



A Decision Tree for Automatic Diagnosis of Parkinson's Disease from Offline Drawing Samples: Experiments and Findings

Antonio Parziale^(✉), Antonio Della Cioppa, Rosa Senatore,
and Angelo Marcelli

Natural Computation Lab, DIEM, University of Salerno, Fisciano (SA), Italy
{anparziale, adellacioppa, rsenatore, amarcelli}@unisa.it

Abstract. We address the problem of designing a machine learning tool for the automatic diagnosis of Parkinson's disease that is capable of providing an explanation of its behavior in terms that are easy to understand by clinicians. For this purpose, we consider as machine learning tool the decision tree, because it provides the decision criteria in terms of both the features which are actually useful for the purpose among the available ones and how their values are used to reach the final decision, thus favouring its acceptance by clinicians. On the other side, we consider the random forest and the support vector machine, which are among the top performing machine learning tool that have been proposed in the literature, but whose decision criteria are hidden into their internal structures. We have evaluated the effectiveness of different approaches on a public dataset, and the results show that the system based on the decision tree achieves comparable or better results that state-of-the-art solutions, being the only one able to provide a plain description of the decision criteria it adopts in terms of the observed features and their values.

Keywords: E-health · Trusted Artificial Intelligence · Parkinson's disease · Machine learning

1 Introduction

Parkinson's disease (PD) is a neurodegenerative disorder that affects dopaminergic neurons in the Basal Ganglia, whose death causes several motor and cognitive symptoms. PD patients show impaired ability in controlling movements and disruption in the execution of everyday skills, due to postural instability, onset of tremors, stiffness and bradykinesia [8–10, 15]. In the last decades, the analysis of handwriting and/or drawing movements has brought many insights for uncovering the processes occurring during both physiological and pathological conditions [2, 16, 17], and for providing a non-invasive method for evaluating the

stage of the disease [14]. A comprehensive survey of the literature on handwriting to support neurodegenerative diseases, including PD, may be found in [4]. As the authors pointed out, the large majority of studies aimed at investigating the handwritten production for the purpose of inferring which are the features that best characterize the production of PD patients with respect to healthy subjects. A variety of tests have been proposed, requiring the subject to write simple letters patterns, geometric figures, words, sentence and so on, but it has been recently shown that the most suitable ones are those requiring the drawing of geometric shapes such as spirals and meanders, with or without a reference pattern printed on the paper [20].

Following these suggestions, Pereira and his collaborators have collected the NewHandPD data set, which include both off-line images and on-line signals of the traces produced by the subjects while drawing 4 samples of spirals and 4 samples of meanders by writing on paper with a digitizing pen. A variety of top performing machine learning algorithms, such as Convolutional Neural Networks (CNN), Support Vector Machine (SVM), Optimal Path Finder (OPF), Random Forest (RF) and Restricted Boltzmann Machines (RBM) have been evaluated, showing that when on-line data are used, the ImageNet architecture could achieve a global accuracy of 87.14% in its best configuration, while the top performance on off-line data were achieved by the SVM, with a global accuracy of 66.72% on the meander drawings [11, 12].

In this framework, the aim of the work reported here is twofold: improving the performance on off-line data and providing an explicit representation of the criteria developed by the system for discriminating between PD patients and healthy subjects. Improving the performance obtained with static features is a goal worth to be pursued for two main reasons: (1) it would allow to analyze old writings or drawings produced by the subject (available only on paper given the age of insurgence of the diseases) for reconstructing the patient's medical history or to date the onset of the disease; (2) it would avoid that some subjects, especially elderly ones, feeling uncomfortable writing on a graphic tablet, may change their usual handwriting/drawing, thus introducing in their production unnatural characteristics that can lead to errors in assessment. Developing systems capable of explaining the decision procedure they have learned and adopted is currently a topic of investigation in the framework of so called explainable artificial intelligence (xAI) [1, 6]. In particular, in medical applications, adopting a decision procedure that can be described in a way that resembles the one routinely followed by clinicians when considering the results of tests, will favour their acceptance as supporting tool for early diagnosis.

To address both the issues, we propose to adopt a decision tree as the machine learning tool to perform the discrimination between handwritten samples produced by PD patients and healthy subjects. This choice is motivated by the structure of the tree that can be naturally "translated" into a chain of if-then decision rules, thus resembling very closely the diagnostic procedure. Other approaches such as fuzzy rule based systems have been also proposed, but they need specialized tools for automatically generating the decision rules [19]. The performance

of the decision tree are then compared with those achieved by a Support Vector Machine, that in the previous studies mentioned above has provided the best performance, and with those exhibited by a Random Forest, as it exploits the same basic idea of decision tree, but whose decision criteria are much more complex and much less explicable than those of the decision tree. In the remaining of the paper, Sect. 2 briefly summarizes the main features of the three machine learning tools we have compared, and the data set we have used during the experimental work. Section 3 describes the experiments we have designed and performed for the automatic learning of the decision trees and reports the results obtained in terms of patient classification, as to show to which extent the proposed tool can be used by clinicians in their daily practice. Eventually, in the conclusion, we summarize the work that has been done, discuss the experimental results and outline our future investigations.

2 Machine Learning Tools and Dataset

We briefly summarize in the following the machine learning tools we have compared and the dataset used for performance evaluation.

2.1 C4.5

C4.5 is a statistical classifier introduced by Quinlan [13] and used in data mining for inducing classification rules in the form of decision trees, which can be employed to generate a decision from a set of training data. At each node of the tree, C4.5 chooses the feature that most effectively splits the set of samples into subsets that best differentiate the instances contained in the training data. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to take the decision. Once C4.5 creates a tree node whose value is the chosen attribute, it creates child links from this node where each link represents a unique value for the chosen attribute and uses the child link values to further subdivide the instances into subsets.

2.2 Random Forest

A random forest [7] is a meta estimator that aggregates a number of decision tree classifiers, i.e., the forest, on various sub-samples of the dataset and use averaging to enhance the predictive accuracy while mitigating the over-fitting. In general, the more trees in the forest the more robust the forest looks like. In random forest algorithm, instead of using information gain for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly.

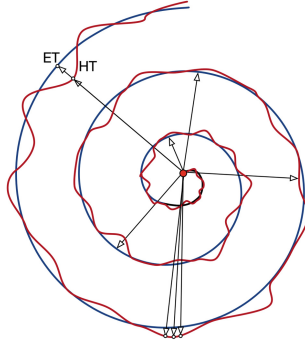


Fig. 1. The feature extraction process for a spiral. The blue line is the template trace ET, while the red line is the handwritten trace HT. The arrows indicate the radii for both ET and HT, the white circles indicate the intersection of a radius with the ET and the HT traces and the red circle represents the center of the ET. In order to compute all the features, the radius is shifted by using a predefined spanning angle. (Color figure online)

2.3 Support Vector Machines

The objective of the support vector machine [3] algorithm is to find a hyper-plane in an n -dimensional space (n is the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyper-planes that could be chosen. The objective is to find an hyper-plane that has the maximum margin, i.e., the maximum distance among data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Support vectors are data points that are closer to the hyper-plane and influence the position and orientation of the hyper-plane. Deleting the support vectors will change the position of the hyper-plane. Using these support vectors, we maximize the margin of the classifier.

2.4 The Dataset

NewHandPD dataset contains handwritten data collected from graphical tests performed by 31 PD patients and 35 healthy subjects. Each subject produced 4 samples of spirals and meanders, and from each sample 9 features, reported in Table 1, were extracted. As a result, the dataset is composed of 264 spirals and 264 meanders drawn by the participants following a printed template on paper with a pen. As a consequence, we have two unbalanced datasets, i.e., spirals and meanders, each of which is composed of 124 samples belonging to PD patients and 140 belonging to healthy subjects. Figure 1 shows, together on one sample, the two main geometric entities, namely the distance between the centre of the template and the template/written trace (ET/HT radius), from which all the features are computed.

Table 1. Features description of the dataset used. HT: handwritten trace, ET: exam template

Feature	Description
x_0	RMS of the difference between HT and ET radius
x_1	Maximum difference between HT and ET radius
x_2	Minimum difference between HT and ET radius
x_3	Standard Deviation of the difference between HT and ET radius
x_4	Mean Relative Tremor
x_5	Maximum HT radius
x_6	Minimum HT radius
x_7	Standard Deviation of HT radius
x_8	Number of times the difference between HT and ET radius changes sign

3 Experiments

To evaluate the performance of the proposed approach in providing explainable yet effective solutions, we performed a patient classification experiment, as described below.

3.1 Patient Classification

We divided the dataset into a Training set and a Test set made of 70% and 30% of the original dataset, respectively, in such a way as to maintain the relative occurrence of patients and healthy subjects. In particular, the Training set was made of 25 healthy subjects and 22 PD patients while the Test set was composed by 10 healthy subjects and 9 PD patients. The minimum and the maximum of each feature were computed on the Training set and a min-max normalization was applied to the whole dataset in order to scale all the features in the range $[0, 1]$.

The health condition of an individual was evaluated by classifying each of his/her handwritten samples and by applying a majority vote: an individual was classified as healthy or patient if the majority of his/her samples were assigned to the “healthy” or to the “patient” class, respectively. A decision about the health condition was rejected when the same number of samples were assigned to both the classes. The experimentation was performed with the aim of understanding: (1) which is the most performing classification schema among the ones described in Sect. 2, (2) if both meanders and spirals are necessary for evaluating the healthy condition, (3) how many samples are required for reaching the best performance. Therefore, we evaluated the health condition of individuals by classifying only the meanders, only the spirals and both meanders and spirals and by varying the number of samples per subject used during the training and the classification phases. In particular, samples were progressively included in

Table 2. Classifiers' parameters.

Classifier	Parameters
Decision Tree	Pruning confidence: 0.25, Minimum number of instances per leaf: 2
Random Forest	Bag size: 100, Number of iterations: 100, Maximum depth of the tree: unlimited
SVM	Kernel: radial basis, Gamma: 0.11, Cost: 1

the datasets following the order in which the subjects traced them. It follows that each experimental configuration differs from the others in the classifier and the type and number of samples belonging to the Training set and the Test set.

The implementation of the classifiers are those provided by Weka [18] as well as their parameter values and have not been fine-tuned to provide a baseline comparison among the selected classifiers. The parameter values are the same for all the experiments presented in this paper and are reported in Table 2.

The effectiveness of the classifiers was evaluated in term of *Accuracy*, *Reject*, *False Negative Rate* (FNR), which measures the percentage of healthy people who are identified as PD patients, and *False Positive Rate* (FPR), which measures the percentage of PD patients who are identified as healthy.

For each experimental configuration, a 6-fold cross validation was performed during the training phase and the classifier obtaining the best performance on the validation set, i.e. the one with the smallest values of FNR and FPR and with the greatest value of Accuracy, was selected for classifying individuals in the Test set.

All the experiments were repeated 15 times by shuffling the individuals between Training and Test set. Results obtained on the Test set when it was made up of only meanders, only spirals and both the patterns are reported in Tables 3, 4 and 5, respectively.

As it is evident from the results, all the classifiers exhibit the worst performance on the dataset containing both meanders and spirals, while the best performance is achieved on the meander dataset. Moreover, if we take into account the performance as a function of the number of samples per subject, it is possible to infer that the top performing scenario is represented by the one with the first 3 samples traced by each subject. Finally, by taking into account all the scenarios together, RF results the most performing classifier. In order to evaluate whether the performance in terms of accuracy by RF is significantly different from that provided by other methods, a statistical analysis, based on a non-parametric statistical test [5], is carried out. Following the results reported in the Tables 3, 4 and 5, the analysis has been performed considering the performance achieved by the classifiers on the datasets with only meanders and only spirals.

Table 3. Results obtained on Meanders. The first column reports the number of meanders for each subject.

# of samples	Accuracy (%)		Error (%)		Reject (%)		FNR (%)		FPR (%)	
	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
<i>Decision Tree</i>										
1	72.98	12.29	27.02	12.29	0.00	0.00	32.00	16.41	21.48	20.48
2	69.82	9.31	4.21	3.94	25.96	9.11	6.00	8.00	2.22	4.44
3	85.97	9.15	14.03	9.15	0.00	0.00	24.67	15.43	2.22	4.44
4	72.63	8.42	11.58	5.83	15.79	6.37	22.00	11.08	0.00	0.00
<i>Random Forest</i>										
1	76.84	6.60	23.16	6.60	0.00	0.00	26.67	11.35	19.26	13.12
2	70.88	6.34	5.96	6.34	23.16	5.70	8.67	11.47	2.96	4.91
3	85.96	8.08	14.04	8.08	0.00	0.00	20.00	14.61	7.41	9.66
4	72.98	9.96	8.77	5.32	18.25	7.66	16.00	10.20	0.74	2.77
<i>SVM</i>										
1	75.79	9.76	24.21	9.76	0.00	0.00	40.67	19.48	5.93	6.87
2	62.45	7.16	15.09	8.57	22.46	9.11	28.67	16.28	0.00	0.00
3	72.98	8.78	27.02	8.78	0.00	0.00	51.33	16.68	0.00	0.00
4	58.95	5.83	18.95	8.11	22.11	8.20	35.33	15.43	0.74	2.77

Table 4. Results obtained on Spirals. The first column reports the number of spirals for each subject.

# of samples	Accuracy (%)		Error (%)		Reject (%)		FNR (%)		FPR (%)	
	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
<i>Decision Tree</i>										
1	64.21	10.73	35.79	10.73	0.00	0.00	36.00	13.06	35.55	18.68
2	51.23	10.43	10.53	5.44	38.25	10.61	14.00	10.20	6.67	8.89
3	77.54	5.25	22.46	5.25	0.00	0.00	27.33	13.40	17.04	17.15
4	60.00	7.88	16.14	6.78	23.86	9.96	22.67	13.89	8.89	10.10
<i>Random Forest</i>										
1	65.61	9.58	34.39	9.58	0.00	0.00	37.33	16.52	31.11	19.12
2	57.19	8.78	5.61	4.89	37.19	8.26	4.67	6.18	6.67	10.58
3	84.91	4.65	15.09	4.65	0.00	0.00	12.67	9.29	17.78	12.03
4	68.77	10.61	7.72	3.78	23.51	8.78	6.67	7.89	8.89	9.25
<i>SVM</i>										
1	61.40	8.74	38.60	8.74	0.00	0.00	39.33	25.94	37.78	29.20
2	53.33	7.89	12.63	7.88	34.04	6.04	21.33	16.68	2.96	6.37
3	68.77	9.51	31.23	9.51	0.00	0.00	55.33	21.25	4.44	8.89
4	61.40	8.74	16.49	8.78	22.11	9.06	28.67	16.68	2.96	11.09

3.2 Statistical Analysis

The statistical analysis has been performed through the Friedman Aligned Ranks test, configured with multiple comparison methods. The null hypothesis H0 for the test states equality of medians between the different algorithms. The goal

Table 5. Results obtained on Spirals (S) and Meanders (M). The first column reports the number of samples, equally divided between spirals and meanders, for each subject.

# of samples	Accuracy (%)		Error (%)		Reject (%)		FNR (%)		FPR (%)	
	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
<i>Decision Tree</i>										
2 (1S+1M)	51.23	13.52	15.79	11.04	32.98	14.44	24.00	21.54	6.67	12.03
4 (2S+2M)	67.72	9.39	9.12	6.22	23.16	5.70	13.33	11.35	4.44	7.91
6 (3S+3M)	75.44	9.15	10.53	6.37	14.04	7.36	15.33	14.54	5.18	6.87
8 (4S+4M)	73.68	7.19	15.44	5.91	10.88	8.48	24.67	13.60	5.18	7.98
<i>Random Forest</i>										
2 (1S+1M)	56.49	9.51	9.47	4.79	34.03	10.50	14.00	10.20	4.44	6.79
4 (2S+2M)	77.54	8.48	5.61	6.22	16.84	7.24	8.67	10.24	2.22	4.44
6 (3S+3M)	82.45	8.74	6.67	4.89	10.88	7.05	8.67	8.06	4.44	5.44
8 (4S+4M)	81.75	6.89	8.77	5.66	9.47	6.14	14.00	10.20	2.96	4.91
<i>SVM</i>										
2 (1S+1M)	58.60	9.19	21.05	9.61	20.35	7.66	38.67	19.96	1.48	3.78
4 (2S+2M)	65.96	8.57	18.24	9.39	15.79	5.77	34.67	17.84	0.00	0.00
6 (3S+3M)	69.82	8.03	20.70	7.05	9.47	6.43	39.33	13.40	0.00	0.00
8 (4S+4M)	62.46	6.04	19.65	6.51	17.90	6.32	36.67	13.00	0.74	2.77

Table 6. The ranking of the classification methods according to the Friedman Aligned Ranks test.

	Rank
Random Forest	6.000
Decision Tree	12.625
SVM	18.875

of the test is to either confirm this hypothesis or reject it, at a given level of confidence α . As it is typically done in scientific literature, we have used here $\alpha = 0.05$.

Table 6 reports the ranking of the methods and highlights that RF is the best-performing method followed by DT and SVM. The statistic for Friedman Aligned Ranks with control method is 9.46, distributed according to a chi-square distribution with 2 degrees of freedom, while the p-value is 0.00881. This value is lower than the chosen level of confidence 0.05, which suggests the existence of statistically significant differences among the algorithms considered. Given that H_0 is rejected, meaning that the statistical equivalence among all the algorithms does not hold true, we can proceed with the post-hoc procedures to investigate where these differences between the algorithms exist. When using these procedures, a new null hypothesis H_0' is used, which states the statistical equivalence for couples of algorithms, rather than for the whole set of algorithms as it was for H_0 .

Table 7 reports the adjusted p-values for the post-hoc procedures for Friedman Aligned Ranks test, when RF is used as control method. In the table the generic (i, j) adjusted p-value represents the smallest level of significance that results in the rejection of H_0' between algorithm i and the control method, i.e., the lowest value for which the algorithm i and the control method are not statistically equivalent, for the j -th post-hoc procedure. The lower the adjusted p-value, the more likely the statistical equivalence can be rejected. A very important feature of such a table is that it is not tied to a pre-set level of significance α . Rather, depending on the value of α we choose, the post-hoc procedures will either accept or reject H_0' . Considered that $\alpha = 0.05$, the table says that the statistical equivalence between RF and SVM can be statistically rejected for all the post-hoc procedures apart Hochberg, while there is a statistical equivalence between DT and RF. A general conclusion coming from the statistical analysis is that DT and RF perform better than SVM.

Table 7. Adjusted p-values for the post-hoc procedures for Friedman Aligned Ranks test (RF is the control method).

		Adjusted p-value				
Algorithm	Statistic	Bonferroni-Dunn	Holm	Finner	Hochberg	Li
SVM	3.642	0.00054	0.00054	0.00054	0.06095	0.00029
Decision Tree	1.874	0.12191	0.06095	0.06095	0.06095	0.06095

4 Conclusions

We have presented an approach for the automatic diagnosis of Parkinson's disease that aims at developing a system that is capable of providing an explanation of its behavior in terms that are easy to understand by the clinicians, thus favouring their acceptance/reject of the machine diagnostic suggestions.

In this frameworks, we have chosen as machine learning tool the Decision Tree, as it provides a description of the decision process in terms of if-then rules applied to the feature values, a decision process clinicians are very familiar with, and explicitly establishes a ranking of the features relevance. To illustrate this point, the best performing decision tree obtained by C4.5 on the meander dataset is reported below (Algorithm 1). It shows that the Mean Relative Tremor (x_4) results the most relevant feature, followed by the Maximum HT radius (x_5), the Minimum difference between the HT and ET radius (x_2) and, eventually, the Minimum HT radius (x_6). Those findings are in accordance with the literature, according to which tremor and deviation from a desired trajectory (either coded into the subject motor plan or provided as task) are among the most distinctive features of PD patients handwritten production. Even more important, it shows how the selected features are combined to reach the final decision. In contrast, for the other classifiers considered in this study, the SVM is unable to provide an explanation of its decision, while the RF provides a global ranking of the

Algorithm 1. Best performing model obtained by C4.5 on the Meander dataset.

```

if ( $x_4 > 0.1729$ ) then
  output = "control"
else
  if ( $x_4 > 0.0905$ ) then
    if ( $x_2 > 0.3016$ ) then
      output = "control"
    else if ( $x_6 > 0.0381$ ) then
      output = "control"
    else
      output = "patient"
    end if
  else if ( $x_5 > 0.2849$ ) then
    output = "patient"
  else
    output = "control"
  end if
end if

```

features relevance, but not a description of the way they are intertwined in the decision making process.

Eventually, the experimental results show that the performance of the Decision Tree is comparable to that of the RF and better than the SVM one. As in a previous study comparing SVM with OPF and NB [12] SVM was the top performing classifier, we conclude that the DT proposed here achieves state-of-the-art performance on off-line data, while being the only one capable of describing its decision process in terms that can be simply and naturally understood by clinicians.

The experimental results also suggest that care should be paid in collecting the data to ensure that the task is neither too difficult nor too easy to carry on, so to avoid that fatigue/boredom may introduce misleading data that can negatively affect the performance.

In our future work we will investigate if and to which extent our method can be used to monitor the progress of the disease, as well as to trace back its onset.

Acknowledgment. The work reported in this paper was partially funded by the "Bando PRIN 2015 - Progetto HAND" under Grant H96J16000820001 from the Italian Ministero dell'Istruzione, dell'Università e della Ricerca.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xAI). *IEEE Access* **6**, 52138–52160 (2018)
2. Broderick, M.P., Van Gemmert, A.W.A., Shill, H.A., Stelmach, G.: Hypometria and bradykinesia during drawing movements in individuals with Parkinson's disease. *Exp. Brain Res.* **197**(3), 223–233 (2009)

3. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)
4. De Stefano, C., Fontanella, F., Impedovo, D., Pirlo, G., Scotto di Freca, A.: Handwriting analysis to support neurodegenerative diseases diagnosis: a review. *Pattern Recogn. Lett.* **121**, 37–45 (2019)
5. Derrac, J., García, S., Molina, D., Herrera, F.: A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol. Comput.* **1**(1), 18 (2011)
6. Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint [arXiv:1710.00794](https://arxiv.org/abs/1710.00794) (2017)
7. Ho, T.K.: Random decision forests. In: *Proceedings of the Third International Conference on Document Analysis and Recognition, ICDAR 1995*, vol. 1, pp. 278. IEEE Computer Society, Washington, DC (1995)
8. Hurrell, J., Flowers, K.A., Sheridan, M.R.: Programming and execution of movement in Parkinson's disease. *Brain* **110**(5), 1247–1271 (1987)
9. Jankovic, J.: Parkinson's disease: clinical features and diagnosis. *J. Neurol. Neurosurg. Psychiatry* **79**(4), 368–376 (2008). <https://doi.org/10.1136/jnnp.2007.131045>
10. Marsden, C.: Slowness of movement in Parkinson's disease. *Mov. Disord.* **4**(1 S), S26–S37 (1989)
11. Pereira, C.R., Passos, L.A., Lopes, R.R., Weber, S.A.T., Hook, C., Papa, J.P.: Parkinson's disease identification using Restricted Boltzmann Machines. In: Felsberg, M., Heyden, A., Krüger, N. (eds.) *CAIP 2017*. LNCS, vol. 10425, pp. 70–80. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64698-5_7
12. Pereira, C.R., et al.: A new computer vision-based approach to aid the diagnosis of Parkinson's disease. *Comput. Methods Programs Biomed.* **136**, 79–88 (2016)
13. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco (1993)
14. Senatore, R., Marcelli, A.: A paradigm for emulating the early learning stage of handwriting: performance comparison between healthy controls and Parkinson's disease patients in drawing loop shapes. *Hum. Mov. Sci.* (2018). <https://doi.org/10.1016/J.HUMOV.2018.04.007>
15. Stelmach, G.E., Teasdale, N., Phillips, J., Worringham, C.J.: Force production characteristics in Parkinson's disease. *Exp. Brain Res.* **76**(1), 165–172 (1989)
16. Tucha, O., et al.: Kinematic analysis of dopaminergic effects on skilled handwriting movements in Parkinson's disease. *J. Neural Transm.* **113**(5), 609–623 (2006)
17. Van Gemmert, A.W.A., Adler, C.H., Stelmach, G.E.: Parkinson's disease patients undershoot target size in handwriting and similar tasks. *J. Neurol. Neurosurg. Psychiatry* **74**(11), 1502–1508 (2003)
18. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington (2016)
19. Yao, J.F.F., Yao, J.S.: Fuzzy decision making for medical diagnosis based on fuzzy number and compositional rule of inference. *Fuzzy Sets Syst.* **120**(2), 351–366 (2001). [https://doi.org/10.1016/S0165-0114\(99\)00071-8](https://doi.org/10.1016/S0165-0114(99)00071-8)
20. Zham, P., Arjunan, S.P., Raghav, S., Kumar, D.K.: Efficacy of guided spiral drawing in the classification of Parkinson's disease. *IEEE J. Biomed. Health Inform.* **22**(5), 1648–1652 (2018)