Citation for the original published paper (version of record):

Sadikov, A., Groznik, V., Možina, M., Žabkar, J., Nyholm, D. et al. (2017)
Feasibility of spirography features for objective assessment of motor function in Parkinson's disease.
Artificial Intelligence in Medicine
https://doi.org/10.1016/j.artmed.2017.03.011

Feasibility of spirography features for objective assessment of motor function in Parkinson's disease

Aleksander Sadikov[1], Vida Groznik[1], Martin Možina[1], Jure Žabkar[1], Dag Nyholm[2], Mevludin Memedi[3,4], and Dejan Georgiev[5]

[1] University of Ljubljana, Faculty of Computer and Information Science,
Večna pot 113, Ljubljana, Slovenia,
`aleksander.sadikov@fri.uni-lj.si`,
[2] Department of Neuroscience, Neurology, Uppsala University, Uppsala, Sweden
[3] Computer Engineering, Dalarna University, Borlänge, Sweden
[4] Informatics, School of Business, Örebro University, Örebro, Sweden
[5] Ljubljana University Medical Centre, Department of Neurology, Zaloška 2, Ljubljana, Slovenia

Abstract. *Objective:* Parkinson's disease (PD) is currently incurable, however proper treatment can ease the symptoms and significantly improve the quality of life of patients. Since PD is a chronic disease, its efficient monitoring and management is very important. The objective of this paper was to investigate the feasibility of using the features and methodology of a spirography application, originally designed to detect early Parkinson's disease (PD) motoric symptoms, for automatically assessing motor symptoms of advanced PD patients experiencing motor fluctuations. More specifically, the aim was to *objectively* assess motor symptoms related to bradykinesias (slowness of movements occurring as a result of under-medication) and dyskinesias (involuntary movements occurring as a result of over-medication).
Materials and methods: This work combined spirography data and clinical assessments from a longitudinal clinical study in Sweden with the features and pre-processing methodology of a Slovenian spirography application. The study involved 65 advanced PD patients and over 30,000 spiral-drawing measurements over the course of three years. Machine learning methods were used to learn to predict the "cause" (bradykinesia or dyskinesia) of upper limb motor dysfunctions as assessed by a clinician who observed animated spirals in a web interface. The classification model was also tested for comprehensibility. For this purpose a visualisation technique was used to present visual clues to clinicians as to which parts of the spiral drawing (or its animation) are important for the given classification.
Results: Using the machine learning methods with feature descriptions and pre-processing from the Slovenian application resulted in 86% classification accuracy and over 0.90 AUC. The clinicians also rated the computer's visual explanations of its classifications as at least meaningful if not necessarily helpful in over 90% of the cases.
Conclusions: The relatively high classication accuracy and AUC demonstrates the usefulness of this approach for objective monitoring of PD patients. The positive evaluation of computer's explanations suggests the potential use of this methodology in a decision support setting.

Keywords: Parkinson's disease; movement disorder; spirography; spirography features; objective monitoring; visualisation

1 Introduction and problem statement

Parkinson's disease (PD) is a chronic neurological disorder associated with a number of motor and non-motor symptoms. Major motor symptoms include bradykinesia (slowness of movements), tremor and rigidity. While currently incurable, proper treatment can significantly ease the symptoms. As the disease progresses, however, patients start to experience motor fluctuations which adversely impact the quality of their life. Therefore, the treatment approaches should be individualised in order to alleviate unwanted symptoms which occur as a result of insufficient levels of medication (Off state) and abrupt, involuntary movements (also known as dyskinesias) which occur as a result of excessive levels of medication. To this end, careful and objective monitoring of the disease is of paramount importance. As the patients usually see the neurologist only few times per year (sometimes also just once per year), the neurologist often has only a very vague picture of their condition in-between the visits and has problems with prescribing the optimal drug dosage. Currently, patients monitor their long-term condition by keeping a simple diary. The entries in the diary, however, are subjective opinions of the patients and are not something measurable. Clinical scales like the Unified Parkinson's Disease Rating Scale (UPDRS) are not suitable for long-term and remote follow-up of the symptoms since they are relatively time-consuming [1], may need to be filled out at a clinical visit, require considerable clinical experience [2] and some of their items have poor inter-clinician reliability [3].

The digitalised spirography is a quantitative method for detection and evaluation of different types of tremors and other movement disorders. The spirographic system is usually composed of a computer with a specialised software, a graphic tablet or a touch-screen measurement device (e.g. a smartphone), and a stylus [14,4]. The patient is required to draw a spiral (or sometimes several spirals). The digitalised spirography enables us to store the exact timestamp of each point of the spiral in a two-dimensional area and thus provides an objective measurement of upper limb motor function. These systems allow extraction of detailed upper limb motor features from the spiral drawing tasks by analysing spatial and temporal artefacts of the spirals.

Our long-term goal is to develop a system for objective monitoring of PD patients that would also be able to automatically detect any significant changes in the patient's condition and report these changes to the neurologist. We envision the same methodology also used for automatic analysis of the patient's diary containing spirographic measurements, and as a decision support system for clinicians (automated spirographic test). To this end, the spirals drawn have to be described mathematically by various features (some described in Table 3 later in the paper) for the machine learning algorithms to be able to analyse them automatically. These features are also needed for visual explanation of the computer's reasoning.

This study presents a step towards our long-term goal. There were two main objectives of this study: (1) to investigate whether the methodology developed for the ParkinsonCheck application can be successfully used for the differentiation between states of bradykinesia (insufficient medication) and dyskinesia (overmedication), and (2) to create a visual aid for the physicians to use with spirography that also serves as an explanation of the computer's reasoning.

The background (previous work of both groups) is as follows. The group in Slovenia previously developed the PARKINSONCHECK application for early detection of signs of Parkinsonian or Essential tremor [11,12]. PARKINSONCHECK is a decision support system based on a spirographic test performed on smartphones. The application involves extensive preprocessing and mathematical descriptions of different features of the spirals drawn by the user. The group in Sweden performed a clinical trial to (also) test the suitability of spirography for objective monitoring of the patients. These were advanced PD patients, and one of the goals was to determine whether spirography can be used to associate the dosage of the medication with the current condition of the patient.

In the present study we applied the PARKINSONCHECK methodology to automatically differentiate between the states of bradykinesia and dyskinesia which is important for drug dosage adjustment, using spirography data collected in a Swedish study [5,6,17]. Note that this application is different from PARKINSONCHECK's objective. The successful application validates the PARKINSONCHECK's preprocessing and features for spiral description on a more general level as well as confirms the usefulness of digitalised spirography for this particular task. In the second part of the study we used the mathematical features with the visualisation methodology to create a decision support system for the physicians.

The work presented here consists of two parts, and the organisation of the paper follows this. In the first part, described in Section 2, we use machine learning to construct a decision model for differentiating bradykinesia from dyskinesia using the descriptive features and methodology developed for PARKINSONCHECK application. The data used and the differences with the data for which the features were originally constructed are also described in this section. The second part (Section 3) tests the comprehensibility and clinical usefulness of the decision model built in the first part. It describes an experiment using a visualisation technique to explain the "reasoning" of the decision model to the clinicians and their acceptance of the model. The results of both experiments (parts) are presented in their respective sections, while the joint discussion, conclusions, and further work are given in Sections 4 and 5.

2 Learning a classification model

2.1 Subjects

In this study, a retrospective analysis was conducted on spirography data of 65 patients with advanced idiopathic PD from eight different clinics in Sweden, recruited from January 2006 until August 2010

[6]. The patients were either treated with levodopa/carbidopa gel intestinal infusion or were candidates for receiving this treatment. There were 43 males and 22 females with median (± inter-quartile range) age of 65 ± 11 years and total UPDRS of 49 ± 20.5. UPDRS is a widely used scale for clinical assessment of Parkinson's disease and consists of four parts: I: mentation, behaviour and mood, II: activities of daily living, III: motor examination, and IV: complications of therapy. All questions have five response options 0–normal, 1–slight, 2–mild, 3–moderate, and 4–severe. The sum over the answers to the questions gives the UPDRS rating of a patient. [16]

2.2 Spirography data collection

During the course of the clinical study, the patients used a touch screen telemetry device in their home environments [5]. On each test occasion, they were asked to trace a pre-drawn Archimedean spiral using the dominant hand. The pre-drawn spiral was shown on the screen of the device and the patients were instructed to use an ergonomic pen stylus to trace it from the center and out, as accurately and fast as possible, supporting neither hand nor arm. The patients were instructed to place the device on a table and to be seated in a chair. The patients were asked to repeat the drawing three times per test occasion. The raw data consisted of x-y coordinates and time-stamps of the pen tip, digitized at a sample rate of 10 Hz. In total, the database consisted of 10,272 test occasions having approximately 30,816 (3 × 10,272) spirals. Some example spiral drawings are shown in Figure 1.

2.3 Clinical assessment of motor impairments

A clinician (D.N.) used a web interface that animated the spiral drawings, allowing him to observe different kinematic features during the drawing process and to rate task performance of the patients [7,17]. The interface retrieved spiral data from the database tables and then animated the drawing in real-time; that is with the same speed as the patients initially drew the spirals. A number of motor features were assessed by the clinician including 'impairment' on a scale from 1 (normal) to 10 (extremely severe), 'speed', 'irregularity', and 'hesitation' on a scale from 0 (normal) to 4 (extremely severe). The motor features were considered specific for the type of upper limb motor movements found in patients with motor fluctuations. Finally, 'cause' of the said dysfunctions was assessed as either tremor, bradykinesia, or dyskinesia. In case the clinician could not decide which category of 'cause' to select, he had the option to skip the rating.

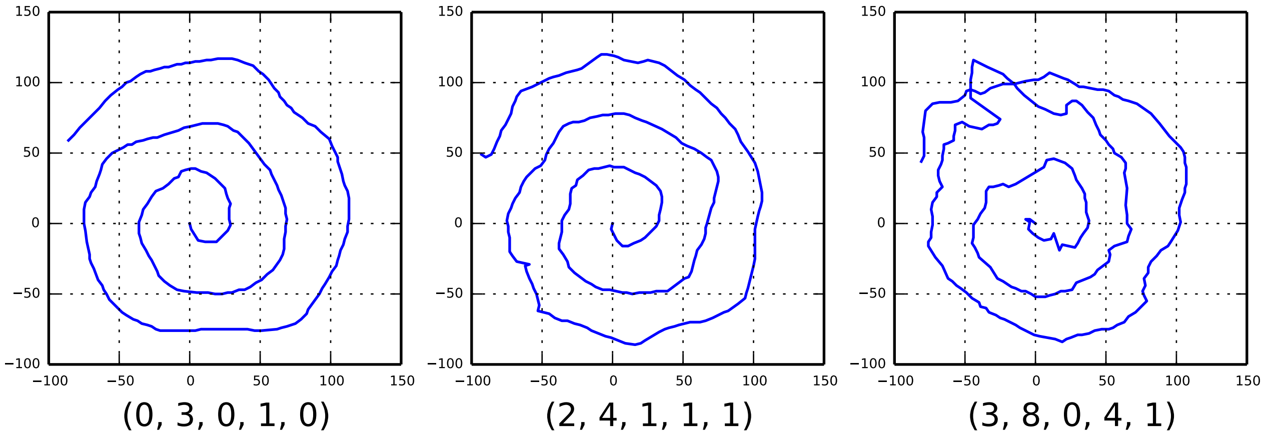


Fig. 1: Three example spiral drawings and ratings as given by the clinician. The values are for cause (0=none, 2=bradykinesia, 3=dyskinesia), impairment (on a scale from 1=normal to 10=extremely severe), speed, irregularity, and hesitation (on a scale from 0=normal to 4=extremely severe), respectively.

2.4 Selection of cases for machine learning

Three test occasions per patient were randomly selected from the database. However, there was incomplete data available for 5 patients and these were excluded from the dataset, leaving 180 cases (three cases per each of 60 patients with complete data) to be assessed by the physician.

Out of 180 cases, there were 38 cases rated as bradykinesia, 119 as dyskinesia, 1 as tremor, and 22 were skipped by the rater. Only cases that were rated as bradykinesia and dyskinesia were included in the subsequent analysis, giving a final dataset of 157 cases to be used in this study.

For the purpose of this study we treated separate test occasions from the same patient as independent cases. Our reasoning is that these were taken a considerable time apart (the clinical trial lasted for over three years) and motor symtoms can progress quite fast in PD patients. [19] The rate of progression depends on many factors, such as the degree of motor impairment and age at commencement of treatment (the higher the impairment and the bigger the age at treatment commencement the faster the progression of the disease). [20] Strictly speaking, however, these samples are not completely independent.

2.5 Features and pre-processing

The starting point of the data analysis phase was the raw data gathered with the Swedish telemetry device. A spiral drawn by a patient was described as a sequence of triples, (t, x, y), where t denotes the time (in ms) and x, y denote a Cartesian coordinate of a point of the spiral (as seen on the screen of the device). We used raw data (t, x, y) to calculate an applicable subset of the features from the PARKINSONCHECK application, also precisely matching all the pre-processing involved. The exact detailed calculation including the pre-processing such as smoothing and normalisation are given in [10]. For each of the three spirals (repetitions), the constructed features included: root mean squared error (RMSE) between the patient's spiral and the optimal spiral, computed in polar coordinates, absolute, radial and tangential speed of drawing, the percentage of the spiral length when the patient is drawing towards the centre, and the parameters of oscillations; altogether over 80 features per spiral. The meaning of the most important features is briefly described in the next section. These features were aggregated over the three repetitions using operators such as min, max, average and range over all three spirals.

Using these features, a learning example is composed of only aggregated features over all three repetitions, and a 'cause' class as appraised by the clinician. The learning situation was thus identical to the one the neurologist had when initially classifying the examples. Only cases rated as bradykinesia or dyskinesia were included for the learning problem, giving the majority class of roughly 76%. Each case (learning example) represents a test occasion.

2.6 Differences with the dataset from PARKINSONCHECK

The two datasets, the PARKINSONCHECK and the Swedish one used in this study, were collected for a different purpose. PARKINSONCHECK is an application for *early* detection of signs of Parkinsonian or Essential tremor (ET, the most common differential diagnosis of PD) and for differentiation between these two types of tremor, while the Swedish data were collected during a longitudinal study aimed at monitoring the *advanced* PD patients.

As straightforward as spirography looks, the two datasets — that is Swedish and Slovenian (PARKINSONCHECK) — were collected with different data collections schemes. The following are the main discrepancies between the two schemes:

- A different equipment (smartphones or tablets vs specialized device) was used for drawing the spirals using a very different sampling rate (> 50 Hz vs 10 Hz);
- The spirals were drawn with a stylus instead of with fingers alone (different biomechanics involved);
- Different direction of drawing (counter-clockwise vs clockwise);
- Completely unsupervised data collection (patients were on their own) and without safeguards for detecting the center of the screen and direction of drawing;
- In the Swedish study, patients were asked to repeat the spiral test (tracing the pre-drawn spiral template on the screen) 3 times per test occasion using the dominant hand, while PARKINSONCHECK data consists of four different drawings: with and without the template visible, using first the right hand and then the left hand;

- The Swedish patients were specifically instructed to complete the spiral drawing test in approximately 10 seconds per drawing, while no such instruction was given for PARKINSONCHECK data.

As the two datasets differ in so many aspects, it is interesting to investigate the extent of symptom information resolution of the previously developed features and symptom scoring methodology of PARKINSONCHECK applied on the Swedish dataset.

2.7 Machine learning methods

We tested four machine learning methods using Orange machine learning suite[15]: logistic regression (LR), naïve Bayes classification (NB), support vector machines (SVM), and random forests (RF). The first two were selected because they are linear and we prefer linear models since their classification models are simple to understand. The second two, that can extract also nonlinear patterns, were used to determine whether linear models are sufficient for this problem or not.

The parameters of the methods were tuned with an internal cross-validation. We have not used any feature subset selection techniques since they did not produce any considerable improvements in our previous work with similar data. However, all continuous attributes were first automatically discretized with the entropy-based discretization [13]. Attributes that could not be discretized, were removed from the learning data, see Sadikov [10] for details. The probability estimates from SVM outputs were estimated with an improved implementation of Platt's method, as used in LIBSVM [18]. In the case of random forests, the predicted class probabilities were obtained by averaging probabilities from each of the trees.

The methods were evaluated with a 10-times repeated tenfold cross-validation procedure with stratified sampling, for the following reasons: (a) using cross-validation will result in less biased estimates, (b) repeating cross-validation procedure on different splits will result in a smaller variance on an accuracy estimate, and (c) stratified sampling decreases the differences between class distributions in the learning and testing sets. To prevent overfitting, the discretization was first applied only on the learning data and then the same thresholds were used on the testing data. To compare the qualities of learned models, we used the following measures: classification accuracy (CA), area under the curve (AUC), and Brier score, which is the quadratic loss of the probability estimates.

2.8 Results

Table 1 summarizes the results for various machine learning algorithms for differentiating between the state of bradykinesia and the state of dyskinesia. There is no significant difference between the algorithms, they differentiate between the two states almost equally well. In the continuation, we will therefore focus on logistic regression (LR) as its model is relatively easy to understand as compared to the other algorithms. Comprehensibility, while not being an unconditional requirement, is a welcome feature, especially so with new methodology still being validated. The confusion matrix for logistic regression model is given in Table 2.

Table 1: The results with standard deviations (classification accuracy, Brier score, area under the curve) for logistic regression (LR), random forests (RF), support vector machines (SVM), and naïve Bayes classifier (NB).

	Majority	LR	RF	SVM	NB
CA	0.758±0.000	0.846±0.013	0.864±0.009	0.841±0.012	0.853±0.009
Brier	0.367±0.000	0.221±0.014	0.195±0.004	0.232±0.017	0.314±0.009
AUC	0.500±0.000	0.914±0.010	0.925±0.008	0.896±0.015	0.848±0.011

Table 2: The confusion matrix for logistic regression model. The results are averaged over all repetitions of 10-fold cross-validation.

	Bradykinesia (predicted)	Dyskinesia (predicted)
Bradykinesia (clinician)	23.2	14.8
Dyskinesia (clinician)	9.4	109.6

Table 3 contains ten most important features for the logistic regression model. The three most important features (avgP.min) all describe variability in speed (radial, tangential, and absolute) of drawing over the whole range of the spiral. The minimum over all three spiral drawings is taken. The features plrErrComCnt (avg and max) describe the level of curvature or smoothness of the spirals. The percentage of time the patient draws towards the centre is also quite important (percNeg feature) as well as the general misfit of the drawing as compared to the ideal spiral in polar coordinates (plrErrFit). The number of times the spiral crosses itself (rot.avgP) is also important — it is a good measure of severe fluctuations during drawing.

3 Clinical appraisal of the model

We envision the application of the methodology presented in this paper in two distinct ways: (a) as the built-in expert system of a monitoring device used independently by the patient with PD, or (b) as a decision support system of a digital spirography application in the hands of a clinician.

Table 3: The coefficients of the logistic regression with pre-discretization of attributes. Only ten most influential attributes (as measured by beta value range) are given.

Attribute	Importance	General description
radSp.avgP.min	1.13	radial speed variability
tangSp.avgP.min	1.02	tangential speed variability
absSp.avgP.min	0.78	absolute speed variability
plrErrComCnt.avg	0.70	level of curvature/smoothness of the spiral
radSp.percNeg005.min	0.69	percentage of time the patient drew towards the centre
plrErrComCnt.max	0.67	level of curvature/smoothness of the spiral
plrErrFit.avg	0.65	general misfit from the ideal spiral (template)
tangSp.avgP.rng	0.65	tangential speed variability
tangSp.avgP.max	0.63	tangential speed variability
rot.avgP.min	0.62	number of times the spiral crosses itself

The self-monitoring application represents a telemedical setting where the patient regularly measures him- or herself with a device for digital spirography; this could also be an app on a smartphone similar to PARKINSONCHECK [11]. These measurements should be done alongside the current practice of keeping a diary of medication intake and noting the general well-being. The latter is quite subjective and in this respect the spirographic measurements provide an additional objective component for the appraisal of the patient's condition over time.

In this type of application the built-in expert system would automatically analyse the condition of the patient over time to detect whether he or she is potentially under- or over-medicated, both conditions being of interest. The objective dosage monitoring is a very serious issue in managing of Parkinson's disease as the disease slowly progresses over a long period of time and the patient typically sees the neurologist only once or twice per year in most countries. The expert system would alarm the patient and/or clinician when it would detect a serious anomaly *over time*. In this setting, (classification) accuracy is probably the most important metric for the appraisal of the system, and comprehensibility is of secondary importance. Having said that, it is still a good idea to visualise the way computer "thinks", though. We continue this thought later in this subsection.

The second application of our methodology is envisioned as a decision support system for physicians in a clinical setting. Often times spirography is performed simply on a piece of paper as part of a clinical examination; digital spirography with the calculation and presentation of various features is still relatively new.

Our methodology could take this one step further and offer a clinician not just calculated results of features of spirographic test, but a computer's opinion on the patient's condition. This would serve for decision support purposes or as an additional opinion and would completely automate the digital spirographic test. Such a decision support system could thus be used either instead of a spirographic test performed on paper or, perhaps more importantly, to analyse patient's drawings over a longer period of time (either drawn at home or in a hospital setting). Currently, it is quite difficult for clinicians to assess the patient's condition over a longer period of time (e.g. to prescribe the correct dosage of the medication) on their (yearly) visit as they rely only on the patient's (subjective) diary. This diary consists of a large volume of entries and it is hard for clinicians to go over that in a limited amount of time. An automated solution (in this case over the spiral drawings, but also potentially including patient's subjective appraisal and medication intakes) would go a long way to help with this problem.

To this end, not only (classification) accuracy is important as a metric of success, but also the comprehensibility of the computer's reasoning and explanation. It is known that well explained suggestions are more easily accepted by the experts.

Our approach to the explanation is to highlight the specific interesting parts of the spiral. This gives an immediate visual clue to the clinician. Such an approach is perhaps quite novel in the sense that we do not highlight the machine learning features in their feature space, but back on the original spiral drawing for static features and on spiral animation for temporal features. This approach is described in more detail in the next subsection.

Visualisation can be very useful for some other purposes as well. As stated in [12]: "The purpose of such visualisation is threefold: (a) it provides the physician with immediate visual clues to be aware of when assessing the spiral, (b) it can serve as a 'visual debugging aid' for the developers of a DSS, and (c) it can trigger generation of new domain knowledge". Visual debugging and new knowledge generation is perhaps even more fitting or relevant in this case as the features and preprocessing methodology was not designed with this specific task (separating bradykinesia from dyskinesia) in mind. Thus, while the first experiment is concerned with model building and its accuracy, the second experiment is geared towards the analysis of the model's comprehensibility and usefulness in a decision support setting.

3.1 Visualisation of the model's decision making

It is often hard, even for a trained eye, to derive all of the important information from a given time series — in our case from a spiral. However, this is not hard for a computer using different algorithms to detect anomalies. Yet people frequently do not trust in the computer's results (although correct) as they do not see the logic behind its decision. Therefore it is important to try to explain the computer's decision. Usually this is done with the written explanation of the decision model, which cannot be always done in a comprehensible way. We tried to make a step further in our explanations to somewhat overcome this problem. We wanted to explain the decisions on the original data — in our case to highlight the parts of the actual spiral that was drawn, and not the feature space (which is often incomprehensible to humans). With this idea in mind, we developed an algorithm which finds the parts of the spirals, which have contributed the most to the computer's decision.

Let us first define some of the basic terms which will be used hereinafter.

A *time series T* is a sequence of values $t_1, t_2, ...t_n$ measured in subsequent time points.

An *attribute A* is a function of a time series *T*: $A = f(T)$.

Let us say we have a time series *T* of the length *n*. An *interval I* of the length $s \leq n$ is a subsequence of consecutive points of the time series *T*.

Suppose a *W* is a set of all of the possible intervals *I*.
The *goal* is to find $W_a = \{I_1, I_2, ...I_k\}$, where $W_a \subset W$, so that W_a is a subset, where $\mid I_1 \cup I_2 \cup ... \cup I_k \mid$ is the smallest, and that with adding new intervals I_j in a subset W_a *benefit* does not rise significantly.

Benefit is defined as a ratio between the sum of interval *contributions* and their *cost*.

If *T'* is a modified time series *T*, where the values of the points on the interval I were replaced with expected (neutral) values of the attribute *A*, then a *contribution* of an interval is defined as $\mid f(T) - f(T') \mid$.
Interval *contributions* are normalised to an interval $[0, 1]$.

Contribution of an interval is a good indicator of which of the intervals are the most important for a given classification. We could actually highlight n intervals with the highest *contributions*. If we have intervals of length one, in the worst case scenario the intervals of the highest values of *contribution* are not adjacent and we would therefore highlight individual points which are scattered around the whole time series. For the majority of the attributes, this is not really useful as we would not understand the computer's reasoning behind the classification any better than without the visualisation. This is the reason why we would like to find adjacent intervals instead of highlighting scattered points. This kind of visualisation can be more beneficial from the cognitive point of view and can give experts more information. We achieved this by introducing interval *cost*.

Cost of the intervals is calculated using the formula:

$$cost = \sum_I (\omega * \mid I \mid + (1-\omega) * log_2(1+ \mid I \mid))$$

The *cost* function is made up of two parts in both of which the length of an interval ($\mid I \mid$) is one of the main arguments. With this we achieved that our method prefers longer intervals to shorter ones. A short example of what we achieve with this *cost* function is shown in a Table 4. The example is also illustrated in Picture 2.

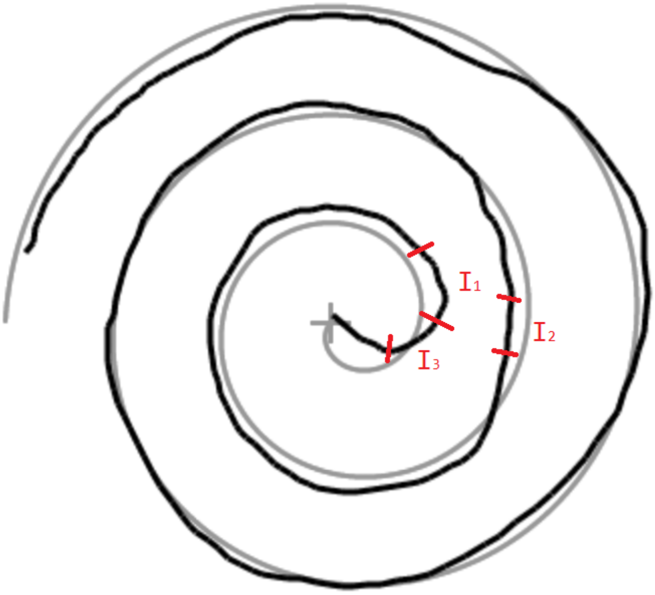


Fig. 2: Illustration of positions of intervals I_1, I_2, and I_3 of the length 2 with the adjacent intervals I_1 and I_3.

We are showing three intervals I_1, I_2, and I_3 of the length 2 with the adjacent intervals I_1 and I_3. From the results we can see that if we would not prefer longer intervals (e.g. joining the adjacent intervals), we would highlight intervals I_1 and I_2. In the case of joining intervals I_1 and I_3 their *cost* is lower than the cost of using I_1 and I_2 (which are not adjacent and cannot be joined). The benefit of the joined intervals I_1 and I_3 is higher than the benefit of interval I_1 or interval I_2 and more than two times higher than the benefit of interval I_3 alone. This results in highlighting the joined interval I_1 and I_3.

Since we did not want to punish shorter intervals unproportionately, we balanced the formula by using a weight ω. We have to set the weight for every attribute, depending mostly on the meaning of the attribute and of our preferences — whether we want longer or shorter intervals. This cannot be

Table 4: An example of costs and benefits of the individual intervals of the same lengths and of the combination of the intervals. Weight ω in the *cost* function is set to 0.5. Values for the benefit were calculated using the formula $benefit = \frac{contribution}{cost}$.

Interval	Contribution	Cost	Benefit
I_1	6.0	1.79	3.35
I_2	5.5	1.79	3.07
I_3	4.0	1.79	2.23
$I_{1,2}$	11.5	3.59	3.21
$I_{1,3}$	10.0	**2.16**	**4.63**
$I_{2,3}$	9.5	3.59	2.65

made automatically since the decision of how long intervals we want to highlight is largely a cognitive decision.

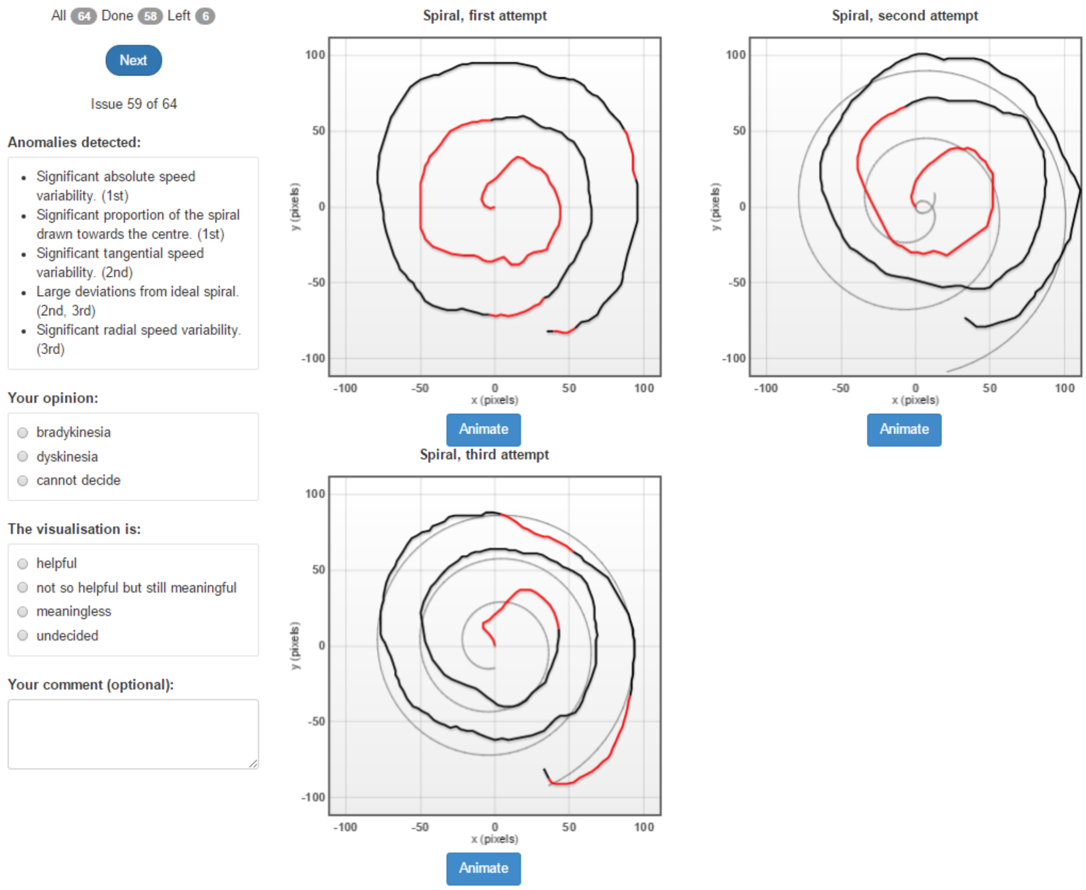
Searching for the most important parts of the time series is done only if the attribute value on the whole time series is high (or low, depending on the model) enough for the attribute to appear in the decision model. If the value is not high enough, we are not interested in highlighting its results on a time series.

3.2 Objectives and experimental setup

The clinical appraisal experiment was conducted to assess the model learned as the result of the machine learning experiment described earlier. As the accuracy of the model was already tested in the first experiment, the objectives of this assessment were focused on evaluating the model's suitability for use as a decision support tool for a clinician. In this respect, we were interested in the appraisal of comprehensibility of the model and its explanations, and general acceptance of the model's "thinking" as seen by clinicians.

The experiment was set up using our web-based spiral drawing viewer application shown in Figure 3. The task was to classify test instances as either bradykinetic or dyskinetic. Each test instance consisted of three spiral drawing repetitions as this is how the patients were asked to perform the individual measurements. The clinicians were also able to see the animation of all the presented spiral drawings. The animations were in real-time, exactly how the patients drew them. In this way the dynamic (speed) characteristics of the drawings could be analysed.

The test consisted of 32 instances, of which 16 were bradykinetic and 16 were dyskinetic. The clinicians (D.N. and D.G.) were first presented with all 32 instances in random order (as to the class) without the visual clues of the model. In the second part of the experiment all the instances were repeated in the same order as before but with the visual clues and textual explanations (red highlighted parts of the drawings and text under the "anomalies detected" heading). The visual clues for dynamic



All 64 Done 58 Left 6

Issue 59 of 64

Anomalies detected:

- Significant absolute speed variability. (1st)
- Significant proportion of the spiral drawn towards the centre. (1st)
- Significant tangential speed variability. (2nd)
- Large deviations from ideal spiral. (2nd, 3rd)
- Significant radial speed variability. (3rd)

Your opinion:

○ bradykinesia
○ dyskinesia
○ cannot decide

The visualisation is:

○ helpful
○ not so helpful but still meaningful
○ meaningless
○ undecided

Your comment (optional):

Fig. 3: Web-based spiral drawing viewer displaying an instance with visual clues as used in the appraisal of visualisation.

characteristics (blue highlights) were also shown when animating the drawings. The number of instances was a compromise between the sample size and the duration of the experiment. Altogether each participant had to assess 64 instances (2 × 32) consisting of three spiral drawings each for a total of 192 drawings.

The clinicians were told in advance that all the instances were either bradykinetic or dyskinetic, but not the distribution of the instances between the two classes. In both parts of the experiment the clinicians were asked to classify the instances also having the option “cannot decide”. In the second part they were not shown their classification from the first part. In the second part of the experiment the clinicians were additionally asked to evaluate the visual (and textual) clues given by the model as to their helpfulness and meaningfulness for the task. The four options for the assessment of the visualisation given were:

- helpful;
- not so helpful, but still meaningful;
- meaningless;

– undecided.

Additionally, it was also possible to leave a comment for each instance.

The selection of instances was geared towards (more) clear cut cases. The logic behind this was that in this experiment we predominantly wanted to evaluate the model's explanations and validity, and not test its predictive accuracy. Another reason for this selection was that the ground truth for the instances was obtained on the basis of spiral drawings alone (no clinical confirmation as the patients took the recordings in an unsupervised setting) and we wanted to use the instances that were most likely correct. The ratio between dyskinetic and bradykinetic cases in the full dataset (see first experiment) was about three to one. However, we decided on an equal distribution between the classes for this experiment since the sample size was not that large and we wanted the bradykinetic class to still be well represented.

We had at our disposal a set of instances assessed for ground truth by four independent clinicians for this experiment [17]. For the reasons above we selected a random selection of 16 instances of each class from this set with the additional stipulation that there had to be a majority of three evaluators to one or better for each instance regarding its class. However, the large majority of the selected instances were unanimously classified by the evaluators.

3.3 Results

The assessment was performed by two clinicians. One of them, D.N. was involved in this research before and was also involved in the ground truth assessment a year ago. The other, D.G. was not involved before, but is very familiar with spirography. However, he never used spirography to look for dyskinesia before, and was thus completely new to this particular task. He did not receive any specific prior training or instructions besides the description of the task and relied on his medical knowledge in the area of movement disorders alone.

Table 5: The confusion matrix for D.N. without visual clues.

	Bradykinesia (predicted)	Dyskinesia (predicted)	Undecided
Bradykinesia (true)	6	10	0
Dyskinesia (true)	0	16	0

The results of the classification task are presented as confusion matrices for each of the clinicians separately (Tables 5–8). There is one matrix for each part of the test, without and with visual clues,

Table 6: The confusion matrix for D.N. with visual clues.

	Bradykinesia (predicted)	Dyskinesia (predicted)	Undecided
Bradykinesia (true)	4	12	0
Dyskinesia (true)	0	16	0

as we were interested in the potential effect the visualisation has on the classification task. From the matrices we can see that both clinicians only very rarely changed their opinion in the second part of test: one changed his mind twice (both times for the worse) and the other once (positively from mistaken to at least undecided). However, all in all, there was minimal change in the classification decisions for both of them.

Table 7: The confusion matrix for D.G. without visual clues.

	Bradykinesia (predicted)	Dyskinesia (predicted)	Undecided
Bradykinesia (true)	16	0	0
Dyskinesia (true)	5	9	2

Table 8: The confusion matrix for D.G. with visual clues.

	Bradykinesia (predicted)	Dyskinesia (predicted)	Undecided
Bradykinesia (true)	16	0	0
Dyskinesia (true)	4	9	3

If taken together, the classification accuracy of both clinicians is approximately 75%. This suggests, that even taking more clear cut examples, the problem is far from trivial and a decision support system could be a very useful tool. Furthermore, it is interesting how the two clinicians differed in their classifications. While one of them perfectly detected all dyskinetic cases and had problems with the bradykinetic ones, it was just the opposite for the other. The latter can be explained by D.G.'s statement that he never used spirography for detecting signs of dyskinesia before.

As a side note the classification accuracy of the model used for visualisation was 100%. This does in no way suggest that the model is perfect. The results from Table 2 convey the real accuracy of the model. However, it is a welcome circumstance as this makes the visual clues more reliable — after all visualisation depends not only on the features used in the model but also on which features were important for the *given* classification.

Table 9: Opinion on the acceptability of visualisation.

	Helpful	Not so helpful, but still meaningful	Meaningless	Undecided
Number of answers	29	32	0	3

The other question dealt with the issue whether the visualisation and thus computer's decision model (reasoning) makes sense to the clinicians. We consider this the main question of this experiment, the accuracy with and without visualisation was tested mainly to see whether it has any significant influence on the decision process (good or bad). The combined results are presented in Table 9. The large majority of "not so helpful, but still meaningful" answers came from D.N., while the large majority of "helpful" answers came from D.G. This probably makes sense as the former was involved in this research from the onset, while it was new to the latter one. The combined results (no cases where the visualisation seemed meaningless and just three undecided cases out of 64) suggest that the model and its features are reasonable for this task.

4 Discussion

As we stated in the introduction the main objective of this research was to analyse whether the descriptive features and methodology developed for PARKINSONCHECK application for early detection of signs of Parkinsonian or Essential tremor can be applied more generally. In this paper we applied it to the clinically important task of monitoring advanced PD patients for signs of bradykinesia and dyskinesia which directly relate to the drug dosage adjustment problem. We have performed two experiments: the first one is a classic application of machine learning methodology to obtain a model to differentiate between the two conditions, and the second one is aimed at validating the model in terms of its comprehensibility which relates also to the model's acceptability in a clinical setting. We discuss both experiments separately in the continuation.

4.1 Learning the classification model

The main observation of the first experiment is that all the algorithms are clearly better than the majority classifier. This indicates that the features of the Slovenian PARKINSONCHECK application contain relevant symptom information for objective assessment of PD motor symptoms when applied to data collected by a telemetry device used in a three years Swedish clinical study. Even though the features were designed for early detection of PD and ET symptoms, the findings indicate that a (sub)set of them can be used for recognising upper limb motor movements specific to Off episodes and peak-dose dyskinesias, which are prominent in advanced patients experiencing motor fluctuations.

In order to cover cases from all the patients involved in the Swedish study and to make it possible for the clinicians to rate the spirals, a sample of spirals from the whole database has been drawn by randomly selecting 3 test occasions per patient [7].

The results in this experiment were obtained using only the aggregated features over all three spiral drawings. While experimenting, we also fitted models using the non-aggregated features, and the observed classification accuracies were up to 90% (with increases in other metrics as well). However, we decided to present the results only for aggregated features as we currently do not have a good explanation why any given spiral (first, second, or third attempt) would be more important than the others. It does indicate, though, that there is more information that can be extracted from this data. From the confusion matrix presented in Table 2, we can conclude that there were more misclassifications for bradykinesia class than dyskinesia. This can probably be because of the unbalanced data design where majority of the spirals (76%) were rated as dyskinesia. This is something that we are looking into while continuing this research. Moreover, it has to be noted that both possible misclassifications carry equal weight as both conditions are unpleasant for the patient and both need to be properly addressed.

In contrast to other technology-based symptom assessment strategies which are mainly based on the use of wearable sensors [8], spirography has been used mainly for assessing the severity of the symptoms. To our knowledge, only [9] have tackled the problem of assessing motor fluctuations using spirography tasks. However, they have mainly focused on quantifying the severity of dyskinesias only, by limiting data processing on frequency bands relevant to dyskinesias, and with no reference to Off symptoms. In contrast, our approach is designed to capture movement patterns exhibited by patients being in Off (bradykinesia) and dyskinesia motor states.

4.2 Evaluation of the model's comprehensibility

The objective of the second experiment was to check the model obtained as the result of the first experiment for comprehensibility. If the model is comprehensible and assessed by the clinicians as meaningful, then it can more convincingly be used in a decision support setting.

We deem this second experiment as particularly important in this particular case as all the methodology (features, preprocessing, etc.) was not developed with this specific task in mind. Comprehensibility was thus far from guaranteed.

The results from this experiment are somewhat mixed. There seem to be no doubt that the clinicians find the visual clues (and consequently "reasoning" of the model) at least meaningful if not downright helpful. On the other hand, we did not observe much influence of the visualisation on the clinicians' decision making. To a certain extent this result surprised us as we did a very similar experiment (for a different task) and there one the main findings was that the visual clues very much affect the decision making of the clinicians (this research is still ongoing and is not published yet).

There could be several reasons for finding no influence of visualisation in this experiment. We specifically selected more clear cut cases to be particularly certain that the class (ground truth) is correct. This, however, might have resulted in clinicians being quite certain of their decision — although their accuracy does not coincide with this. There might also be an issue that 32 instances are not enough and that they were able to remember their classifications from the first part when presented with the same instance again. Perhaps this is also the result of a possible bias of both clinicians: one was aware of the uneven distribution of classes from the first experiment (about 76% of the cases were dyskinetic) and this could somewhat influence him while the other was familiar with observing bradykinesia in the drawings, but completely unfamiliar with using spirography for observing dyskinesia. As it stands, repeating the experiment with more clinicians and more perhaps more instances makes sense.

We believe, though, that we can take away from this experiment that the model is sensible and the features used are useful for this task in a telemedical setting and perhaps even as an automatic analysis tool for a patient's diary of spirographic measurements. On the other hand, it remains to be seen whether the model/features can be useful in a decision support setting.

5 Conclusions and Future Work

The present study demonstrates that a lot of information about the PD patient's condition can be extracted from his or her spirographic data. The main conclusion is that it is feasible to apply PARKINSONCHECK's features and pre-processing methodology to a substantially different set of spirographic data and obtain a comprehensible model with good predictive accuracy. The features can thus be thought of as quite general for the description of spiral drawings.

In the long term this result suggests that spirography could be used as a valid method for objective monitoring of PD patients, especially combined with other tests that could detect those conditions that were currently misclassified. The latter could be improved with new features as well, however, we suspect that tests complementary to spirography will have an even more significant impact.

The obvious future work is to look into the misclassified cases and try to either improve the model or get an understanding of why the misclassifications occurred. It would be interesting for a clinician to review the misclassified cases, perhaps changing his opinion or pointing the reasons for the misclassification. This could lead to potential new features for describing the spirals. Or it could hint at the lack of information in the spirals for this particular problem. On this topic, there were a few comments from the clinicians in the second experiment describing some cases as difficult or expressing surprise with the lack of some feature being highlighted. Interestingly in all those cases, they made a classification error. However, these are exactly the cases that are most fruitful for further analysis.

In the future, the plan is to collect more ratings on animated spirals from more clinicians, but also to collect more spiral drawings in a controlled environment. This would allow us to investigate the feasibility of time-space reconstruction of the spirals to clinicians and to investigate how the machine learning approach would work when the ground truth is more reliable. The long-term plan is also to include multiple objective motor function measurements with different sensors including wearable sensors, eye tracking and upper limb touch screen tests (tapping and spirography) as well as video recording the patients while performing the tests and executing standardised motor tasks. This would allow us to investigate the relationship of the objective measures and blinded video ratings as well as their relationship to plasma levodopa concentration.

Acknowledgements

This work was partly supported by Swedish Knowledge Foundation and Sweden's Innovation Agency (VINNOVA) as part of FLOAT and MuSyQ projects, and partly by Slovenian Research Agency (Artificial Intelligence and Intelligent Systems research programme), Slovenian Ministry of Education, Science and Sport, and European Regional Development Fund (PARKINSCHECK project). M. Memedi and D. Nyholm are shareholders in Jemardator AB, holder of IP of a test battery for collecting symptom data.

References

1. Martinez-Martin, P., Gil-Nagel, A., Gracia, L. M., Gomez, J. B., Martinez-Sarries, J., Bermejo, F.: Unified Parkinson's disease rating scale characteristics and structure. Movement Disorders, vol. 8, pp. 76–83, (1994).
2. Taylor Tavares, A. L., Jefferis, G. S., Koop, M., Hill, B.C., Hastie, T., Heit, G., Bronte-Stewart, H. M.: Quantitative measurements of alternate 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. Movement Disorders, vol. 20, pp. 1286–1298, (2005).
3. Hagell, P., Whalley, D., McKenna, S. P., Lindvall, O.: Health status measurement in Parkinson's disease: validity of the PDQ-39 and Nottingham health profile. Movement Disorders, vol. 18, pp. 773–783, (2003).

4. Haubenberger, D., Kalowitz, D., Nahab, F.B., Toro, C., Ippolito, D., Luckenaugh, D.A., Wittevrongel, L., Hallett, M.: Validation of digital spiral analysis as outcome parameter for clinical trials in essential tremor. Movement Disorders, vol. 26, pp. 2073–2080, (2011).
5. Westin, J., Dougherty, M., Nyholm, D., Groth, T.: A home environment test battery for status assessment in patients with advanced Parkinson's disease. Computer Methods and Programs in Biomedicine, vol. 98(1), pp. 27–35, (2010).
6. Palhagen, S.E., Dizdar, N., Hauge, T., Holmberg, B., Jansson, R., Linder, J., Nyholm, D., Sydow, O., Wainwright, M., Widner, H., Johansson, A.: Interim analysis of long-term intraduodenal levodopa infusion in advanced Parkinson disease. Acta Neurologica Scandinavica, vol. 126, pp. e29–33, (2012).
7. Memedi, M., Bergqvist, U., Westin, J., Grenholm, P., Nyholm, D.: A web-based system for visualizing upper limb motor performance of Parkinson's disease patients. Movement Disorders, vol. 28, pp. S112–S113, (2013).
8. Maetzler, W., Domingos, J., Srulijes, K., Ferreira, J.J., Bloem, B.R.: Quantitative wearable sensors for objective assessment of Parkinson's disease. Movement Disorders, vol. 28, pp. 1628–2637, (2013).
9. Liu, X., Carroll, C.B., Wang, S.Y., Zajicek, J., Bain, P.G.: Quantifying drug-induced dyskinesia in the arms using digitised spiral-drawing tasks. Journal of Neuroscience Methods, vol. 144, pp. 47–52, (2005).
10. Sadikov, A., Žabkar, J., Možina, M., Groznik, V., Georgiev, D., Bratko, I.: PARKINSONCHECK: A decision support system for spirographic testing. Tech. Rep., University of Ljubljana, Faculty of Computer and Information Science, (2014).
11. Sadikov, A., Groznik, V., Žabkar, J., Možina, M., Georgiev, D., Pirtošek, Z., Bratko, I.: PARKINSONCHECK smart phone app. Proceedings of European Conference on Artificial Intelligence, pp. 1213–1214, (2014).
12. Groznik, V., Sadikov, A., Možina, M., Žabkar, J., Georgiev, D., Bratko, I.: Attribute Visualisation for Computer-Aided Diagnosis: A Case Study. Proceedings of the IEEE International Conference on Healthcare Informatics, pp. 294–299, (2014).
13. Fayyad U.M., Irani K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022–1029, (1993).
14. Elble, R.J., Sinha, R., and Higgins, C.: Quantification of tremor with a digitizing tablet. Journal of Neuroscience Methods, vol. 32, pp. 193–198, (1990).
15. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinović, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B.: Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, vol. 14, pp. 2349–2353, (2013).
16. Fahn, S., Elton, R., and UPDRS Development Committee: Unified Parkinson's disease rating scale. Recent Developments in Parkinson's Disease, vol. 2, pp. 153–163, 293–304, (1987).
17. Memedi, M., Sadikov, A., Groznik, V., Žabkar, J., Možina, M., Bergquist, F., Johansson, A., Haubenberger, D., Nyholm, D.: Automatic spiral analysis for objective assessment of motor symptoms in Parkinson's disease. Sensors, vol. 15, pp. 23728–23744, (2015).
18. Chih-Chung Chang and Chih-Jen Lin: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.
19. Jankovic, J., Kapadia, A.S.: Functional Decline in Parkinson Disease. Archives of Neurology, vol. 58, pp. 1611–1615, (2001).
20. Ding, C., Ganesvaran, G., Alty, J.E., Clissold, B.G., McColl, C.D., Reardon, K.A., Schiff, M., Srikanth, V., and Kempster, P.A.: Study of levodopa response in Parkinson's disease: Observations on rates of motor progression. Movement Disorders, vol. 31(4), pp. 589–592, (2016).

Feasibility of spirography features for objective assessment of motor function in Parkinson's disease

Highlights

- A method for self-monitoring the motor function in Parkinson’s disease is presented.
- Slowness of movement (bradykinesia) is typically associated with under-medication.
- Involuntary movements (dyskinesia) can be the result of over-medication.
- A machine learning model that detects bradykinesia and dyskinesia is proposed.
- The model’s visual explanatory power is evaluated.