Classification of Patients with Parkinson's Disease Using Free Handwriting Features Collected through a Smart Ink Pen

Simone TOFFOLI[*a], Francesca LUNARDINI*[b], Monica PARATI[a,c], Matteo GALLOTTA[c], Manuel MULETTI[c], Chiara BELLONI[a], Maria Elisabetta DELL'ANNA[c], and Simona FERRANTE[a]
[a] *Department of Electronics, Information, and Bioengineering, Politecnico di Milano*
Via Ponzio, 34/5
20133, Milano, ITALY
[b] *Child Neuropsychiatry Unit, Fondazione IRCCS Istituto Neurologico Besta*
Via Giovanni Celoria, 11
20133, Milano, ITALY
[c] *Istituti Clinici Scientifici Maugeri IRCCS*
Via Camaldoli, 64
20138, Milano, ITALY
simone.toffoli@polimi.it, francesca.lunardini@istituto-besta.it, monica.parati@polimi.it, matteo.gallotta@icsmaugeri.it, manuel.muletti@icsmaugeri.it, chiara1.belloni@mail.polimi.it, me.dellanna@libero.it, simona.ferrante@polimi.it

Abstract. Systems for monitoring of Parkinson's disease (PD) patients, able to complement clinical assessment, are needed. These solutions should be objective, based on technology that captures physical characteristics of the pathology, and capable of providing frequent measures conducted both on-site and remotely. Since one of the most typical clinical hallmarks of PD is handwriting deterioration, we devised an innovative smart ink pen for quantitative and reliable handwriting monitoring, without altering the natural writing conditions. 30 PD patients and 30 age-matched controls performed two unconstrained writing tasks (free text and grocery list) with the smart ink pen. A series of 47 writing and tremor indicators were computed and used to classify patients from age-matched controls. Catboost and Logistic Regression classifiers were used, and the SHAP model explanation technique was applied to explore the contribution of the features in the classification. Very good performances were obtained through the Catboost classifier when combining features extracted from both tasks (Accuracy: 93%, Precision: 96%, Recall: 90%; F1: 93%; AUC: 98.9%). We achieved a classification performance in line with previous work, with two main advantages: writing data acquisition through an ink pen used on common paper, and proposition of an unconstrained protocol mimicking daily-life writing.

1. Introduction

Parkinson's disease is the second most common neurodegenerative disease. It affects 7 million people and cases are expected to double in the next 20 years (Dorsey et al., 2018). In its most typical phenotype, PD cardinal motor symptoms are tremor at rest, bradykinesia, rigidity, and postural instability (Jankovic, 2008). As first observed by J. Parkinson (Parkinson, 1817), when motor symptoms affect the dominant limb, patients may report worsening of handwriting as one of the initial signs. Furthermore, changes in handwriting, such as writing akinesia, may even precede the appearance of gait akinesia. Thus, handwriting analysis might be a critical tool in PD monitoring.

Today, disease assessment is mainly clinical, inferential, and based on rating scales. In addition, pen-and-paper handwriting examinations are commonly performed during clinical assessments (Alty & al., 2017). Notwithstanding the importance of clinical judgement, this approach suffers some limitations since it is subjective, dependent on patients' memories, and infrequent, providing only a snapshot of the patient's condition (AlMahadin et al., 2020). These aspects explain the importance of developing support systems for the monitoring of PD patients able to complement the clinical assessment. In turn, these tools should be objective, and capable of providing frequent measures conducted both on-site and remotely. Against this background, the use of technologies such as digitizers and tablets has been introduced to objectively study handwriting characteristics, thus enriching the qualitative clinical examination based on the classic pen-and-paper approach. The use of such technology has made possible the quantitative study of pressure and kinematic features of the writing tasks, resulting in a surge of research work successfully investigating graphomotor impairment in patients with PD (Drotár & al., 2015, 2016).

However, some limitations hampered their adoption both at clinical premises and remotely at the patient's home. These limits include the undermined naturalness of the writing gesture performed on the small and frictionless surface of a tablet. Moreover, its use may not be straightforward - particularly when dealing with patients and elder users - often requiring technical support of an operator. To combine the advantages of the digitizer technology with the natural "feel" and simplicity of the traditional pen-and-paper approach, we devised a novel smart ink pen instrumented with force and motion sensors (Lunardini & al., 2021). The smart ink pen was

[*] These authors equally contributed to this work.

designed to achieve quantitative and reliable monitoring without altering the natural writing conditions. For this reason, it completely replaces a daily-life object: i) it looks like a common pen; ii) allows writing on plain paper; iii) activation is automated upon use; iv) no supervision is required. These factors are key to facilitate its adoption both in clinical practice - where ease and speed are essential for the strict time constraints - and in home monitoring - where transparency and intuitiveness are crucial for technology acceptance. To adapt to both scenarios, the data analysis software is implemented to analyse standard writing tests performed during clinical assessment, like Archimedes' Spiral (Toffoli et al., 2021), and free handwriting typically executed during daily-life.

In this work, 30 PD patients and 30 age-matched healthy subjects performed two types of free writing tasks with the smart ink pen. From the raw data, a series of handwriting and tremor indicators were computed and used to classify patients (class 1) from age-matched controls (class 0).

2. Materials and Methods

2.1 The Smart Ink Pen

Tasks were performed using the smart ink pen (Lunardini & al., 2021). The signals (3D acceleration, 3D angular velocity, force) are sampled at 50Hz, saved on the pen on-board memory, and downloaded offline for analysis.

2.2 Participants and Protocol

Patients affected by PD and age-matched healthy subjects were included in the study, after the signing of an informed consent. Patients were recruited and evaluated (Unified Parkinson's Disease Rating Scale (UPDRS) and Mini Mental State Examination (MMSE)) by the IRCCS Istituti Clinici Scientifici (ICS) Maugeri (Milan, Italy). Exclusion criteria: MMSE < 24; presence of disorder (other than PD) affecting handwriting skills. Politecnico di Milano (Milan, Italy) enrolled the control group. The same exclusion criterium for MMSE was considered, coupled with the absence of diseases. The protocol, approved by the Ethical Boards (2457 CE 06/07/2020 ICS Maugeri; n. 10/2018 Politecnico di Milano), consisted in writing a 7-element grocery list and a 6-line free text on a sheet of paper, with the dominant hand. The content was left free to the subject, to resemble daily handwriting one can perform in the home setting. Being this the first work employing the pen with patients, the acquisition was managed by an operator through an iOS application in a supervised scenario.

2.3 Feature extraction

In MATLAB® 2020b, *writing* and *tremor* features extraction was performed. 32 *writing* indicators – divided into 4 categories - were computed: i) *Time*. As for pauses, we retained the total and relative numbers, the duration mean and coefficient of variation (CV). We computed the mean in-air time and on-sheet time, and their CVs. The air-sheet ratio (in-air time divided by on-sheet time) and the on-sheet ratio (time spent on-sheet normalized by the task duration) were extracted. The number of strokes per second was also computed. ii) *Force*. We retained the mean force exerted on the writing surface. Then, indicators related to force variability were computed: force CV, mean and CV of the difference between consecutive force extrema. iii) *Tilt*. We computed mean and CV of the pen tilt angle during writing. iv) *Smoothness*. The number of inversions was extracted for force, acceleration (*NCA*), and angular velocity. Irregular oscillations were measured by the logarithmic dimensionless jerk (LDLJ) for acceleration (*LDLJ_A*), the LDLJ and Spectral Arc Length (Balasubramanian & al., 2012) for angular velocity.

Fifteen *tremor* indicators (2 categories) were computed from the acceleration and angular velocity signals: v) *Tremor amplitude* was quantified in the time domain through the root mean square of the signal for acceleration (*RMS_A*) and angular velocity, also including the RMS for the signal filtered around the spectral peak for angular velocity (*RMS_G_Peak*), and mean (*Cons_Peak_Diff_G_Mean*) and CV of the difference between consecutive angular velocity extrema. As for the frequency domain, after estimating the power spectral density, the peak frequency was retained and the relative power distribution was evaluated in the following frequency bands for acceleration: 0-2Hz, 4-7Hz and 8-12Hz (*RPW_f1-f2*). For angular velocity, the relative power was evaluated in a window of 1Hz around the peak. vi) *Tremor regularity*. Acceleration tremor was inspected by means of entropy, stability index (*TSI_DB* and *TSI_Luft*) (Di Biase & al., 2017; Luft & al., 2019), mean harmonic power (*MHP*) (Muthuraman, & al., 2011). Following the approach in (Oung & al., 2018), Shannon Entropy was applied to the decomposition coefficients and the instantaneous amplitude of the angular velocity tremor. Its variability over time was also estimated.

2.4 Classification

We investigated the ability of unconstrained handwriting to discriminate subjects belonging to the patient (class 1) or to the control (class 0) groups through machine learning (ML) classification techniques.

We define D_T the Text dataset, and D_L the List dataset, both composed of 60 samples and 48 attributes (47 indicators and the group label). Merging D_T and D_L, we built the D_{TL} dataset composed of 60 samples and 95 attributes (94 indicators - 47 for each task - and the group label). Two ML algorithms were used to compare different classification logics: i) logistic regression, to set a baseline performance measure since it is one of the

simplest and most used linear classifiers; ii) a more recent boosting algorithm, Catboost, since it is known to achieve notable performance while avoiding data overfitting even with small datasets (Prokhorenkova & al., 2018). After data standardization, for each dataset, we evaluated the models through Accuracy, Precision, Recall, F1, and Area Under the ROC Curve (ROC-AUC). For both models, we used scikit-learn default parameters (Leave-One-Out Cross Validation (LooCV) and early stopping set at 100 epochs). The LooCV technique returns an estimate of performance on unlearned data and the best number of learning iterations for each model.

2.5 Model explanation

We used a model explanation technique to overcome the limitations of the black-box nature of the Catboost algorithm and gain insights into model decisions. We used SHAP (Lundberg & al., 2017) a model explanation library that computes the Shapley Values (Roth, 1988) of the features according to their impact on its predictions. If a feature has a positive impact, it influences the prediction favouring class 1, and vice versa. This step was useful to understand how much each indicator has an effect of the prediction to class 0 or 1.

3. Results and Discussion

A total of 30 PD patients (14 Males; 30 Right-handed; Age = 72.8 (mean) ± 7.40 (SD) yo; MMSE = 27.67 ± 1.68; Years since Diagnosis = 7.3 ± 4.86; UPDRS III score = 19.50 ± 7.75) and 30 age-matched healthy controls (11 Males; 30 Right-handed; Age = 72.7 ± 8.48 yo; MMSE = 28.07 ± 1.56) participated in the study.

Table 1: Performance Metrics for Logistic Regression and Catboost Classifier

	Accuracy [%]			Precision [%]			Recall [%]			F1 [%]			ROC-AUC [%]		
	D_T	D_L	D_{TL}	D_T	D_L	D_{TL}	D_T	D_L	D_{TL}	D_T	D_L	D_{TL}	D_T	D_L	D_{TL}
Logistic Regression	**78.3**	75	**78.3**	**79.3**	75.9	**79.3**	**76.7**	73.3	**76.7**	**78**	74.6	**78**	**87**	79.3	84.1
Catboost	83	88	**93**	86	93	**96**	80	83	**90**	83	88	**93**	96.3	97.2	**98.9**

Table 1 reports the performance metrics for each dataset (D_T, D_L, D_{TL}) for the Logistic Regression and the Catboost Classifier, evaluated through LooCV for the binary classification between the patient and the control groups. For each model, the best scores are highlighted in bold. As expected, the Catboost algorithm achieved the best performance for all datasets. In addition, we can notice that the best scores were obtained when merging the indicators extracted from the two writing tasks (D_{TL} dataset).

Since the best scores were obtained from the Catboost algorithm applied to the D_{TL} dataset, a closer look to the classification performance achieved by combining the two tasks is presented in Figure 1. Panel (a) reports the confusion matrix; (b) presents the SHAP feature ranking. Panel (b) lists features in decreasing order of importance and explains how the learned samples were predicted according to the values of their features. Each dot represents the Shapely Value of the indicator for a particular sample: negative Shapley values push the prediction towards class 0 (control group), while positive values push toward class 1 (PD group). The value of the indicator is represented in a blue-red colour scale (low: blue; high: red).

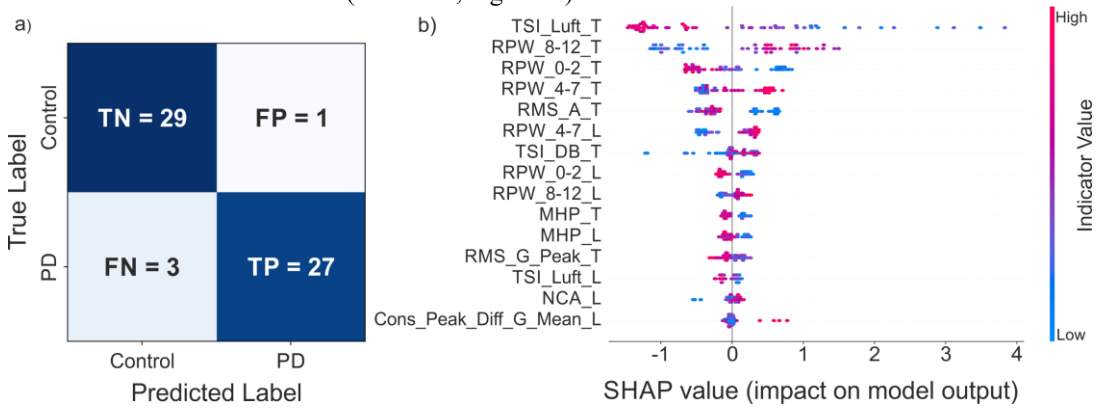

Figure 1: Classification performances and model explanation plots for the Catboost algorithm applied to D_{TL}.

As can be noted from Table 1 and Figure 1 - (a), the classification achieves a notable performance, with only 1 FP and 3 FN. The result is interesting if compared to other ML classification studies, which typically employ standardized, content-constrained tests, performed on a digitizer. In (Drotár & al., 2015, 2016) features derived from pre-defined loops, bi-grams, tri-grams, words and sentences were combined to solve the classification

problem, achieving accuracy from 82 to 88%. From the SHAP feature ranking (Figure 1 - (b)), we can notice that both list ("_L") and text ("_T") indicators contribute to the classification. As for the types of indicators, we can see how *tremor* indicators play a more important role. A possible explanation may be found in the nature of the protocol. The Tremor Stability Index (*TSI*) during text writing emerged as the most important feature; it captures the regularity of the tremor frequency over time and, in line with literature, lower values (higher tremor regularity) pushed the classification toward the PD group. For both tasks, the Tremor Relative Power in the 4-7Hz Band (*RPW_4-7*) emerges in the feature ranking: higher values favour the classification toward the PD group; indeed, PD tremor frequency is reported in this band (Elble, 1996). On the other hand, the greater Relative Power in the band associated with voluntary motion (*RPW_0-2*) was associated with the classification in class 0. As for *writing* indicators, the presence of the Number of Changes in Acceleration (*NCA*) suggests that a reduced smoothness pushed the prediction toward the PD group.

To sum up, this is the first classification work based on indicators derived exclusively from unconstrained handwriting, acquired with an innovative smart ink pen. We achieved a classification performance in line with previous work, with the undoubtable advantages of 1) acquiring writing data through an ink pen writing on common paper – thus ensuring the naturalness of the gesture – and 2) proposing an unconstrained protocol mimicking daily-life writing. These aspects pave the way toward remote daily-life handwriting monitoring in patients with PD.

References

AlMahadin, G., Lotfi, A., Zysk, E., Siena, F. L., Carthy, M. M., & Breedon, P. (2020). Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective. *BMC Neurology*, *20*(1), 1–13. https://doi.org/10.1186/s12883-020-01996-7

Alty, J., Cosgrove, J., Thorpe, D., & Kempster, P. (2017). How to use pen and paper tasks to aid tremor diagnosis in the clinic. *Practical Neurology*, *17*(6), 456–463. https://doi.org/10.1136/practneurol-2017-001719

Balasubramanian, S., Melendez-Calderon, A., & Burdet, E. (2012). On the analysis of movement smoothness SPARC: A modified SPectral ARC length. *IEEE Transcations Biomedical Engineering*, *59*(8), 2126–2136.

Di Biase, L., Brittain, J. S., Shah, S. A., Pedrosa, D. J., Cagnan, H., Mathy, A., … Brown, P. (2017). Tremor stability index: A new tool for differential diagnosis in tremor syndromes. *Brain*, *140*(7), 1977–1986. https://doi.org/10.1093/brain/awx104

Dorsey, E. R., Elbaz, A., Nichols, E., Abbasi, N., Abd-Allah, F., Abdelalim, A., … Murray, C. J. L. (2018). Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *17*(11), 939–953. https://doi.org/https://doi.org/10.1016/S1474-4422(18)30295-3

Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., & Faundez-Zanuy, M. (2015). Decision support framework for Parkinson's disease based on novel handwriting markers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *23*(3), 508–516. https://doi.org/10.1109/TNSRE.2014.2359997

Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., & Faundez-Zanuy, M. (2016). Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. *Artificial Intelligence in Medicine*, *67*, 39–46. https://doi.org/10.1016/j.artmed.2016.01.004

Elble, R. J. (1996). Central mechanisms of tremor. *Journal of Clinical Neurophysiology*, *13*(2), 133–144. https://doi.org/10.1097/00004691-199603000-00004

Jankovic, J. (2008). Parkinson's disease: Clinical features and diagnosis. *Journal of Neurology, Neurosurgery and Psychiatry*, *79*(4), 368–376. https://doi.org/10.1136/jnnp.2007.131045

Luft, F., Sharifi, S., Mugge, W., Schouten, A. C., Bour, L. J., van Rootselaar, A. F., … Heida, T. (2019). A power spectral density-based method to detect tremor and tremor intermittency in movement disorders. *Sensors (Switzerland)*, *19*(19). https://doi.org/10.3390/s19194301

Lunardini, F., Febbo, D. DI, Malavolti, M., Cid, M., Serra, M., Piccini, L., … Ferrante, S. (2021). A Smart Ink Pen for the Ecological Assessment of Age-Related Changes in Writing and Tremor Features. *IEEE Transactions on Instrumentation and Measurement*, *70*. https://doi.org/10.1109/TIM.2020.3045838

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.

Muthuraman, M., Hossen, A., Heute, U., Deuschl, G., & Raethjen, J. (2011). A new diagnostic test to distinguish tremulous Parkinson's disease from advanced essential tremor. *Movement Disorders*, *26*(8), 1548–1552. https://doi.org/10.1002/MDS.23672

Oung, Q. W., Muthusamy, H., Basah, S. N., Lee, H., & Vijean, V. (2018). Empirical Wavelet Transform Based Features for Classification of Parkinson's Disease Severity. *Journal of Medical Systems*, *42*(2). https://doi.org/10.1007/s10916-017-0877-2

Parkinson, J. (2002). An essay on the shaking palsy. 1817. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *14*(2), 223–236; discussion 222. https://doi.org/10.1176/jnp.14.2.223

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *2018-Decem*(Section 4), 6638–6648.

Roth, A. E. (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

Toffoli, S., Lunardini, F., Parati, M., Gallotta, M., Maria, B. De, Anna, E. D., & Ferrante, S. (2021). A smart ink pen for spiral drawing analysis in patients with Parkinson ' s disease. In *EMBC 2021* (pp. 6475–6478).