Machine Learning-based Classification of Simple Drawing Movements in Parkinson's Disease

C. Kotsavasiloglou, N. Kostikis, D. Hristu-Varsakelis, and M. Arnaoutoglou

***Abstract*—**

This work explores the use of a pen-and-tablet device to study differences in hand movement and muscle coordination between healthy subjects and Parkinson's disease patients. We let volunteers draw simple horizontal lines and recorded the trajectory of the pen's tip on the pad's surface. The signals thus obtained were then processed to compute various features which correspond to the variability of the pen tip's velocity, the deviation from the horizontal plane, and the trajectory's entropy. Our goal was to establish simple and objective metrics which can be used to differentiate between normal and pathological movement. In a small-scale clinical trial, 44 age-matched subjects were divided in two groups, namely 20 healthy subjects (H), and 24 Parkinson's disease (PD) patients. We applied a comprehensive machine learning approach to build a model that could classify unknown subjects based on their line-drawing performance. We were able to achieve an average prediction accuracy of 91% (88% sensitivity [TP], 95% specificity [TN]). Our results show that the proposed method is a good candidate for differentiating between healthy and Parkinson's disease individuals, and shows promise in the context of telemedicine applications and tracking of the disease's symptoms via inexpensive, widely available hardware.

***Index Terms*— Movement disorders; Parkinson's disease; handwriting; machine learning; normalized velocity.**

I. Introduction

PARKINSON'S disease (PD) is one of the most representative and frequently encountered movement disorders; it is a neurodegenerative disorder with initial clinical features caused predominately by the loss of dopaminergic function in the substantia nigra pars compacta in the midbrain. PD is the second most common neurodegenerative disorder after Alzheimer's disease, and affects more than 1% of individuals over 55 and more than 3% of those over 75 years of age [1]. The cardinal features establishing a PD diagnosis are bradykinesia, tremor, rigidity and postural instability [2]. The disease includes non-motor symptoms as well which are not addressed directly in this study. The evaluation of patients' clinical status and response to medication is currently achieved via clinical assessment (neurological examination, clinical assessment scales). The careful application of diagnostic criteria, such as tremor, bradykinesia and rigidity derived from existing clinicopathologic studies can increase the positive predictive value of diagnosis to over 95% [3]. Nevertheless, a purely clinical assessment of the disease is inevitably a subjective procedure, and although additional factors can be used to increase the certainty of diagnosis, maximizing the specificity of the criteria leads to a significant decrease in the sensitivity of the diagnosis, sometimes excluding as many as one-third of "true" cases [4].

These considerations have led to various efforts aimed at quantifying aspects of the motor system and its disorders. The primary laboratory method used for the evaluation of motor disorders (including PD) is EMG, which involves recording the electrical activity of muscle fibers [5]. However, over the last few years, there has been a significant volume of research on alternative, simpler and user-friendlier devices used to measure aspects of movement disorders. One prominent example involves small-sized accelerometers [6], [7], [8], which are mounted on the patient's limbs and record during rest or while the patient executes a specified movement. Accelerometers have not yet transitioned into clinical practice; however, their widespread availability, in most smartphones, has led to increased research activity in that area (e.g., see [9] and references within).

C. Kotsavasiloglou, Thessaloniki, Greece (chkot@cteam.gr)

N. Kostikis is the corresponding author of this manuscript. He is with the Department of Applied Informatics, University of Macedonia, Thessaloniki, 54006, Greece (e-mail: nikkostikis@gmail.com).

D. Hristu-Varsakelis is with the Department of Applied Informatics, University of Macedonia, Thessaloniki, 54006, Greece (e-mail: dcv@uom.gr, tel. +30-2310891721).

M. Arnaoutoglou is with the School of Medicine, Aristotle University of Thessaloniki, 54124, Greece (e-mail: marnaout@med.auth.gr).

Other hardware used for the evaluation of motor disorders includes electronic tablet devices. They are inexpensive and do not require expertise in order to use. In [10] the authors used spectral analysis to identify PD tremor expressed as moment-to-moment fluctuations in the pen's position during a shape-tracing task. In [11] the hypometria and bradykinesia of PD patients was identified mainly during writing tasks with longer strokes. Aiming movements were performed on a digitizer tablet by the authors of [12], where it is shown that the decay in PD patients' positional sense affects the motor planning process. In [13], dopamine depletion in PD was found to affect sensorimotor adaptation.

The authors of [14] showed that spatial and temporal characteristics extracted from predefined handwritten text strokes can objectively discriminate between PD patients and healthy individuals; a later study [15] came to solidify those results, comparing results from micrographia (the abnormal reduction in the size of written text, associated with PD patients) versus spatiotemporal and kinematic variables extracted from digitized handwriting, defined generally as dysgraphia. The authors of [15] argue that with the advent of graphic tablets researchers should focus even more on dysgraphia, which pertains to all handwriting deficits characterizing PD, not only on the traditional measure of writing size. In other studies [16], [17] pressure and in-air trajectory during handwriting are proven to be good potential markers for PD binary classification.

Writing is a skill that is developing later in a child's life. It involves a complex feedback system, integrating continuous information from the writing hand, i.e., proprioceptive sensory stimuli from all muscles involved, and sensory information from the fingers and the visual system. Also, writing is a task that implicates the participation of various degrees of cognitive processes. For this reason the majority of the related papers used tasks such as the Archimedes spiral [18], single letters or simple words, in order to investigate purely motor aspects of handwriting. The work in [19] took a different approach and showed that simpler tasks, namely drawing point-to-point trajectories, contained useful features that helped detect motor blocks in PD patients.

This paper's contribution is two-fold. First, we explore the use of a simple line-drawing task for classifying PD subjects using machine learning techniques with good results in terms of classification accuracy. The task in question is simpler than, for example, writing or drawing spirals, lasts only a few seconds, involves fewer muscular systems, no cognitive effort, and low coordination control effort. As it turns out, it is also unaffected by the dexterity of the participant's dominant hand, making it easy to perform using either hand, something which is not true in the case of writing letters or words. Second, our classification model is aided by the introduction of a new metric that characterizes the variability of the subject's drawing velocity and which, as we will see, is "rich" in information, more so than other "standard" markers that have been used in similar studies.

More specifically, we investigate the kinematics of hand motion during line drawing task, measuring hand movement at a timescale where there is no conscious control of the motion, so that what we detect is the balance of the tone of agonist versus antagonist muscular systems. This balance is altered in Parkinson's disease and in other pathological conditions. Unlike other studies on the subject, we aimed to make the task as simple as possible for the participants, and thus had them draw simple horizontal lines, hypothesizing that any imbalance in agonist-antagonist coordination should be present even in simple drawings. Our results validate that hypothesis, as we shall see. Furthermore, with lines, participants can draw starting from either side of the writing surface, extending or flexing their arms, allowing us to check the performance of two different groups of muscular systems.

Our hypothesis is that due to impaired coordination in patients with movement disorders, certain features, such as the velocity variability of the pen's tip (to be made precise shortly) or the "excursions" from the horizontal, should be more pronounced compared to healthy subjects. Towards that end, our approach involves recording the position of the pen on the tablet and computing a vector of metrics, namely the Normalized Velocity Variability (NVV), the velocity's Standard Deviation (SDV) and Mean (MV), and the signal Entropy (ETP), to characterize the pen's trace spatiotemporally. Using data acquired from healthy and PD individuals, we test the hypothesis that these metrics are statistically different between the two groups. Our hypothesis is based on research proving that velocity- and acceleration-based metrics of voluntary movements can separate pathological from healthy subjects. This includes [11] where PD patients showed a reduced ability to modulate acceleration, leading to smaller than required movements and micrographia, and [20] which showed that that dopamine depletion in PD leads to smaller than normal pallidothalamic gating signals, which in turn affect the ability to control variable movement speed. The variability in handwriting velocity in patients with PD was also noted in [21] where the patients showed multiple peaks in their velocity signal whereas the controls showed just one peak. In [22], the velocity and acceleration profiles of PD patients were different in relation to healthy subjects while writing circles. In a different experimental design investigating the effect of the dopamine on handwriting movements [23], researchers found lower values for maximum and minimum velocity in ascending and descending strokes in PD patients than in healthy subjects. They also found that patients had significantly more inversions of velocity and acceleration than healthy people. There was also a difference between patients on medication and off medication, where the number of inversions in velocity and acceleration was statistically significant.

The remainder of this paper is organized as follows: In Section II we describe our experimental setup. In Section III we present the data analysis and introduce a machine learning model for classifying subjects. We conclude our paper in Section IV.

TABLE I
INFORMATION FOR THE SUBJECTS' AGE & GROUPING

Group (according to health & age)	Group size	Age Statistics	
		Mean±StDev	StError
YH	15	36.40±5.94	1.53
H	20	66.35±7.91	1.61
PD	24	70.91±5.74	1.17
Total	59		

II. EXPERIMENTAL SETUP

A. Subjects

Fifty-nine subjects in total participated in this study. All were right-handed, and had normal or corrected-to-normal vision. Their right-handedness was established based on what hand they used to write and eat with, as well as an evaluation of the muscular force of their two hands. The subjects agreed to participate in this study after a detailed explanation of its purposes and procedures. They were divided in three groups based on their health status and age (Table I).

TABLE II
INFORMATION FOR THE PD SUBJECTS

Number	Age	Sex	UPDRS*	H&Y	PD Duration (Years)**
1	70	F	9	2	2
2	61	M	15	3	2
3	63	F	13	2	2
4	66	F	29	2.5	2
5	75	F	8	2	2.5
6	65	F	9	2	1.5
7	78	M	23	3	3
8	72	M	28	3	3
9	69	M	15	2	3
10	72	M	29	2	3
11	72	M	19	2	2.5
12	70	M	1	2	2
13	75	M	11	2.5	2
14	67	M	20	2	2
15	72	M	25	3	4
16	80	M	10	3	5
17	66	M	7	3	2
18	73	M	16	4	15
19	74	M	23	4	14
20	78	M	10	2	2.5
21	79	F	14	2.5	3
22	76	F	8	2.5	2
23	71	F	4	2	2
24	58	M	4	2	2

*This represents a modified UPDRS III total score, where only the components regarding upper limb symptoms have been included.
**Time from first diagnosis.

Group H included 20 healthy persons aged 56-89. All subjects had a detailed neurological examination in order to screen for any movement disorders that would exclude them from the study. None had a first-degree relative with PD or some kind of tremor. Also, none had hypertension or diabetes. Most of the healthy subjects came from retirement facilities.

Group PD had 24 subjects aged 58-80, all under medication. They were recruited from the Parkinson's disease outpatient clinic of the 1st Neurology Department of the Aristotle University of Thessaloniki, Greece. All had been under periodic evaluation and levodopa and/or dopamine agonist treatment for more than a year. Undeniably, patients who are on medication improve on some of their clinical signs and symptoms. Nevertheless, even when a patient is on medication, most PD signs and symptoms never disappear. The symptoms that are mostly alleviated are bradykinesia and rigidity, whereas tremor in most cases is "drug-resistant" [24]. Because PD subjects participating in this study were all tested in an outpatient setting, they were kept on

medication for ethical and safety reasons (i.e., drug deprivation could lead to injuries). Additionally, most of our patients were not newly-diagnosed (their Hoehn and Yahr rating was above 1) and despite treatment their signs and symptoms were present. Their information is provided in Table II. We introduced a third group of 15 young healthy volunteers (YH) with a mean age 36.40 years. They were included in this study to establish a "baseline" for the various metrics, and their scores were compared to those of older healthy volunteers (group H) to test for possible age effects.

B. Experimental Procedure

Our hardware setup was based on a commercially available Wacom pen-tablet device, model Bamboo CTE-450, although any digital tablet would be suitable as well. Ours has an active surface of 147.6 x 92.3 mm and a resolution of 100 dots per mm. The tablet connects to a PC through a USB port. We used MATLAB to create a custom software tool for data collection. The raw data consisted of the coordinates of the pen's tip (measured in pixels), recorded at a rate of 60 Hz. Given the time scale involved in the movement of the upper limbs (including the hand/wrist) and the manifestations of PD which are in the under-10 Hz range [25], anything above a 20 Hz sampling rate seems to be sufficient to capture both regular and PD impaired motion.

In each session, the subject sat at a table with the tablet placed at under-chest level, at a position comfortable for drawing. Each subject was instructed to draw a horizontal line on the tablet's surface, keeping the pen's velocity as constant as possible. While drawing, the PC's screen would turn black and the digital trace of the line drawn would appear on the screen to provide visual feedback. Drawing lines at a constant speed is more natural than other velocity profiles. The effects of PD are dominant particularly at rest and during steady movements and that is why we encouraged our volunteers to try for a constant drawing speed. All subjects went through a practice phase where they drew a few lines and familiarized themselves with the procedure. The movement was such that the hand was away from the body, and the forearm was not supported by the table or the pad; the only contact with the pad's surface was through the pen, and subjects drew in such a way that their wrist was fixed in relation to the forearm. The trace of each line was shown on the PC's screen. The overall positioning of the arm and forearm was the same for all persons. Every participant drew a set of at least 10 horizontal lines from left to right and another 10 from right to left with each hand, i.e., a total of 40 movements per participant. The lines drawn had a length of approximately 145 mm (determined by the working area on the pad). We opted to discard the initial and final data (150 pixels at the start and the end of the line drawn), taking into consideration only the middle 125 mm of the line, because we wanted to focus on the intentionally steady portion of the movement without the initial/final acceleration/deceleration effects.

Volunteers were instructed to draw the line in about 2 seconds. In reality, their drawing times fell between 1.5 and 3.5 seconds. Having cut the initial (acceleration) and final (deceleration) portions of the line (150 pixel at the start and end) resulted in 1.47sec average time for the H group and 3 seconds for the PD group.

C. Data and Processing

We proceed to describe the calculations involved in quantifying each subject's performance by a set of five metrics, which will be referred to as their *score vector*. Let $(t_i,\ x_i,\ y_i)$, i=1…N be the sequence of samples in a subject's digitized path on the tablet along the horizontal and vertical direction, at temporal intervals $\Delta t = t_i - t_{i-1}$ and total duration $T = t_N - t_1$. We computed the pen's horizontal velocity along the path by taking first differences, $v_i = (x_i - x_{i-1})/\Delta t$, and characterized the path by the following metrics:

$$MV = \frac{1}{N}\sum_{i=1}^{N} v_i \quad (1)$$

$$NVV = \frac{1}{T|MV|}\sum_{i=1}^{N-1}|v_{i+1} - v_i| \quad (2)$$

$$SDV = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|v_i - MV|^2} \quad (3)$$

$$ETP_x = -\sum_{i=1}^{N} P(x_i)\log_2\big(P(x_i)\big) \quad (4)$$

$$ETP_y = -\sum_{i=1}^{N} P(y_i)\log_2\big(P(y_i)\big) \quad (5)$$

where (1) is the mean horizontal velocity, (2) is the normalized velocity variability, in units of 1 per second, (3) is the standard deviation of the horizontal velocity, and (4) and (5) are the entropies of the horizontal and vertical components of the signal, with

the histograms of the x-component, $P(x_i)$, or y-component, $P(y_i)$, respectively, being the estimators of the signal's probability density function. To calculate the histograms, we used MATLAB's *hist* function, which automatically creates "bins" of uniform width, depending on the signal's element range and distribution shape.

In general, as expected, and observed in our experiments, smoother movements incur a lower NVV and SDV compared to more irregular movements (e.g., a trajectory with constant speed in the x-direction would have NVV and SDV equal to zero).

For each line drawn we calculated the score vector (from (1)-(5)). Then, subsets of each subject's score vectors were averaged in six ways, resulting in different metrics' components: ALL (mean over all of a subject's efforts), LH and HH (mean over lowest-scoring and highest-scoring hand, respectively), HLH=HH-LH (difference between highest- and lowest-scoring hand), LD and HD (mean over lowest-scoring and highest-scoring movement direction i.e. extension or flexion, respectively). When no component is subscripted, ALL is implied.

In the next section we begin with a preliminary analysis of the data collected, in an effort to determine the basic statistical patterns of the metrics calculated from (1)-(5). The significance of these metrics in distinguishing healthy from PD subjects will be established later on, during feature selection for a machine learning model.

III. Analysis and Results - The Prediction Model

The NVV metric defined in (2) is novel and less explored in the context of movement disorders compared to the other metrics and handwriting markers, which have been used with success in [22], [23], [25], and [26]. At a neurological functional level, we expected differences in NVV scores between PD an H volunteers, based on the fact that the substantia nigra is one of the key structures which participate in the regulation of the muscular tone [27]. The NVV attempts to capture some expression of the balance, or lack thereof, of the muscular tone between opposing muscular systems, given the fact that low-level control of those systems occurs on a time scale which is on the order of milliseconds, while conscious control of movement cannot be done at such high a frequency.

Ultimately, the classification analysis (later in this Section) should confirm these expectations. Before using the NVV as a potential classification feature, we conducted a preliminary statistical analysis of the NVV scores between the H and PD populations.

A. NVV Validity - Preliminary Analysis

We conducted normality (Jarque-Berra and Liliefors tests), homoscedasticity (Levene and Brown-Forsythe tests) and means testing, both parametric (T-Test) and non-parametric (Mann-Whitney), depending on normality, within the PD and H populations, and their components, individually, as well as between them. The p-value for all tests was set to 0.01.

We found NVV scores to be consistent in multiple sessions of single subjects: Seven of the H subjects were re-tested a second or third time, over a period of several days. In six of the seven cases, means testing revealed no statistically significant differences in the subjects' means over different testing sessions. The one volunteer who showed significant difference between the two sessions reported sleeping for only three hours the night before the second test session. Sleep deprivation is known to favor abnormal brain responses.

Scaling the velocity with which a path is traversed leaves the NVV invariant, by definition. This is also borne out experimentally, in the fact that there was no correlation found between the duration of each effort and its NVV score ($r=0.11$ with $p<0.001$ for H and $r= -0.06$ with $p=0.04$ for PD, where r is the Pearson product-moment correlation coefficient calculated for effort duration and NVV score).

The NVV value of healthy people is also not affected by their age. Using means testing for YH and H groups we found no significant differences between the two healthy subgroups' NVV values. Although aging is accompanying with a deterioration of the performances of the various components of the central nervous system, the NVV shows that the coordination between agonists and antagonists is well preserved, at least in the upper limbs.

By examining the NVV values between PD and H groups we found that the NVV means for the PD group were not normally distributed for all six components described in Section II.C, whereas the NVV means for the H group were normally distributed for every component.

We found statistically significant differences between the PD and H populations' NVV mean scores for components ALL, HH, HD, LD, HLH, suggesting that there is no overlap in the NVV measures of the two health-status-related groups.

In general, the symptoms of PD appear with different intensity between the patient's two sides (left-right). We were interested in exploring if the NVV values would reflect the laterality of the patients.

We found that there was 1 out of 20 (5%) healthy subject with a statistically significant difference between his two hands. The corresponding proportion in the PD group was 14 out of 24 (58%). Although the populations do not satisfy the standard binomial

requirement and we cannot determine the significance of the difference between these two percentages, their values are not close. The NVV detected the pathology laterality correctly for the same 58% of the patients. That means that the clinically observed (through UPDRS evaluation) most affected hand matched the NVV identified highest-performing hand for 14 out of 24 PD patients.

The fact that we found no significant difference between hands for 19 of 20 healthy subjects (95%) is to be expected. The NVV index is an expression of the balance between agonists-antagonists muscular systems and this balance is similar in both sides, considering good health. On the other side, we found that within the PD group the means are statistically significantly different for the HH and LH components. Combined with the significant difference in the means of the NVV_{HLH} component between H and PD detected earlier, this means that the NVV difference between hands, is a kind of marker of the disease and that will be useful for the classification later on.

NVV_{HLH} values are the differences in NVV between the highest and lowest performing hand for each subject. Having established that there is a statistically significant difference in NVV_{HLH} values between the H and PD groups (the latter group having on average higher NVV_{HLH} values), we went on to compare the performance dissimilarity between the best-performing, (i.e. lowest NVV) hand of the healthy persons to the worst-performing (i.e. highest NVV) hand of PD patients. We found that the best-performing hand of PD subjects is still worse, in terms of NVV score, than the worst-performing hand of H subjects. To our knowledge, this is a result not mentioned in the literature in papers using high sample rate data.

Focusing on the H and PD participants, we used each subject's mean NVV_{ALL} score to categorize them as positive (PD) or negative (H). The receiver operating characteristic (ROC) curve had an area-under-the-curve value of 0.9354. The best cutoff threshold was NVV_{ALL}=0.0165 resulting in TN=90% (18 subjects true negative), and TP=88% (21 subjects true positive). These high percentages along with the results from the means testing prove that there is little overlap in the NVV values between the two populations and the NVV and its components can justifiably be used as a feature in a PD versus H classification model.

B. Metrics' correlation to the UPDRS

There have been studies of the writing behavior of PD patients where the calculated digital metrics do not correlate well with the UPDRS scores [18]. It is true that the UPDRS is a rather coarse grain grading approach. It is common for many of our patients to have similar H&Y scores but different score vectors.

For our correlation tests we used each patient's H&Y, UPDRS component III summation, and UPDRS upper limb tremor components summation scores, and the score vectors of the PD population. As expected none of the five metrics correlated well with the UPDRS scores, regardless of the data set (i.e. HH, LH, HD, etc.) We calculated low Pearson product-moment correlation coefficient values combined with high p values. It appears that the small and sparse variations of the UPDRS scores cannot be reflected by the more sensitive pen-tablet trajectory-generated metrics.

However, we ran the correlation test once more, this time including the H population's score vectors full data set and zero padded their UPDRS related scores. As shown in Table III, the calculated correlation coefficients were indicative of similar trends between the score vectors and the UPDRS scores. All p values were less than 0.01.

Worth noting in Table III is the lack of correlation between MV and the UPDRS or H&Y scores. A deeper look into the data though, reveals that there is actually a *positive* correlation ($r \cong 0.61$, $p \cong 0$) between the H&Y scores and the MV metric restricted to the lowest-scoring direction (MV_{LD}), and a symmetric negative correlation between the H&Y and the MV_{HD} values of the highest-scoring direction. This suggests that, as one's health status deteriorates, MV_{LD} and MV_{HD} tend to converge, i.e. the slowest movement direction (extension versus flexion, potentially different for each subject) becomes relatively faster while the fastest movement direction becomes relatively slower. This observation could reflect how as PD progresses over the years, its laterality starts to fade and the symptoms of the least affected side aggravate. As we will see, MV_{HD} and MV_{LD} will prove to be useful as classifier features.

TABLE III
CORRELATION COEFFICIENT VALUES

	NVV	MV	SDV	ETPx	ETPy	MV_{HD}	MV_{LD}
H&Y	0.53	0.08	-0.47	-0.46	0.5	-0.61	0.61
UPDRS (III)	0.49	0.01	-0.32	-0.26	0.23	-0.5	0.48
UPDRS Tremor	0.49	-0.04	-0.32	-0.25	0.23	-0.5	0.47

C. Feature Selection for a Machine Learning Model

The final step in our analysis was to build a classification model that would be trained to read unlabeled data, i.e., data missing class -PD or H- information, and decide successfully on their classification. To specify the significance of the features we ran multiple tests using various algorithms and search methods to define their optimal subset, which would help us avoid over-fitting our model to redundant information. As mentioned in Section II.C, our full feature set contains the 5 metrics defined in (1)-(5), averaged in six ways, namely All, High, Low, and High-Low Hands (ALL, HH, LH, and HLH respectively), and High, Low Direction (HD, LD), resulting in a total of 30 features. Although the number of features (30) compared to the sample size (44) could save our classifier from the curse of dimensionality, the feature selection procedure was essential because many of these features could prove to be highly correlated due to the nature of calculations and common data pools. In Table IV we present a summary of the results from the selection process, including selection methods used and features selected by each method. All methods were run over a ten-fold cross-validation. The features in Table IV are listed in order of importance based on the votes they received across the ten folds.

TABLE IV
FEATURE SELECTION TESTS - TOP FEATURES ACCORDING TO VARIOUS METHODS

Test Number	Method basis	Selects	Search Algorithm	Selected Subset / Features
1	Correlation	Subset	Greedy SW	**NVV**$_{HH}$, **MV**$_{LD}$, **MV**$_{HD}$, **ETPy**$_{LD}$, **ETPx**$_{HLH}$, **NVV**$_{HD}$, **ETPy**$_{HH}$, **ETPy**$_{HD}$
2	Consistency	Subset	Greedy SW	**NVV**$_{HD}$, **ETPy**$_{LD}$, **MV**$_{HD}$, **ETPx**$_{HLH}$, **NVV**$_{HH}$, **ETPx**$_{HH}$, **NVV**$_{HLH}$, **NVV**$_{ALL}$
3	Consistency	Subset	Exhaustive	**MV**$_{HD}$, **ETPy**$_{LD}$, **NVV**$_{HH}$, **ETPy**$_{HH}$, **ETPx**$_{HLH}$, **ETPx**$_{HH}$, **NVV**$_{HD}$, **NVV**$_{HLH}$
4	Classifier (J48)	Subset	Greedy SW	**NVV**$_{HD}$, **NVV**$_{HH}$, **SDV**$_{HH}$, **ETPx**$_{HH}$, **MV**$_{LH}$, **MV**$_{HH}$, **MV**$_{HD}$, **ETPx**$_{HD}$, **ETPy**$_{HD}$
5	Wrapper (Naïve Bayes)	Subset	Greedy SW	**NVV**$_{HH}$, **ETPx**$_{LD}$, **ETPy**$_{HLH}$, **NVV**$_{ALL}$, **MV**$_{LD}$, **MV**$_{LH}$, **NVV**$_{LD}$, **ETPy**$_{LD}$, **MV**$_{HD}$
6	Correlation	Attributes	Ranker	**MV**$_{LD}$, **MV**$_{HD}$, **NVV**$_{HH}$, **NVV**$_{ALL}$, **SDV**$_{LH}$, **NVV**$_{HD}$, **ETPx**$_{LH}$, **NVV**$_{LH}$, **SDV**$_{LD}$
7	Variation Clustering	Attributes	Ranker	**SDV**$_{HH}$, **SDV**$_{HD}$, **ETPx**$_{HLH}$, **MV**$_{LD}$, **MV**$_{HD}$, **SDV**$_{ALL}$, **SDV**$_{LD}$, **ETPy**$_{HD}$
8	Gain Ratio	Attributes	Ranker	**MV**$_{LD}$, **NVV**$_{HH}$, **NVV**$_{HD}$, **NVV**$_{ALL}$, **MV**$_{HD}$, **NVV**$_{LH}$, **NVV**$_{LD}$, **ETPx**$_{HLH}$
9	Information Gain	Attributes	Ranker	**NVV**$_{HH}$, **NVV**$_{HD}$, **MV**$_{LD}$, **NVV**$_{ALL}$, **MV**$_{HD}$, **NVV**$_{LH}$, **NVV**$_{LD}$, **SDV**$_{LH}$, **SDV**$_{LD}$
10	One-R	Attributes	Ranker	**MV**$_{LD}$, **MV**$_{HD}$, **NVV**$_{HH}$, **NVV**$_{HD}$, **NVV**$_{LH}$, **NVV**$_{ALL}$, **SDV**$_{LH}$, **NVV**$_{LD}$, **SDV**$_{LD}$
11	Relief	Attributes	Ranker	**MV**$_{LD}$, **MV**$_{HD}$, **NVV**$_{HD}$, **NVV**$_{HH}$, **NVV**$_{ALL}$, **SDV**$_{LH}$, **SDV**$_{LD}$, **SDV**$_{ALL}$
12	SVM	Attributes	Ranker	**MV**$_{LD}$, **NVV**$_{HD}$, **NVV**$_{HH}$, **MV**$_{HD}$, **NVV**$_{ALL}$, **ETPy**$_{HH}$, **SDV**$_{LH}$, **ETPx**$_{HLH}$
13	Symmetrical Uncertainty	Attributes	Ranker	**NVV**$_{HH}$, **MV**$_{LD}$, **NVV**$_{HD}$, **NVV**$_{ALL}$, **MV**$_{HD}$, **NVV**$_{LH}$, **NVV**$_{LD}$, **SDV**$_{LH}$, **ETPx**$_{HLH}$

We used feature selection methods that were based on different criteria to define the value of single features and subsets of features. As is argued in [28], a single variable may not be important on its own, but contribute to the performance of the classifier when used in a subset. One good such example is the ETP, which is ranked low by the attribute-ranking-based methods but is among the first selections when it comes to subset-based methods. The SDV on the other hand is considered more important than ETP by most attribute-ranking-based methods but seems trivial when exploring subsets during feature selection. Regardless of the method used, the NVV index is almost always deemed important during the feature selection tests, suggesting that it is useful as a discriminating factor (the variation clustering approach was an exception).

D. The Classification Model

Following feature selection, we proceeded to test different classifiers to establish the best performing one for our model. As shown in Table IV, the subset space feature selection approaches have similar results and most of the attribute space selection approaches as well, but there are noticeable differences between the two methodologies. Therefore, we used the outcome of every feature selection algorithm we ran, with forward selection up to the point where adding more features would not improve

the classifier's performance. The best subset and performance of each classifier is presented in Table V. All of them were tested using 10-fold cross-validation and using forward selection of features from the subsets defined in Table IV. Although most of the feature selection methods ranked the features included in Table V high, the ones that explore the subset space rather than ranking individual attributes were the most consistent in optimizing the model's subset. The best performing classifier was Naïve Bayes, achieving 91% average accuracy, 88% sensitivity, 95% specificity, and 0.952 AUC value, using a selection of features from the feature subset defined by a wrapper of the same learning scheme. Although our data and features do not justify the assumptions made by the Naïve Bayes algorithm, its performance is theoretically proven [29] and justifies its use. The boosted tree, logistic regression and SVM methodologies had slightly worse results for feature subsets comprising mainly the NVV, ETP and MV metrics. We also experimented with other classifiers, such as random forests, with the results being similar or slightly worst. Table V is not an exhaustive record of our features-classifiers combinations, only the best-performing ones.

TABLE V
CLASSIFICATION TESTS

Classifier	Average Accuracy	AUC	TP	TN	Feature Subset	Feature Selection Test Number from Table IV
Naïve Bayes	90.90%	0.95	0.88	0.95	$\mathbf{NVV}_{HH}$, $\mathbf{ETPx}_{LD}$, $\mathbf{ETPy}_{LD}$	5
AdaBoost (J48)	88.63%	0.91	0.88	0.9	$\mathbf{NVV}_{HD}$, $\mathbf{ETPy}_{LD}$, $\mathbf{MV}_{HD}$, $\mathbf{ETPx}_{HLH}$	2
Log. Regression	86.36%	0.91	0.84	0.9	$\mathbf{NVV}_{HD}$, $\mathbf{ETPy}_{LD}$, $\mathbf{MV}_{HD}$, $\mathbf{ETPx}_{HLH}$	2
SVM	86.36%	0.85	1.00	0.7	$\mathbf{NVV}_{HH}$, $\mathbf{MV}_{LD}$, $\mathbf{MV}_{HD}$, $\mathbf{ETPy}_{LD}$	1

What is interesting is that the NVV, as a classification feature, contributes massively in achieving high accuracy percentages. In fact, when we repeated the training for the same classifiers after excluding all instances of the NVV metric, they all suffered a 5-10% decline on average accuracy.

One of our goals in this study was to introduce a new handwriting marker that could be extracted from a simple and unobtrusive task, and to confirm its merit when combined with the other well-established metrics included in our analysis. Because these metrics could (and do) lead to highly correlated features it is not appropriate for all of them to be used in the same classification model. This is why, during our first treatment of the data with machine learning techniques, we opted against the complexity of re-introducing feature selection inside the classification training, and assessed the performance of the individual feature subsets, as they are given in Table IV, applying various classifiers after the feature selection procedure, as shown in Table V and described previously in this section.

To eliminate the potential bias from the application of cross-validation both for optimizing the model through feature selection and for evaluating its performance [30], [31], we processed the data again, this time with cross-validation being the outermost "loop". Feature selection was applied to every data fold separately and then, the performance of each classifier was assessed using each fold's optimal feature subset. We applied a stratified 20-fold cross-validation 10 times, where each time, each one of the 20 random test sets consisted of one PD and one H sample. For each fold (200 folds in total) we applied WEKA's *CfsSubsetEval*, a filtering subset feature selection algorithm, which tends to prefer subsets of features that are highly correlated within the class, while having low intercorrelation [32]. Because each fold was optimized separately, the features selected were not always the same. Table VI shows the most "popular" features, where we observe that the NVV_{HH} metric introduced in this work was selected as an important feature in *all* 200 folds leaving little doubt as to its value as a marker for pathological hand motion patterns.

TABLE VI
FEATURES SELECTED DURING 10-TIME 20-FOLD STRATIFIED CROSS-VALIDATION

Feature	$\mathbf{NVV}_{HH}$	$\mathbf{MV}_{LD}$	$\mathbf{ETPy}_{LD}$	$\mathbf{MV}_{HD}$	$\mathbf{ETPx}_{HLH}$	$\mathbf{NVV}_{HD}$	$\mathbf{NVV}_{ALL}$	$\mathbf{NVV}_{HLH}$
Selected in number of folds	200	199	169	147	103	58	40	38

The features were selected for each one of the 200 folds separately using WEKA's *CfsSubsetEval*.

Inside the cross-correlation loop and after the feature selection, we applied 6 different classifiers, which were trained and tested separately using the training and testing sets of each fold. The average accuracy results are shown in Table VII. As expected, there is a drop in the accuracy percentages compared to the single-feature-set classifier described above; however, the (average)

performance of the classifiers was still among the best regarding handwriting markers, which we find remarkable considering the simplicity and parsimonious nature of the task involved.

TABLE VII
CLASSIFICATION TESTS FOR 200 FOLDS

Classifier	Average Accuracy	AUC	TP	TN
Naïve Bayes	88.63%	0.931	0.90	0.87
AdaBoost (J48)	81.81%	0.879	0.85	0.80
Log. Regression	84.09%	0.931	0.85	0.83
J48	86.36%	0.859	0.85	0.87
SVM	84.09%	0.842	0.85	0.83
Random Forest	79.54%	0.901	0.80	0.80

All numeric values are averaged over 200 folds, through a 10-time, random, stratified, 20-fold cross-validation.

IV. CONCLUSIONS

In this experiment we defined a vector of metrics based on handwriting markers that differentiates the healthy from PD patients while they are drawing simple lines. Other authors used interesting and original approaches to assess PD, other than writing words, by having participants draw a modified Archimedes spiral [18], or circles [22]. Our approach requires subjects to draw straight lines, a simple movement involving few muscular systems and low coordination control effort. Apart from its short duration and, as a consequence, minimum annoyance for the participants, our approach allows for both hands to be used and assessed, constituting the dexterity of the dominant hand irrelevant to the assessment.

The choice of keeping the drawing task as simple as possible was made with the expectation that one should be able to detect differences in the H versus PD groups even in very simple tasks, because the impairment of the coordination of antagonist muscular systems at a low level of control is expressed independently of the complexity of the task. Our findings justify that choice. A limitation of our procedure is that it can be used to check for impairment in motor coordination in the upper limbs only. On the other hand, the strength of the proposed method is that it is an objective measure of the velocity microvariations at a few milliseconds level.

For each digitized line of each subject, we calculated four metrics which have previously been used for similar purposes, and a new, information-rich metric, termed Normalized Velocity Variability (NVV), which quantifies the variability of the pen's horizontal speed as the line is drawn, independently of the average drawing speed. With each subject being represented by a score vector we designed a machine learning model which achieved maximum average classification accuracy 91% for unlabeled PD and H data. The achieved accuracy of the model at the higher end of all other similar studies known to the authors, having to do with predictive classification of PD volunteers through handwriting markers, while using more complex and cumbersome tasks and setups. In a second, more accurate machine learning experiment, where a 10-time, 20-fold cross-validation was the outermost "loop", surrounding both the feature selection and the classification evaluation loops, the NVV was shown to be the most important feature by far, having been selected in *all* 200 random folds.

The inclusion of the NVV metric alone, as well as the machine learning model proposed in this work can have immediate implications in medical practice. Touch sensitive tablets or pen-tablet digitizers are ubiquitous and can be used to quantify characteristics of PD manifestation, create profiles for patients and help their physicians track their progress in a consistent, even remote way. Our method can be adapted for remote patient monitoring or facilitating the long-term management of PD. Apart from individual benefits, the easy and non-invasive quantification of a disease's symptoms, particularly one's whose assessment is generally subjective, can only be of value to the scientific community and assist in further research and understanding of the disease and its symptoms.

Amongst our future goals is to use our tool to record simple lines' trajectories drawn by individuals with non PD-related impaired movement patterns. That could provide insight as to whether our approach could be used to specifically diagnose Parkinson's disease by identifying its characteristic symptoms.

V. ETHICAL STANDARDS & CONFLICT OF INTEREST STATEMENT

This study was performed by approval of the appropriate ethics committee and have therefore has performed in accordance

with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All subjects gave their informed consent prior to their inclusion in the study.

The authors also declare that they have no conflict of interest.

REFERENCES

[1] S. K. Van Den Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter, D. A. Bloch and L. M. Nelson. Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. American Journal of Epidemiology, 157, pp. 1015-1022, 2003.

[2] A. J. Hughes, S. E. Daniel, L. Kilford and A. J. Lees. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. Journal of Neurology, Neurosurgery & Psychiatry, 55, pp. 181-184, 1992.

[3] A. J. Hughes, S. E. Daniel, Y. Ben-Shlomo and A. J. Lees. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. Brain, 125, pp. 861-870, 2002.

[4] A. J. Hughes, Y. Ben-Shlomo, S. E. Daniel and A. J. Lees. What features improve the accuracy of clinical diagnosis in Parkinson's disease: a clinicopathological study 1992. Neurology, 57, pp. 34-38, 2001.

[5] S. L. Pullman and R. Saunders-Pullman. Assessing Disability in Movement Disorders: Quantitative Techniques and Rating Scales. In Movement Disorders by R. L. Watts, D. Standaert and J. A. Obeso, 3rd edition, McGraw-Hill Professional, 2001.

[6] R. LeMoyne, C. Coroian and T. Mastroianni. Quantification of Parkinson's disease characteristics using wireless accelerometers. In Proceedings IEEE / ICME, pp. 1-5, 2009.

[7] S. Patel, K. Lorincz, R. Hughes, N. Huggins, J. Growdon, D. Standaert, M. Akay, J. Dy, M. Welsh and P. Bonato. Monitoring Motor Fluctuations in Patients With Parkinson's Disease Using Wearable Sensors. Transactions on Information Technology in Biomedicine, 13, pp. 864-873, 2009.

[8] M. Yang, H. Zheng, H. Wang, S. McClean, J. Hall and N. Harris. Assessing accelerometer based gait features to support gait analysis for people with complex regional pain syndrome. In Proceedings PETRA, Samos, Greece, 2010.

[9] N. Kostikis, D. Hristu-Varsakelis, M. Arnaoutoglou, C. Kotsavasiloglou and S. Baloyiannis. Towards remote evaluation of movement disorders via smartphones. In Proceedings IEEE / EMBC, Boston, MA, pp. 5240-5243, 2011.

[10] N. M. Aly, J. R. Playfer, S. L. Smith and D. M. Halliday. A novel computer-based technique for the assessment of tremor in Parkinson's disease. Age and Ageing, 36, 4, pp. 395-399, 2007.

[11] M. P. Broderick, A. W. Van Gemmert, H. A. Shill and G. E. Stelmach. Hypometria and bradykinesia during drawing movements in individuals with Parkinson's disease. Experimental Brain Research, 197, 3, pp. 223-233, 2009.

[12] D. H. Romero, A. W. Van Gemmert, C. H. Adler, H. Bekkering and G. E. Stelmach. Altered aiming movements in Parkinson's disease patients and elderly adults as a function of delays in movement onset. Experimental Brain Research, 151, pp. 249-261, 2003.

[13] F. Paquet, M. A. Bedard, M. Levesque, P. L. Tremblay, M. Lemay, P. J. Blanchet, P. Scherzer, S. Chouinard and J. Filion. Sensorimotor adaptation in Parkinson's disease: evidence for a dopamine dependent remapping disturbance. Experimental Brain Research, 185, pp. 227-236, 2008.

[14] S. Rosenblum, M. Samuel, S. Zlotnik, I. Erikh and I. Schlesinger. Handwriting as an objective tool for Parkinson's disease diagnosis. Journal of neurology, 10, 1007, 2013.

[15] A. Letanneux, J. Danna, J. L. Velay, F. Viallet and S. Pinto. From micrographia to Parkinson's disease dysgraphia. Movement Disorders, 10, 1002, 2014.

[16] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. Artificial Intelligence in Medicine, 67, pp. 39-46, 2016.

[17] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, A new modality for quantitative evaluation of Parkinson's disease: In-air movement. In Proceedings IEEE 13th International Conference on Bioinformatics and Bioengineering, pp. 1-4, 2013.

[18] R. Saunders-Pullman, C. Derby, K. Stanley, A. Floyd, S. Bressman, R. B. Lipton, A. Deligtisch, L. Severt, Q. Yu, M. Kurtis, S. L. Pullman. Validity of spiral analysis in early Parkinson's disease. Movement disorders, 23, 4, pp. 531-537, 2008.

[19] M. B. Popovic, E. Dzoljic and V. Kostic. A method to assess hand motor blocks in Parkinson's disease with digitizing tablet. The Tohoku journal of experimental medicine, 216, 4, pp. 317-324, 2008.

[20] J. L. Contreras-Vidal and G. E. Stelmach. A neural model of basal ganglia-thalamocortical relations in normal and parkinsonian movement. Biological Cybernetics, 73, 5, pp. 467-476, 1995.

[21] A. W. Van Gemmert, C. H. Adler and G. E. Stelmach. Parkinson's disease patients undershoot target size in handwriting and similar tasks. Journal of Neurology, Neurosurgery & Psychiatry, 74, pp. 1502-1508, 2003.

[22] T. E. Eichhorn, T. Gasser, N. Mai, C. Marquardt, G. Arnold, J. Schwarz and W. H. Oertel. Computational analysis of open loop handwriting movements in Parkinson's disease: a rapid method to detect dopamimetic effects. Movement Disorders, 11, 3, pp. 289-297, 1996.

[23] O. Tucha, L. Mecklinger, J. Thome, A. Reiter, G. L. Alders, H. Sartor, M. Naumann and K. W. Lange. Kinematic analysis of dopaminergic effects on skilled handwriting movements in Parkinson's disease. Journal of Neural Transmission, 10, 1007, 2005.

[24] J. Carr. Tremor in Parkinson's disease. Parkinsonism & Related Disorders, 8, 4, pp. 223-234, 2002.

[25] X. Liu, C. B. Carroll, S. Y. Wang, J. Zajicek and P. G. Bain. Quantifying drug-induced dyskinesias in the arms using digitised spiral-drawing tasks. Journal of Neuroscience Methods, 144, 1, pp. 47-52, 2005.

[26] P. Drotar, J. Mekyska, Ir. Rektorova, L. Masarova, Zd. Smekal, and M. Faundez-Zanuy. Decision Support Framework for Parkinson's Disease based on novel handwriting markers. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 23, 3, pp. 508-516, 2014.

[27] A. Korchounov, M. F. Meyer and M. Krasnianski. Postsynaptic nigrostriatal dopamine receptors and their role in movement regulation. Journal of Neural Transmission, 117, pp. 1359-1369, 2010.

[28] I. Guyon, A. Elisseeff. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, pp. 1157-1182, 2003.

[29] H. Zhang. The Optimality of Naïve Bayes. In Proceedings FLAIRS, Miami, FL, 2004.

[30] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. Bioinformatics, 23, 19, pp. 2507-2517, 2007.

[31] G. C. Cawley, N. L. C. Talbot. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research, 11, pp. 2079-2107, 2010.

[32] M. A. Hall. Correlation-based feature subset selection for Machine Learning. Hamilton, New Zealand, 1998.