information

MDPI

Article

Dynamic Handwriting Analysis for Supporting Earlier Parkinson's Disease Diagnosis

Donato Impedovo, Giuseppe Pirlo and Gennaro Vessio *

Computer Science Department, University of Bari, 70125 Bari, Italy; donato.impedovo@uniba.it (D.I.); giuseppe.pirlo@uniba.it (G.P.)

* Correspondence: gennaro.vessio@uniba.it

Received: 15 September; Accepted: 28 September; Published: 3 October 2018

Abstract: Machine learning techniques are tailored to build intelligent systems to support clinicians at the point of care. In particular, they can complement standard clinical evaluations for the assessment of early signs and manifestations of Parkinson's disease (PD). Patients suffering from PD typically exhibit impairments of previously learned motor skills, such as handwriting. Therefore, handwriting can be considered a powerful marker to develop automatized diagnostic tools. In this paper, we investigated if and to which extent dynamic features of the handwriting process can support PD diagnosis at earlier stages. To this end, a subset of the publicly available PaHaW dataset has been used, including those patients showing only early to mild degree of disease severity. We developed a classification framework based on different classifiers and an ensemble scheme. Some encouraging results have been obtained; in particular, good specificity performances have been observed. This indicates that a handwriting-based decision support tool could be used to administer screening tests useful for ruling in disease.

Keywords: Parkinson's disease; e-health; computer aided diagnosis; artificial intelligence; dynamic handwriting analysis

1. Introduction

Neurodegenerative disorders affect the structure and functions of brain regions resulting in a progressive cognitive, functional and behavioral decline. Among them, Parkinson's disease (PD) is one of the most common and most disabling. It is mainly characterized by motor symptoms, including akinesia, bradykinesia, rigidity and tremor, as well as non-motor deficits, such as depression, sleep disorders and dementia [1]. Today, there is no cure and a precise diagnosis is possible only *post-mortem*. Nevertheless, an early diagnosis of PD would be crucial in the perspective of the proper medical treatment to be administered and for evaluating the effectiveness of new drug treatments at prodromal stages.

A growing interest has thus arisen in the scientific community in e-health, particularly in computer aided diagnosis (CAD) systems. Such systems, in fact, have the potential to assist clinicians at the point of care, providing novel diagnostic tools, while reducing the expenditure of public health care. Of course, CAD systems are not expected to replace standard techniques, but to provide complementary approaches to standard evaluations that are non-invasive and very low-cost.

As handwriting difficulties in PD patients have been documented since a long time, a promising *biomarker* concerns the changes in handwriting due to the concomitant impairment [2,3]. Handwriting, in fact, is a complex activity which involves several aspects including fine motor control, eye-hand coordination, visuo-spatial abilities, and so on [4]. Evidence exists about the effectiveness of machine learning approaches aimed at discriminating between unhealthy and healthy subjects based on simple and easy-to-perform handwriting tasks, e.g., [5–7].

The most employed approach to studying the diagnostic potential of handwriting tasks consists in exploiting *dynamic* features of the handwriting process. This approach relies on the analysis of time series data characterizing handwriting, which can be acquired through the use of digitizing tablets provided with electronic pens (see, for example, [8]). Such a device enables the collection of the geometric position of the pen at certain time stamps, as well as the pressure exerted over the writing surface, pen inclination, and if the movement of the pen is performed "in the air".

Some recent studies adopted dynamic handwriting analysis for the purpose of PD classification showing encouraging results, e.g., [9,10]. However, all of them focused on the binary discrimination healthy/unhealthy independently of the degree of disease severity shown by the PD cohort. In other words, the Parkinsonian sample is considered as a single cluster in which all subjects share the same degree of severity. The aim of this paper is to investigate if and to which extent dynamic features of the handwriting process can help discriminate people suffering from PD at earlier stages.

To this end, a freely available dataset, i.e., PaHaW [11], has been used, as it includes several patients at different degrees of disease severity. Dynamic features have then been extracted from the handwriting tasks performed by these subjects. Finally, a classification framework has been developed, based on both the analysis of the overall feature vector resulting from all tasks and the analysis of each task taken individually.

The rest of this paper is organized as follows. Section 2 describes the data used for the present study. Section 3 focuses on the experimental set up. Section 4 analyzes the results obtained. Section 5 concludes the work.

2. Dataset

The "Parkinson's disease handwriting database" (PaHaW) [11] collects data of handwriting tasks performed by 75 subjects, 37 PD patients and 38 age and gender-matched healthy control (HC) subjects, enrolled at the First Department of Neurology, Masaryk University and at the St. Anne's University Hospital, in Brno, Czech Republic. All subjects were right-handed, had completed at least ten years of education, and reported Czech as their native language. PD patients were examined only in their ON-state while on dopaminergic medication and, prior to acquisition, they were evaluated by a clinical neurologist. In addition, the HC group was examined by a clinician to make sure that there was no movement disorder or injury that could have significantly affected handwriting. Note that, since three subjects (1 PD, 2 HC) did not complete all tasks, we excluded them.

Information about participants are summarized in Table 1. The table shows age, disease duration, levedopa equivalent dose (LED) and gender. Note that this information is grouped depending on the degree of severity given by the UPDRS (part V) score [12]. This score corresponds to the Modified Hoehn and Yahr Scale, which is a commonly used rating scale for describing how PD symptoms evolves during time [13]. The scale provides stages from 1 to 5, including 1.5 and 2.5 as intermediate stages. In PaHaW, all stages are represented, except for stage 1.5. In accordance with this score, the PD group can be divided into two subgroups: in the first one, the disease severity ranges from 1 to 2.5, i.e., from early to mild; in the second one, it ranges from 3 to 5, i.e., from mild to moderate or severe. The aim of this paper is to study if handwriting can support an earlier diagnosis, so we focused only on the first subgroup, which includes 29 patients.

Participants performed eight handwriting tasks in accordance with a pre-filled template:

1. Drawing an Archimedes spiral;
2. Writing in cursive the letter *l*;
3. The bigram *les*;
4. The trigram *les*;
5. Writing in cursive the word *lektorka* ("female teacher" in Czech);
6. *porovnat* ("to compare");
7. *nepopadnout* ("to not catch");
8. Writing in cursive the sentence *Tramvaj dnes už nepojede* ("The tram won't go today").

Note that the spiral task, as well as the single word tasks, can be performed fluently, without lifting the pen from the surface.

The signals were recorded using the Intuos 4M digitizing tablet (Wacom technology), overlaid with an empty paper. The sampling rate was 200 Hz. The time series raw data captured by the device are the x- and y-coordinates of the pen position and their timestamps. Moreover, the device is able to record the pressure exerted over the writing surface and measures of pen inclination, i.e., azimuth and altitude. The last signal is the so-called button status, which is a binary variable evaluating 1 when the pen is on the surface, 0 when the pen is no more than 1 cm from the surface (thus allowing the acquisition of "in-air movements"). All measures have the same length, varying from execution to execution.

Table 1. PaHaW dataset. A line divides the Parkinson's disease (PD) group into two subgroups depending on the disease severity: early to mild vs. mild to severe. In the present work, only the healthy control (HC) group and the early to mild subgroup have been taken into account.

	Age		Disease Duration		LED		M/F	Stage
	Mean	Std	Mean	Std	Mean	Std		
HC	62.5	11.5	–	–	–	–	19/17	–
PD	62.2	9.7	3.8	1.1	986.2	611.2	3/2	1
	–	–	–	–	–	–	–	1.5
	70.6	12.0	7.0	3.4	1385.1	634.9	6/12	2
	68.0	7.9	12.0	4.8	1674.4	478.1	4/2	2.5
	69.0	7.2	12.5	4.1	1349.9	610.8	3/1	3
	78.5	8.5	9.5	2.5	1383.3	66.6	2/0	4
	75	–	18	–	1370.0	–	0/1	5

3. Experimental Set Up

The experimental methodology includes several steps which are described in the following subsections. Note that this methodology is similar to the one proposed in [9], in which state-of-the-art results have been obtained on the same dataset. The main differences are listed below:

- In [9], all PD subjects were considered independently of their degree of disease severity. In the present work, we focused only on those patients exhibiting earlier manifestations of the disease;
- In [9], the features coming from each task were combined into a single high dimensional feature vector that was fed into a single machine learning algorithm. In this work, in addition to this approach, we also investigated the discriminating power of every single task and combinations of them, by using several machine learning algorithms and an ensemble approach.

3.1. Feature Extraction

From the time series raw data sampled by the acquisition device, several features capturing the dynamics of handwriting have been extracted. The comprehensive list of features is reported in Table 2. It is worth noting that the feature extraction stage resulted in either a single value feature or a vector feature. For all of the resulting vector features, the following basic statistical measures have been calculated: mean; median; standard deviation; 1st percentile; 99th percentile; 99th–1st percentile (outlier robust range).

The horizontal and vertical components of the pen position have been segmented into on-surface and in-air strokes, in accordance with the button-status. A *stroke* is a single connected and continuing trait of the handwritten pattern: on-surface strokes correspond to the trace left on the pad surface; in-air strokes are imaginary traces that express the pauses and hesitations between characters and words. Based on this segmentation, several kinematic features have been computed: number of strokes; tangential, horizontal and vertical displacement, velocity, acceleration and jerk; number of changes of velocity/acceleration (NCV/NCA), NCA and NCV relative to writing duration. Displacement corresponds to the straight-line distance between two consecutive sampled points: given the high sampling frequency of

the device (200 Hz), it provides a good approximation of the pen trajectory. Displacement can be simply calculated as follows:

$$d_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}, i = 1, \ldots, N-1,$$

where N is the number of points sampled within a stroke and $d_N = d_{N-1}$. From displacement, the other kinematic measures are easily calculated.

In addition, spatio-temporal features have been extracted: stroke size and duration; speed and stroke speed; stroke height and width; on-surface and in-air time; normalized on-surface and in-air time; in-air/on-surface ratio. The calculation of these features is quite straightforward.

Table 2. List of features. Unless otherwise specified, they are intended both on-surface and in-air.

Feature	*s*/*v*	Description
Stroke number	*s*	Number of strokes
Displacement	*v*	Tangential trajectory during handwriting
Velocity	*v*	Rate of change of position whit respect to time
Acceleration	*v*	Rate of change of velocity with respect to time
Jerk	*v*	Rate of change of acceleration with respect to time
Horizontal/vertical displacement	*v*	Displacement in the horizontal/vertical direction
Horizontal/vertical velocity	*v*	Velocity in the horizontal/vertical direction
Horizontal/vertical acceleration	*v*	Acceleration in the horizontal/vertical direction
Horizontal/vertical jerk	*v*	Jerk in the horizontal/vertical direction
NCV	*s*	Mean number of local extrema of velocity
NCA	*s*	Mean number of local extrema of acceleration
Relative NCV	*s*	NCV relative to writing duration
Relative NCA	*s*	NCA relative to writing duration
Stroke size	*v*	Path length of each stroke
Stroke duration	*v*	Movement time per stroke
Speed	*s*	Trajectory during handwriting divided by writing duration
Stroke speed	*v*	Trajectory during stroke divided by stroke duration
Stroke height	*v*	Height of each stroke
Stroke width	*v*	Width of each stroke
On-surface time	*s*	Overall time spent on-surface
In-air time	*s*	Overall time spent in-air
Total time	*s*	On-surface time plus in-air time
Normalized on-surface time	*s*	On-surface time normalized by total time
Normalized in-air time	*s*	In-air time normalized by total time
In-air/on-surface ratio	*s*	Ratio of time spent in-air/on-surface
Mean pressure	*v*	Average pressure over all on-surface strokes
NCP	*s*	Mean number of local extrema of pressure
Relative NCP	*s*	NCP relative to writing duration
Horizontal/vertical Shannon entropy	*v*	Shannon entropy of the horizontal/vertical component of the pen position
Horizontal/vertical Rényi entropy	*v*	Second and third order Rényi entropy of the horizontal/vertical component of the pen position
Horizontal/vertical signal-to-noise ratio	*v*	Signal-to-noise ratio of the horizontal/vertical component of the pen position
Horizontal/vertical intrinsic Shannon entropy	*v*	Shannon entropy of the first/second IMF obtained by the EMD of the horizontal/vertical component of the pen position
Horizontal/vertical intrinsic Rényi entropy	*v*	Second and third order Rényi entropy of the first/second IMF obtained by the EMD of the horizontal/vertical component of the pen position
Horizontal/vertical signal-to-noise ratio	*v*	Signal-to-noise ratio of the first/second IMF obtained by the EMD of the horizontal/vertical component of the pen position

Abbreviations: *s* = scalar value; *v* = vector of elements.

To make use of the pressure signal, we also computed: mean pressure; number of changes of pressure (NCP); relative NCP.

The above-mentioned features are suited to our classification problem. Several studies, in fact, reported impairments of Parkinsonian handwriting in terms of writing time, writing size, pressure applied, velocity fluctuations, and so on (e.g., [14–16]).

In order to uncover hidden complexities of handwriting, the following features have also been computed for both the on-surface and in-air horizontal and vertical components of handwriting: Shannon and Rényi entropy; signal-to-noise ratio (SNR) and empirical mode decomposition (EMD). These measures are likely to provide information about the randomness and irregularity of fine movements due to the concomitant impairment. The classic formula for Shannon entropy is:

$$H_S(X) = -\sum_{x \in X} p(x) \log_2 p(x),$$

where $p(x)$ is the probability density function estimated with a Gaussian kernel. Instead, Rényi entropy is given by:

$$H_{R,r}(X) = \frac{1}{1-r} \log_2 \left(\sum_{i=1}^{n} p_i^r \right),$$

where r is the Rényi entropy order ($r \geq 0$ and $r \neq 1$) and $p_i = \Pr(X = i)$. In this paper, the second and third order Rényi entropy have been considered. SNR has been simply calculated as the ratio between the median of the signal and the standard deviation of the estimated background noise. Finally, EMD has been applied. The method iteratively decomposes the signal into so-called intrinsic mode functions (IMFs), which are functions that satisfy two requirements: (1) the number of extrema and the number of zero crossings are either equal or differ at most by one; and (2) the mean of their upper and lower envelopes equals zero. In this paper, Shannon and Rényi entropy and SNR have been applied to only the first and second IMF resulting from the decomposition.

All features have been normalized before classification to have zero mean and unit variance.

3.2. Model Fitting

Some state-of-the-art supervised machine learning algorithms have been employed: K-Nearest Neighbours, Support Vector Machines, Gaussian Naïve Bayes, Linear Discriminant Analysis, Random Forest and AdaBoost. They are briefly described in the following paragraphs. These algorithms are tailored to the small dataset here adopted and the high dimensionality of the feature space. It is worth remarking that, for each algorithm, the scikit-learn implementation has been used [17]. We did not consider other advanced techniques, such as Neural Networks and Deep Learning, because they typically require large sets of data for training.

3.2.1. K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a very simple approach to classification which outputs, for any given test example, the most commonly-occurring class among the K closest examples in the training set [18]. In the present paper, the usual Euclidean distance has been used as distance metric, while K has been set to 5.

3.2.2. Support Vector Machines

Support Vector Machines (SVMs) work by constructing a separating hyperplane between the two classes so that the minimal distance from the closest data points of either classes is the largest [19]. Test examples are predicted to belong to a class based on which side of the hyperplane they fall. To mitigate overfitting, the *margin* of the hyperplane is chosen so that most of the training examples are separated correctly, while some of them are misclassified. To learn nonlinear decision boundaries, the data points are mapped to a higher dimensional space via a kernel function: a popular choice is

the radial basis function (RBF) kernel. In the present work, we employed both the linear kernel and the RBF kernel. Note that the bias-variance trade-off of the algorithm depends on the fine tuning of the penalty parameter C and the kernel coefficient γ in the case of RBF kernel [18]. We set $C = 1$ and $\gamma = \frac{1}{n}$, where n is the number of features. These values represent a typical setting.

3.2.3. Gaussian Naïve Bayes and Linear Discriminant Analysis

These are *generative* algorithms that directly model the class conditional distribution for each class $k \in K$. Predictions on unseen examples are obtained by applying the Bayes rule and by outputting the class for which the following estimated probability is the largest:

$$\Pr(Y = k \mid X = x) = \frac{\Pr(X = x \mid Y = k) \Pr(Y = k)}{\Pr(X = x)}.$$

$\Pr(X = x)$ is simply a scale factor; $\Pr(Y = k)$ can be easily calculated as the fraction of the training examples that are in class k; $\Pr(X = x \mid Y = k)$ can be estimated in several ways, in particular:

- In Gaussian Naïve Bayes (NB), a univariate Gaussian density for each class and conditional independence among classes are assumed;
- In Linear Discriminant Analysis (LDA), a multivariate Gaussian with shared variance across the different classes is assumed.

3.2.4. Random Forest

Random Forest (RF) is an ensemble method for classification which builds a multitude of decision trees at training time and outputs the mode of the classes predicted by each individual tree at test time [20]. In particular, RF repeatedly (B times) selects a random sample with replacement from the training set and fits a decision tree to this sample. Every decision tree is built on a subset of randomly selected features. The final predictions are obtained via majority voting of the B single predictions. B is a free parameter: in the present work, we chose $B = 500$ trees, which is a common choice in the literature.

3.2.5. AdaBoost

AdaBoost (ADA) is a *meta-classifier* which relies on the *boosting* technique: a sequence of "base learners" is fitted on the training set; then, additional copies of the classifier are fitted on the same set but with the weights of the incorrectly classified examples updated [21]. The predictions from all the base learners are finally combined through a weighted majority voting scheme. More precisely, the data modifications at each boosting iteration consist in applying weights $w_1, w_2, \ldots, w_N$ to each of the training examples. Initially, those weights are all set to $w_i = \frac{1}{N}$. Then, for each iteration, the sample weights are individually modified and the learning algorithm is reapplied to the re-weighted data. At a given step, those training examples that were incorrectly predicted at the previous step have their weights increased; those examples that were correctly predicted have their weights decreased. Examples that are difficult to predict receive ever-increasing influence; in other words, each subsequent base learner focuses on the examples that were missed by the previous learners in the sequence [18]. In the present work, 500 decision trees have been used as base learners.

3.2.6. Ensemble

Similar or conceptually different classifiers can be combined through a voting scheme, so that the individual weakness of each single classifier is mitigated [22]. Combining the classes predicted by different classifiers is likely to provide better predictions, due to diversification. In this work, a majority voting scheme has been adopted: the final class label is the most-occurring class label predicted by each individual classifier in the ensemble. First, every classification model has been trained on the features coming from each handwriting task and their performances have been evaluated. Then, since different classifiers are likely to provide different results on the same data, the best models, i.e., those showing

the best result per task, have been pooled in the ensemble scheme to achieve the final classification. In addition, since some tasks may be less useful for diagnosis than others, and their presence may introduce additional bias in the data, we also investigated the ensemble obtained by combining only the best three tasks, i.e., the tasks obtaining the highest prediction accuracy among all tasks.

3.3. Model Validation

The classification performances have been validated with a 10-fold cross-validation. This validation scheme is typically preferred when the set of examples is small, as in our case. Briefly speaking, the entire dataset is split into ten subsets/folds: nine folds are used as training set; the remaining fold represents the test set. The entire procedure is repeated ten times, by using different assignments of the examples into the folds, until all examples have been tested once. Note that we employed a *stratified* cross-validation: each fold was constructed to have approximately the same number of subjects from each diagnostic group (healthy vs. Parkinsonian).

3.4. Feature Selection

Since the number of features was disproportionately higher than the cardinality of the dataset, to reduce the dimensionality of the feature space and so to mitigate overfitting, a feature selection technique has been applied before classification [23]. We used a *filter* method which evaluates the discriminating power of every single feature based on the performance of a classifier built upon each of them. More precisely, every feature has been evaluated by taking into account its prediction accuracy in the classification PD vs. HC when used as a single input feature to a linear SVM classifier. All features have been ranked in accordance with this score and only the n features providing the highest score have been selected for the final model fitting. This feature selection technique has been chosen because it provided improved results on the same dataset [9].

It is worth remarking that feature selection has been *nested* within cross-validation, so that the most important features have been chosen based only on the training set, blind to the test set. In other words, different features have been selected based on the random assignment of data samples to the training folds within each cross-validation iteration. This procedure provides more reliable results, considering that an aprioristic supervised selection of features on the entire dataset introduces a bias in the classification model which may lead to overoptimistic results [18].

4. Results and Discussion

In the following subsections, the results of two experiments are reported:

1. The first experiment investigates the predictive potential of dynamic handwriting by merging the features coming from all tasks into a single high dimensional feature vector;
2. The second experiment investigates the predictive potential of each individual task and a combination of all the best tasks via ensemble learning.

The experimental results are expressed in terms of some traditional classification performance metrics: accuracy, area under the ROC curve (AUC), sensitivity and specificity. For each measure, the mean value, averaged over all the cross-validation iterations, is reported.

4.1. Merging All Tasks

Table 3 reports the results obtained by the different classification models over the combined feature vector obtained by merging all tasks. Generally speaking, all classifiers show low sensitivity and good specificity, indicating that such an approach may be better in identifying the absence of illness in the healthy population rather than the presence of illness in the pathological group. Another possible interpretation is that the unbalanced dataset (36 HC vs. 29 PD) may have led to a predictive preference for the majority class, even if the over-representation of the healthy sample is quite small. Secondly, it can be noted that the best results have been obtained by SVM with RBF kernel

and, in particular, Random Forest, indicating that, merging all tasks, the classes are better separated by a nonlinear decision boundary. Moreover, the better performances of these classifiers were expected, as it is well known that their effectiveness with high dimensional data, as in this case. It is worth noting that NB achieved the best specificity over all classification models. Nevertheless, such a performance is counterbalanced by the very low value of sensitivity.

The confusion matrix for the best performing classification model, namely Random Forest, is provided in Table 4.

Table 3. Classification performance with features merged from all tasks. The best results are in bold.

Classifier	Accuracy	AUC	Sensitivity	Specificity
KNN	67.90%	72.22%	48.28%	83.33%
SVM-RBF	71.33%	73.89%	55.17%	83.33%
SVM-linear	68.24%	68.33%	58.62%	75.00%
NB	57.29%	68.75%	17.24%	**88.89 %**
LDA	66.81%	70.83%	58.62%	72.22%
RF	**73.38%**	**75.00%**	**62.07%**	83.33%
ADA	61.81%	69.71%	48.28%	72.22%

Table 4. Confusion matrix for the RF classifier.

	HC (*Predicted*)	**PD (*Predicted*)**
HC (*true*)	27	7
PD (*true*)	11	18

4.2. Ensemble of Tasks

Table 5 reports the results obtained by each classification model on every task. In order to achieve the best combination of tasks to be pooled in the ensemble scheme, a performance-driven task selection has been carried out based on prediction accuracy: so, only the mean accuracies per classifier are reported in the table. Below chance accuracies have been obtained in some cases, indicating that the classification task at hand is difficult. In this case, SVM-RBF and RF provided worst results: this may have been due to the observation that the dimensionality of the feature space of every task is much lower than the feature vector obtained with all tasks. Nevertheless, there is an agreement between the classifiers about the overall best tasks, that are writing the bigram *le* (task 3), writing the word *porovnat* (task 6) and writing an entire sentence (task 8). The discriminating power of the sentence task was expected and confirmed the findings already reported in [11]: writing a long sentence, in fact, requires more cognitive load, so can exacerbate the effects of PD on handwriting. In addition the good performance achieved by *le le le* was expected: a pause between a character and the following one can point out the necessity to re-plan the fine motor activity, while fluid writing typically reveals the presence of an anticipated motor planning. Parkinsonian handwriting is known to be more segmented, so difficulties in anticipating the forthcoming movement can be observed [24]. These two tasks achieved performances comparable to those obtained by merging all tasks, indicating that the ensemble of tasks is likely to improve prediction accuracy. Conversely, very low performance can be observed in the spiral task, even if this exercise is a common practice to evaluate tremor (see, for example, [25]). This confirmed the findings reported in [11], where the task was employed without having a significant impact on classification. Such a degradation may have been due to the use of features only tailored to handwriting. In [26], instead, different findings have been reported on the same task by using visual features automatically learned by convolutional neural network models.

In Table 6, the results obtained by the ensemble of all tasks and the best three tasks are shown. The first one exhibited perfect specificity, which positively affected the mean AUC value. The second one, which is made up of only the best tasks, so filtering out probably noisy data, exhibited a more balanced behavior

between sensitivity and specificity, positively affecting accuracy. There is usually a trade-off between sensitivity and specificity and high values for both are obtained simultaneously only when the classifier approximates well the optimal Bayes classifier.

The confusion matrices for the ensemble of all tasks and for the ensemble of the best three tasks are provided in Tables 7 and 8, respectively.

Table 5. Accuracy performance task by task. The best results per task are in bold; the best three tasks overall are in italics.

Task	KNN	SVM-RBF	SVM-lin.	NB	LDA	RF	ADA
(1) Spiral	48.85%	53.69%	50.67%	**54.67%**	49.23%	51.95%	53.00%
(2) *l l l*	57.52%	57.28%	57.19%	56.61%	56.09%	56.38%	**61.80%**
(3) *le le le*	59.09%	61.90%	***72.28 %***	56.71%	66.57%	62.67%	61.47%
(4) *les les les*	40.76%	47.42%	50.38%	**55.28%**	48.38%	51.95%	47.80%
(5) *lektorka*	51.76%	45.57%	49.23%	49.57%	47.57%	45.04%	**59.80%**
(6) *porovnat*	52.19%	61.80%	62.00%	44.23%	***63.71%***	56.14%	60.33%
(7) *nepopadnout*	47.57%	46.14%	54.80%	45.80%	52.19%	59.09%	**60.28%**
(8) Sentence task	58.28%	71.09%	59.23%	***71.95%***	64.23%	66.85%	59.47%

Table 6. Classification performance of the ensemble of tasks. The best results are in bold.

Ensemble Scheme	Accuracy	AUC	Sensitivity	Specificity
All tasks	69.52%	**83.06%**	31.03%	**100.00%**
Three best tasks (task 3, 6 and 8)	**74.76%**	82.78%	**68.97%**	77.78%

Table 7. Confusion matrix for the ensemble of all tasks.

	HC (*Predicted*)	PD (*Predicted*)
HC (*true*)	36	0
PD (*true*)	20	9

Table 8. Confusion matrix for the ensemble of the best three tasks.

	HC (*Predicted*)	PD (*Predicted*)
HC (*true*)	28	8
PD (*true*)	9	10

5. Conclusions

Changes in handwriting are promising as a discriminant factor for neurodegenerative disease assessment, so handwriting analysis can be employed to develop intelligent systems to assist clinicians at the point of care. In particular, as handwriting difficulties in Parkinsonian patients are well known, such a system could be useful to support the diagnosis of Parkinson's disease at earlier stages. In this paper, the predictive potential of dynamic features of the handwriting process for this purpose have been investigated. Encouraging results have been achieved on the publicly available PaHaW dataset, which provides data of different handwriting tasks performed by PD patients and healthy subjects as well. In particular, good specificity and low sensitivity have been observed both by merging features from all tasks and with combinations of tasks based on an ensemble approach. These results suggest that a screening test based on such a tool may be able to correctly exclude the disease from the healthy population, thus may be useful for ruling in disease when resulting in a positive response.

It is worth remarking that state-of-the-art results on the PaHaW dataset have been obtained in [9] and [11]. However, these results concern the discrimination healthy/Parkinsonian independently of the degree of disease severity of the pathological group. Differences between groups may have been so exacerbated by the presence of patients with a severe degree of severity, which are likely to suffer a more severe degeneration of the neuromuscular system performances. In this work, instead, we focused only on patients showing early to mild degree of disease severity: thus, the main contribution of this paper has been to investigate the effectiveness of dynamic handwriting analysis for supporting PD diagnosis at earlier stages.

The major limitation of this study is of course the small size of the population under investigation, which makes the results obtained less generalizable. Unfortunately, the availability of a large benchmark dataset is still a gap in the scientific community working on this topic. Future work should address this issue. In addition, future developments of the present research may explore novel classification strategies to further improve prediction performances. For example, combining dynamic to *static* features of handwriting, namely the ones based on images of the patterns acquired, may provide additional insights and better performance on the data. Novel insights could also be obtained by considering other kinematic properties of the human motor control as features. For instance, it would be interesting to analyze the data through the calculation of the proportional relationships between the velocity and the motion curvature, known as two-third power law [27]. This law applies to a large variety of trajectories and has been recently applied with successful results to the teleoperation of a mobile robot [28].

Author Contributions: Conceptualization, D.I., G.P. and G.V.; Formal Analysis, G.V.; Funding Acquisition, D.I. and G.P.; Investigation, D.I., G.P. and G.V.; Methodology, D.I., G.P. and G.V.; Project Administration, D.I. and G.P.; Software, G.V.; Supervision, G.P.; Validation, D.I., G.P. and G.V.; Writing—Original Draft, G.V.

Funding: This research was funded by the Italian Ministry of Education, University and Research within the PRIN2015-HAND Project under Grant H96J16000820001.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lang, A.E.; Lozano, A.M. Parkinson's disease. *N. Engl. J. Med.* **1998**, *339*, 1130–1143.
2. De Stefano, C.; Fontanella, F.; Impedovo, D.; Pirlo, G.; di Freca, A.S. Handwriting analysis to support neurodegenerative diseases diagnosis: A review. *Pattern Recognit. Lett.* **2018**, in press.
3. Impedovo, D.; Pirlo, G. Dynamic handwriting analysis for the assessment of neurodegenerative diseases: A pattern recognition perspective. *IEEE Rev. Biomed. Eng.* **2018**, in press.
4. Tseng, M.H.; Cermak, S.A. The influence of ergonomic factors and perceptual–motor abilities on handwriting performance. *Am. J. Occup. Ther.* **1993**, *47*, 919–926.
5. Rosenblum, S.; Samuel, M.; Zlotnik, S.; Erikh, I.; Schlesinger, I. Handwriting as an objective tool for Parkinson's disease diagnosis. *J. Neurol.* **2013**, *260*, 2357–2361.
6. Kotsavasiloglou, C.; Kostikis, N.; Hristu-Varsakelis, D.; Arnaoutoglou, M. Machine learning-based classification of simple drawing movements in Parkinson's disease. *Biomed. Signal Process. Control* **2017**, *31*, 174–180.
7. Pereira, C.R.; Pereira, D.R.; Rosa, G.H.; Albuquerque, V.H.; Weber, S.A.; Hook, C.; Papa, J.P. Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification. *Artif. Intell. Med.* **2018**, *87*, 67–77.
8. Sybenga, S.; Rybarczyk, Y. Using machine learning and image processing for character recognition: An application for teaching handwriting. In Proceedings of the 28th International Conference on Computer Applications in Industry and Engineering, San Diego, CA, USA, 12–14 October 2015.
9. Drotár, P.; Mekyska, J.; Rektorová, I.; Masarová, L.; Smékal, Z.; Faundez-Zanuy, M. Decision support framework for Parkinson's disease based on novel handwriting markers. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2015**, *23*, 508–516.
10. Zham, P.; Arjunan, S.; Raghav, S.; Kumar, D.K. Efficacy of guided spiral drawing in the classification of Parkinson's Disease. *IEEE J. Biomed. Health Inform.* **2017**, *5*, 1648–1652.

11. Drotár, P.; Mekyska, J.; Rektorová, I.; Masarová, L.; Smékal, Z.; Faundez-Zanuy, M. Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. *Artif. Intell. Med.* **2016**, *67*, 39–46.
12. Goetz, C.G.; Tilley, B.C.; Shaftman, S.R.; Stebbins, G.T.; Fahn, S.; Martinez-Martin, P.; Poewe, W.; Sampaio, C.; Stern, M.B.; Dodel, R.; et al. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov. Disord. Off. J. Mov. Disord. Soc.* **2008**, *23*, 2129–2170.
13. Goetz, C.G.; Poewe, W.; Rascol, O.; Sampaio, C.; Stebbins, G.T.; Counsell, C.; Giladi, N.; Holloway, R.G.; Moore, C.G.; Wenning, G.K.; et al. Movement disorder society task force report on the hoehn and yahr staging scale: Status and recommendations the movement disorder society task force on rating scales for Parkinson's disease. *Mov. Disord.* **2004**, *19*, 1020–1028.
14. Teulings, H.L.; Stelmach, G.E. Control of stroke size, peak acceleration, and stroke duration in Parkinsonian handwriting. *Hum. Mov. Sci.* **1991**, *10*, 315–334.
15. Van Gemmert, A.; Teulings, H.L.; Contreras-Vidal, J.L.; Stelmach, G. Parkinson's disease and the control of size and speed in handwriting. *Neuropsychologia* **1999**, *37*, 685–694.
16. Broderick, M.P.; Van Gemmert, A.W.; Shill, H.A.; Stelmach, G.E. Hypometria and bradykinesia during drawing movements in individuals with Parkinson's disease. *Exp. Brain Res.* **2009**, *197*, 223–233.
17. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
18. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: Berlin, Germany, 2009.
19. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin, Germany, 2013.
20. Breiman, L. Random forests. In *Machine Learning*; Springer: Berlin, Germany, **2001**, *45*, 5–32.
21. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139.
22. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **2010**, *33*, 1–39.
23. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
24. Bidet-Ildei, C.; Pollak, P.; Kandel, S.; Fraix, V.; Orliaguet, J.P. Handwriting in patients with Parkinson disease: Effect of L-dopa and stimulation of the sub-thalamic nucleus on motor anticipation. *Hum. Mov. Sci.* **2011**, *30*, 783–791.
25. Pullman, S.L. Spiral analysis: A new technique for measuring tremor with a digitizing tablet. *Mov. Disord.* **1998**, *13*, 85–89.
26. Moetesum, M.; Siddiqi, I.; Vincent, N.; Cloppet, F. Assessing visual attributes of handwriting for prediction of neurological disorders—A case study on Parkinson's disease. *Pattern Recognit. Lett.* **2018**, in press.
27. Lacquaniti, F.; Terzuolo, C.; Viviani, P. The law relating the kinematic and figural aspects of drawing movements. *Acta Psychol.* **1983**, *54*, 115–130.
28. Rybarczyk, Y.; Carvalho, D. Effect of the implementation of the two-third power law in teleoperation. In *Advances in Human Factors and System Interactions*; Springer: Berlin, Germany, 2017; pp. 283–292.