flow timeseries by two neurologists (SW, JA) identified several potential contributors to this misclassification. First, several videos showed overall hand movement while fingers were held closed between taps, usually a swinging wrist movement preparing for the next tap (and in one case tremor). This created additional small peaks and a more irregular timeseries in videos that showed otherwise regular, smooth finger tapping. Conversely, moving all the fingers ‘en masse’ tended to create large smooth peaks of optical flow, that reduced the optical flow effect of underlying irregularities in the tapping itself.

Second, a large difference between the speed of finger opening (slower) and closing (quicker) created two distinct optical flow peak sizes/shapes, and a less uniform timeseries, even though the actual tapping was not clearly bradykinetic by UPDRS-FT. Third, our timeseries have a 15s duration, similar to several other objective measures, e.g. [41], but the UPDRS-FT asks raters to judge only the first 10 finger taps. When tapping rate is fast, only a small initial section of the time series is judged by raters, while later tapping changes contribute to the optical flow timeseries.

Finally, it is known that raters prioritise amplitude and rhythm when judging finger tapping, but pay less attention to speed [4]. With this is mind, we noted that slow but large amplitude movements tended to be classified as UPDRS-FT 0–1 by raters, but UPDRS-FT > 1 by SVM-R, whereas fast but smaller amplitude movements tended to be classified as UPDRS-FT > 1 by raters, but UPDRS-FT 0–1 by SVM-R.

5. Discussion

In a pilot sample of 70 finger-tapping test videos, we showed reasonable predictive performance for mild/moderate/severe bradykinesia (UPDRS-FT > 1). The estimated test accuracy of 0.8 (using SVM-R) is promising in light of the level of agreement between expert clinical raters ($\kappa = 0.46$). We also note that disagreement between the automated method and clinical experts may be caused when either (i) the clinician is correct and the automated test is wrong, or (ii) the clinician is incorrect and the automated test is right. Given that prior literature casts doubt on the ability of human experts to accurately evaluate subtle traits [4,39], (ii) is highly feasible; such that the reported accuracy may underestimate how well we truly classify bradykinesia. Further improvements in accuracy and generalisability may be achieved by using classification algorithms that account for uncertain labels, such as probabilistic SVM [40]. However, in our case with only two raters, such approaches may still be fragile, as an individual rater will have a large effect on the probabilistic labels.

The method was less successful at predicting the presence of Parkinson’s diagnosis: NB obtained an estimated test accuracy of 0.67 using 5 principal components. In fact, for all classifiers, the p-values from the permutation test indicate that similar accuracies may be obtained by chance. While this does not invalidate the result, a much larger training sample is required to determine whether the classifiers are learning true structure in the data.

This poorer performance is to be expected. A degree of bradykinesia is often detected in control hands when clinical raters are blinded to diagnosis status, particularly among older age groups [20]. While bradykinesia is a necessary component of the Parkinson’s diagnostic criteria, it is not sufficient in isolation [3]. In practice, finger tapping bradykinesia is only one of a more comprehensive set of clinical assessments used to diagnose Parkinson’s.

The clinician ratings were based on 10 finger taps, as per UPDRS, whereas the optical flow time series was 15 s duration, similar to some existing studies [41]. Some misclassification may have resulted from this difference in assessment time period. Future work could isolate individual tapping epochs [42]. Future work to separate overall hand movement from finger-thumb tapping might also improve classification.

Our novel approach to finger tap measurement cannot be easily compared with previous literature for several reasons. First, previous studies use clinically recognisable features (e.g. tap distance) rather than overall optical flow, but they require special equipment or patient interaction with an app. Second, the results of previous studies vary widely in terms of strength of correlation or accuracy of discrimination, despite apparently similar methods [18,28,43,44]. Finally, in contrast to our work, many previous studies involve measurements after patients have been instructed to withhold medication, artificially creating more severe bradykinesia and thus larger differences [4,21]. With these caveats, our accuracy of 0.8 is broadly comparable to previous work.

The single previous computer vision video study involved a small sample of 13 Parkinson’s patients, who all had advanced disease [10]. We note that their most predictive feature for UPDRS was a measure of tapping rhythm. This corresponds to our results in which the first principal component feature was primarily composed of a rhythm measure (peak to peak variation). Other studies also suggest rhythm measures may be particularly important [4,19].

The approach used here has potential to provide widely available, low-cost bradykinesia detection; without the requirement for new hardware or for patients to directly interact with smartphone apps or computer programs. This is a fundamental difference from previously published methods [4,8]. An automated method broadens access to the measurement of bradykinesia (currently the preserve of a small group of clinicians, principally neurologists). For example, allowing family doctors and medical nurse practitioners to screen for and monitor the phenomenon has potential resource benefits. Furthermore, the use of ubiquitous technology means that the approach may be suitable in a home setting to monitor progression of Parkinson’s. In addition, it might also be useful for monitoring other conditions in which there are changes in movement over time such as rheumatoid arthritis, in which common signs include decreased range of motion and joint stiffness [45,46].

Whilst initial results appear promising, our estimate of accuracy may be optimistic, as our small sample size meant that there was insufficient data to test on an independent subset of data. In addition, the small sample size means that classification using LR, SVMs, and NB produced conservative decision boundaries. A large sample would allow us to determine whether there was any true local structure in the feature space. A larger sample would also allow us to improve the usefulness of the system by estimating the UPDRS score directly, rather than the binary categorisation undertaken here. A larger validation study is therefore necessary and has been initiated by the study team.

In addition, the continuum of finger tapping performance means that in reality there is a soft boundary between UPDRS-FT grade 1 and grade 2, but the use of a binary classifier (e.g. SVM) creates a harder boundary between these classifications, contributing to errors. In future work, we can investigate 'fuzzy' or multi-class neural networks to address this.

Furthermore, the approach taken here is likely sub-optimal in two respects. First, spatial and angular information is discarded at each frame. This has the advantage of reducing the dimensionality of the signal so that real-time processing, even on modest hardware, is practicable. Second, the hand-selection of candidate features was entirely subjective and may have missed important characteristics in the time series. Additional data would allow more sophisticated approaches to automatically learn pertinent features (c.f. [47]).

Finally, it is possible that we may introduce bias by analyzing data on a per-hand, rather than per-patient basis. We do not believe that this was an important factor for the analysis presented here. In supplementary material, we further describe the expert-rated UPDRS of left and right hands of the control and patient population, showing no evidence of systematic difference between hands. We also performed a sensitivity analysis in which the 'partnering' hand was omitted from Leave One Out Cross Validation training, in which the results remained consistent.

6. Conclusion and future work

We have described and demonstrated an automated method to classify the presence of bradykinesia via smartphone video signals. In our pilot study we have shown good agreement with expert clinicians. Further improvements may be possible via more sophisticated analyses, but this requires further training data. A larger validation study of this technology is currently under development.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank Dr. Jeremy Cosgrove for providing clinical ratings of the videos. We also thank Parisa Patel and Paschal O'Gorman for contributing to video capture.

References

[1] Parkinson's UK. The incidence and prevalence of Parkinson's in the UK. London, UK; 2018.

[2] Lees A. Parkinson's disease. Pract Neurol 2010;10(4):240–6.

[3] Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease. A clinico-pathological study of 100 cases. JNNP 1992;55(3):181–4.

[4] Heldman D, Guiuffrida JP, Chen R, Payne M, Mazzella F, Duker AP, et al. The modified bradykinesia rating scale for Parkinson's disease: reliability and comparison with kinematic measures. Mov Disord 2011;26(10):1859–63.

[5] Martinez-Martin P, Gil-Nagel A, Morlan Gracia L, Balserio Gomez J, Martines-Sarries J, Bermejo F. Unified Parkinson's disease rating scale characteristics and structure. Mov Disord 1994;9(1):76–83.

[6] Martinez-Manzanera O, Roosma E, Beudel M, Borgemeester RW, van Laar T, Maurits NM. A method for automatic and objective scoring of bradykinesia using orientation sensors and classification algorithms. IEEE Trans Biomed Eng 2016;63 (5):1016–24.

[7] Gao C, Smith S, Lones M, Jamieson S, Alty J, Cosgrove J, et al. Objective assessment of bradykinesia in Parkinson's disease using evolutionary algorithms: clinical validation. Transl Neurodegen 2018;7(1):18.

[8] Hasan H, Athauda DS, Foltynie T, Noyce AJ. Technologies assessing limb bradykinesia in Parkinson's disease. J Parkinson's Dis 2017;7(1):65–77.

[9] Gao C, smith S, Lones M, Jamieson S, Alty J, Cosgrove J, et al. Objective assesssment of bradykinesia in Parkison's disease using evolutionary algorithms: clinical validation. Transl Neurodegen 2018;7:18.

[10] Khan T, Nyholm D, Westin J, Dougherty M. A computer vision framework for finger-tapping evaluation in Parkinson's disease. Artif Intell Med 2014;60(1): 27–40.

[11] Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Mov Disord 2008;23(15):2129–70.

[12] Wong DC, Relton SD, Fang H, Alty J, Qhajawi R, Graham CD, et al. Supervised classification of bradykinesia for Parkinson's disease diagnosis from smartphone videos. Proc IEEE int symp on computer-based med syst 2019.

[13] Taylor Tavares AL, Jefferis GSXE, Koop M, Hill BC, Hastie T, Heit G, et al. Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. Mov Disord 2005;20(10): 1286–98.

[14] Lee CY, Kang SJ, Hong S-K, Ma H-I, Lee U, Kim YJ. A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease. PLOS ONE 2016;11(7).

[15] Kassavetis P, Saifee TA, Roussos G, Drougkas L, Kojovic M, Rothwell JC, et al. Developing a tool for remote digital assessment of Parkinson's disease. Mov Disord Clin Pract 2016;3(1):59–64.

[16] Homann CN, Suppan K, Wenzel K, Giovannoni G, Ivanic G, Horner S, et al. The bradykinesia akinesia incoordination test (BRAIN TEST), an objective and user-friendly means to evaluate patients with Parkinsonism. Mov Disord 2000;15(4): 641–7.

[17] Costa J, González HA, Valldeoriola F, Gaig C, Tolosa E, Valls-Solé J. Nonlinear dynamic analysis of oscillatory repetitive movements in Parkinson's disease and essential tremor. Mov Disord 2010;25(15):2577–86.

[18] Kim J-W, Lee J-H, Kwon Y, Kim C-S, Eom G-M, Koh S-B, et al. Quantification of bradykinesia during clinical finger taps using a gyrosensor in patients with Parkinson's disease. Med Biol Eng Comput 2011;49(3):365–71.

[19] Sano Y, Kandori A, Shima K, Yamaguchi Y, Tsuji T, Noda M, et al. Quantifying Parkinson's disease finger-tapping severity by extracting and synthesizing finger motion properties. Med Biol Eng Comput 2016;54(6):953–65.

[20] Růžička E, Krupička R, Zárubová K, Rusz J, Jech R, Szabó Z. Tests of manual dexterity and speed in Parkinson's disease: not all measure the same. Parkinsonism Relat Disord 2016;28:118–23.

[21] Di Biase L, Summa S, Tosi J, Taffoni F, Marano M, Rizzo AC, et al. Quantitative analysis of bradykinesia and rigidity in Parkinson's disease. Front Neurol 2018;9 (MAR):1–12.

[22] Heldman DA, Espay AJ, LeWitt PA, Giuffrida JP. Clinician versus machine: reliability and responsiveness of motor endpoints in Parkinson's disease. Parkinsonism Relat Disord 2014;20(6):590–5.

[23] Maetzler W, Ellerbrock M, Heger T, Sass C, Berg D, Reilmann R. Digitomotography in Parkinson's disease: a cross-sectional and longitudinal study. PLOS ONE 2015;10 (4):e0123914.

[24] Lee MJ, Kim SL, Lyoo CH, Rinne JO, Lee MS. Impact of regional striatal dopaminergic function on kinematic parameters of Parkinson's disease. J Neural Transm 2015;122(5):669–77.

[25] Arora S, Venkataraman V, Zhan A, Donohue S, Biglan KM, Dorsey ER, et al. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study. Parkinsonism Relat Disord 2015;21(6):650–3.

[26] Zhan A, Mohan S, Tarolli C, Schneider RB, Adams JL, Sharma S, et al. Using smartphones and machine learning to quantify Parkinson disease severity, the mobile Parkinson disease score. JAMA Neurol 2018;75(7):876–80.

[27] Stamate C, Magoulas GD, Kppers S, Nomikou E, Daskalopoulos I, Luchini MU, et al. Deep learning Parkinson's from smartphone data. Proc IEEE Int conf on pervasive computing and communications 2017:31–40.

[28] Bank PJM, Marinus J, Meskers CGM, de Groot JH, van Hilten JJ. Optical hand tracking: a novel technique for the assessment of bradykinesia in Parkinson's disease. Mov Disord Clin Pract 2017;4(6):875–83.

[29] Bambach S, Lee S, Crandall DJ, Yu C. Lending a hand: detecting hands and recognizing activities in complex egocentric interactions. Proc IEEE int conf on computer vision 2015:1949–57.

[30] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: inverted residuals and linear bottlenecks. Proc IEEE conf on computer vision and pattern recognition 2018:4510–20.

[31] Huang J, Rathod V, Chow D, Sun C, Zhu M. TensorFlow object detection API. https://github.com/tensorflow/models/tree/master/research/object_detection.

[32] Rother C, Kolmogorov V, Blake A. Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Trans Graphics 2004;23(3):309–14.

[33] Horn BK, Schunck BG. Determining optical flow. Artif Intell 1981;17:185–203.

[34] Malik M, Bigger JT, Camm AJ, Kleiger RE, Malliani A, Moss AJ, et al. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Circulation 1996;93(5):1043–65.

[35] Orphanidou C, Wong DC. Machine learning models for multidimensional clinical data. Handbook of large-scale distributed computing in smart healthcare. Springer; 2017.

[36] Golland P, Fischl B. Permutation tests for classification: towards statistical significance in image-based studies. Biennial international conference on information processing in medical imaging 2003:330–41.

[37] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. OSDI 2016;16:265–83.

[38] MATLAB release 2017b. The MathWorks Inc. Massachusetts, USA.

[39] Bajaj NPS, Gontu V, Birchall J, Patterson J, Grosset DG, Lees AJ. Accuracy of clinical diagnosis in tremulous parkinsonian patients: a blinded video study. JNNP 2010;18:1223–8.

[40] Niaf E, Flamary R, Rouviere O, Lartizien C, Canu S. Kernel-based learning from both qualitative and quantitative labels: application to prostate cancer diagnosis based on multiparametric MR imaging. IEEE Trans Image Proc 2013;23(3):979–91.

[41] Bologna M, Leodori G, Stirpe P. Bradykinesia in early and advanced Parkinson's disease. J Neurol Sci 2016;369:286–91.

[42] Stamatakis J, Ambroise J, Cremers J, Sharei H, Delvaux V, Macq B, et al. Finger tapping clinimetric score prediction in Parkinson's disease using low-cost accelerometers. Comput Intell Neurosci 2013;717853. 13pp.

[43] Lones M, Smith S, Alty J, Lacy S, Possin K, Jamieson D, et al. Evolving classifiers to recognize the movement characteristics of Parkinson's disease patients. IEEE Trans Evol Comp 2014;18(4):559–76.

[44] Lones M, Smith S, Tyrrell A, Alty J, Jamieson D. Characterising neurological time series data using biologically motivated networks of coupled discrete maps. Biosystems 2013;112(2):94–101.

[45] Pani D, Barabino G, Dessi A, Tradori I, Piga M, Mathieu A, et al. A device for local or remote monitoring of hand rehabilitation sessions for rheumatic patients. IEEE J Trans Eng Health Med 2014;2:1–11.

[46] Connolly J, Condell J, O'Flynn B, Sanchez JT, Gardiner P. IMU sensor-based electronic goniometric glove for clinical finger movement analysis. IEEE Sens J 2018;18(3):1273–81.

[47] Andreotti F, Carr O, Pimentel MA, Mahdi A, De Vos M. Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG. Proc Comput Cardiol 2017;4.