Artificial Intelligence in Medicine 87 (2018) 67–77

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification

Clayton R. Pereira[a], Danilo R. Pereira[b], Gustavo H. Rosa[c], Victor H.C. Albuquerque[d], Silke A.T. Weber[e], Christian Hook[f], João P. Papa[c,*]

[a] *UFSCAR – Federal University of São Carlos, Department of Computing, São Carlos, Brazil*
[b] *UNOESTE – University of Western São Paulo, Presidente Prudente, Brazil*
[c] *UNESP – São Paulo State University, School of Sciences, Bauru, Brazil*
[d] *UNIFOR – Graduate Program in Applied Informatics, Fortaleza, Brazil*
[e] *UNESP – São Paulo State University, Botucatu Medical School, Botucatu, Brazil*
[f] *OTH – Ostbayerische Technische Hochschule, Regensburg, Germany*

ARTICLE INFO

Article history:
Received 20 July 2017
Received in revised form 5 April 2018
Accepted 7 April 2018

Keywords:
Parkinson's disease
Convolutional neural networks
Handwritten dynamics

ABSTRACT

Background and objective: Parkinson's disease (PD) is considered a degenerative disorder that affects the motor system, which may cause tremors, micrography, and the freezing of gait. Although PD is related to the lack of dopamine, the triggering process of its development is not fully understood yet.
Methods: In this work, we introduce convolutional neural networks to learn features from images produced by handwritten dynamics, which capture different information during the individual's assessment. Additionally, we make available a dataset composed of images and signal-based data to foster the research related to computer-aided PD diagnosis.
Results: The proposed approach was compared against raw data and texture-based descriptors, showing suitable results, mainly in the context of early stage detection, with results nearly to 95%.
Conclusions: The analysis of handwritten dynamics using deep learning techniques showed to be useful for automatic Parkinson's disease identification, as well as it can outperform handcrafted features.

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disease that affects millions of people worldwide, and it has no cure. The main reported symptoms are often related to the freezing of gate, tremors, and alterations in gait and speech, to name a few. Such illness usually impacts daily activities and reduces the quality of life concerning patients and their families [1–4].

A number of drugs have been developed to cope with the disease, but their usage along the years might hasten neurodegeneration [5]. The main problem regarding PD concerns its detection in early stages, since it is unknown the real situations that trigger Parkinson's Disease. Therefore, researchers from different areas aim at pushing together their skills and helping each other to better understand such illness. In this context, machine learning techniques seem to be quite useful since they can learn intrinsic information that sometimes are not perceived by humans.

Das [6], for instance, presented a comparison among some classification techniques concerning PD diagnosis, achieving around 92.2% of classification accuracy by means of Neural Networks. Spadotto et al. [7] introduced the Optimum-Path Forest (OPF) [8,9] in the context of automatic PD identification, and Gharehchopogh and Mohammadi [10] used Artificial Neural Networks with Multi-Layer Perceptron to diagnose the effects caused by Parkinson's Disease. Spadotto et al. [11] also considered using a meta-heuristic-driven feature selection aiming at recognizing such illness.

Memedi et al. [12] measured the disease progression in PD patients, which were asked to perform some handwritten exams at home, and Drotár et al. [13] and Taleb et al. [14] also considered handwriting features for PD evaluation, but focused on finding a subset of that features that really matter when diagnosing PD. Lones et al. [15] employed evolutionary algorithms for combining classifiers aiming at the automatic identification of Parkinson's Disease. Pan et al. [16] evaluated the performance of support vector machines with radial basis function to compare the onset of tremors in patients with PD.

* Corresponding author.
E-mail addresses: clayton@fc.unesp.br (C.R. Pereira), papa@fc.unesp.br (D.R. Pereira), gth.rosa@uol.com.br (G.H. Rosa), silke@fmb.unesp.br (C. Hook), christian.hook@oth-regensburg.de (J.P. Papa).

https://doi.org/10.1016/j.artmed.2018.04.001
0933-3657/© 2018 Elsevier B.V. All rights reserved.

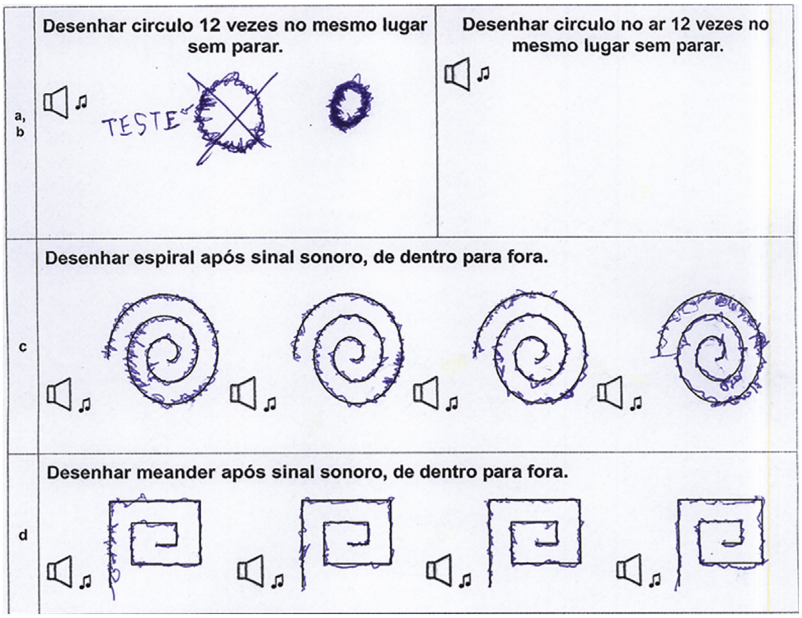

a, b
Desenhar circulo 12 vezes no mesmo lugar sem parar.
TESTE
Desenhar circulo no ar 12 vezes no mesmo lugar sem parar.
c
Desenhar espiral após sinal sonoro, de dentro para fora.
d
Desenhar meander após sinal sonoro, de dentro para fora.

Fig. 1. Handwritten exam adapted from Pereira et al. [21].

Peker et al. [17] employed information from biomedical sound measurements and complex-valued neural networks to aid PD diagnosis as well, and Hariharan et al. [18] developed a new feature weighting method using Model-based clustering (Gaussian mixture model) to enrich the discriminative ability of some dysphonia-based features, achieving 100% of classification accuracy. Very recently, Drotár et al. [19] presented the PaHaW Parkinson's disease handwriting database, which consists of handwriting samples from Parkinson's disease patients and healthy controls. Their goal was to show that both kinematic and pressure features in handwriting can be used for automatic PD diagnosis, and Sadikov et al. [20] aimed at detecting early Parkinson's disease motoric symptoms by means of spirography, i.e., the task of drawing geometrical figures.

Pereira et al. [21] proposed to aid PD diagnosis using images obtained from handwriting movements, as well as they designed a public dataset with hundreds of images containing handwriting drawings from healthy individuals and patients. Later on, the same research group also presented a study that employed Convolutional Neural Networks to learn features from handwritten exams for automatic PD identification [22–24]. Peuker et al. [25] used the signals extracted from a smartpen to perform PD identification, obtaining very suitable results. However, the authors extracted features using a sequential-driven feature selection algorithm, which may be quite costly in terms of computational burden for large datasets.

In this work, we proposed to learn features obtained from handwritten dynamics using a convolutional neural network (CNN) [26], which can process information through a set of layers, being each one in charge of learning a different and finer representation. Another contribution is to make available a dataset composed of the signals extracted from patients and healthy individuals through the smartpen, which is called "NewHandPD"[1]. Additionally, we showed how to improve PD identification by means of an ensemble of CNNs, which were trained over six different handwritten exams: (i) drawing circles on the paper, (ii) drawing circles in the air, (iii) spirals, (iv) meanders, (v) left-wrist movements and (vi) right-wrist movements.

The remainder of this paper is organized as follows. Section 2 presents the methodology employed in this work, as well as the proposed dataset. Section 3 presents the experimental results, and Section 4 states conclusions and future works.

2. Methodology

In this section, we present the methodology used to create the dataset, as well as the proposed approach to analyzing the handwritten dynamics based on Convolutional Neural Networks.

2.1. HandPD dataset

The writing of parkinsonian patients usually faces the so-called micro-graphing, with reduced movement amplitudes, slowness, and rigidity [27]. Also, it is not straightforward to highlight a specific exam that can identify an early-stage patient. Another problem concerns the fact that PD is usually misidentified with other brain disorders quite oftenly.

Recently, Pereira et al. [21] made available a dataset concerning images acquired during handwriting exams, as the one depicted in Fig. 1. The idea of the form is to ask a person to perform some specific tasks that are supposed nontrivial to PD patients, including to trace some geometric shapes and conducting the so-called "diadochokinesis test", which is a test where the individual holds the pen with straight arms and perform hand-wrist movements.

Table 1
Description of the dataset used in this work.

Dataset			
Control		Patient	
Male	Female	Male	Female
6	12	59	15
Average age			
44.22 ± 16.53		58.75 ± 7.51	

[1] http://wwwp.fc.unesp.br/~papa/pub/datasets/Handpd.

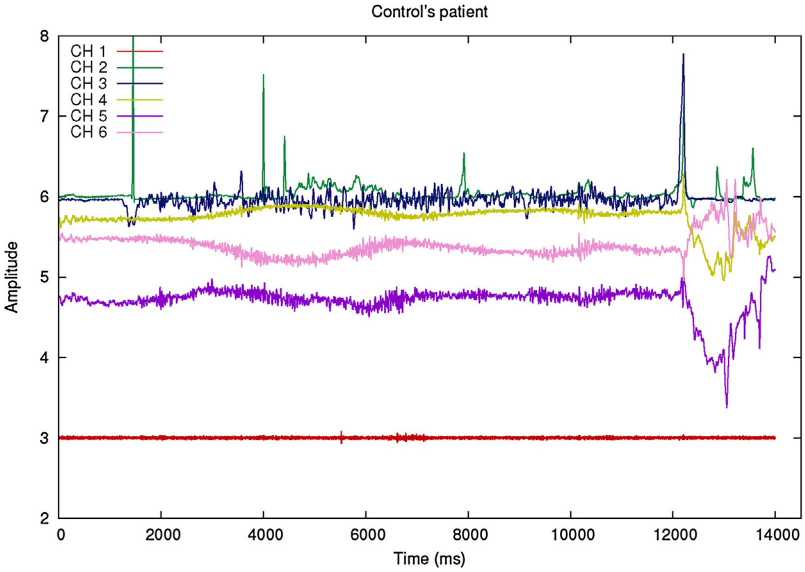


Fig. 2. Signals recorded by the pen from a control individual when drawing a spiral.

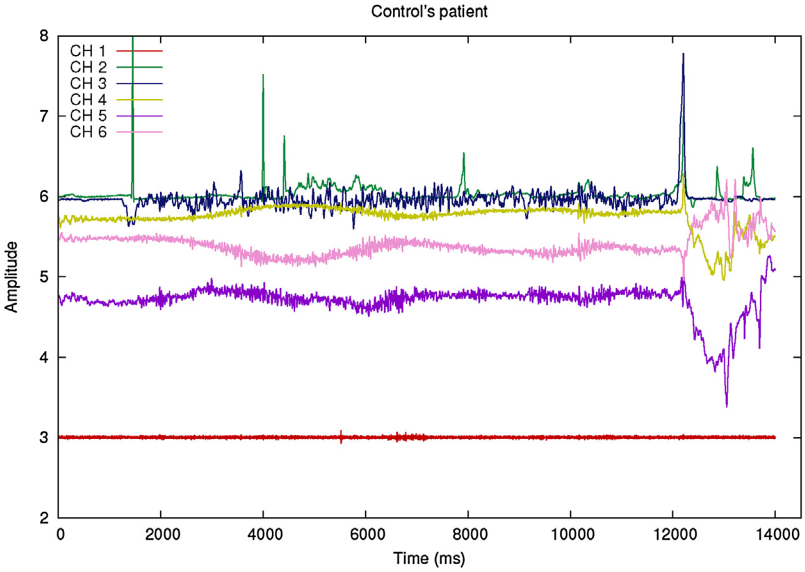


Fig. 3. Signals recorded by the pen from a PD patient when drawing a spiral.

From Fig. 1, one can observe the six handwritten exams: (i) drawing circles on the paper (exam ‘a’), (ii) drawing circles in the air (exam ‘b’), (iii) drawing spirals (exam ‘c’), (iv) drawing meanders (exam ‘d’), (v) right-wrist movements (exam ‘e’), and (vi) left-wrist movements (exam ‘f’). Notice both “spirals” and “meanders” exams are performed four times, and the individual is asked to draw the circle in the air and on the paper twelve times each.

The former HandPD dataset was collected at the Faculty of Medicine of Botucatu, São Paulo State University, Brazil, and it is composed of images extracted from handwriting exams of individuals divided into two groups: (i) healthy people and (ii) PD patients. The dataset comprises 92 individuals that are divided into two groups, as presented in Table 1. Additionally, the control group is composed of 16 right-handed and 2 left-handed individuals, and the patients group contains 69 right-handed and 5 left-handed individuals.

The smartpen contains six sensors, as described below:

- CH 1: Microphone;
- CH 2: Fingergrip;
- CH 3: Axial Pressure of ink Refill;
- CH 4: Tilt and Acceleration in “*X* direction”;
- CH 5: Tilt and Acceleration in “*Y* direction”; and
- CH 6: Tilt and Acceleration “*Z* direction”.

The difference between the exams of healthy individuals and patients are due to a dysfunction of movement disorders. Some parkinsonian patients may present high levels of tremor during drawing tasks. Since each sensor outputs the whole signal acquired during the exam,[2] one can represent such data as a time series, as depicted in Fig. 2, which depicts the output of an exam from a healthy individual when drawing a spiral (e.g., Fig. 5a), and each colored signal stands for a different channel. We can observe the drawing is pretty much the standard form of the image, while the

[2] The extension of the exam is defined as the time interval between a computer beep (a start call) and the end of the drawing process.

signal extracted from the patient seems to figure too much noise, as displayed in Fig. 3 (e.g., Fig. 5b). Also, the microphone (i.e., the red channel) seems to provide little discriminative information between patients and healthy individuals. The "sounds of writing" produced a slightly different output for both types of individuals, but we included them as well.

In order to build the dataset, we used signals extracted from the form tests 'a', 'b', 'c', 'd', 'e' and 'f'. The new dataset comprises 34 individuals, being 14 patients (10 males and 4 females) and 20 control (healthy) individuals (10 males and 10 females). Each person is asked to fill the form out using the smartpen. This activity concerns the analysis of the movement provided by circles, spirals, meanders drawings and diadochokinesis; quantifying the normal motor activity in a healthy individual, as well as the dysfunction of PD patients.

The form tests are divided into six exams to facilitate the dataset organization, as follows:

- Exam 1: drawing a circle twelve times in the same place (row 'a' in Fig. 1);
- Exam 2: drawing a circle twelve times in the air (row 'b' in Fig. 1);
- Exam 3: drawing four spiral from inside to outside (row 'c' in Fig. 1);
- Exam 4: drawing four meanders from inside to outside (row 'd' in Fig. 1);
- Exam 5: diadochokinesis test with the right hand (row 'e' in Fig. 1); and
- Exam 6: diadochokinesis test with the left hand (row 'f' in Fig. 1).

2.2. Modeling time series in CNNs

We modeled the problem of computer-assisted PD identification as an image recognition task using CNNs, where the signals provided by the smartpen are mapped into pictures. Each exam is composed of r rows (exam time in milliseconds) and 6 columns, which stand for the 6 signal channels (e.g., sensors). Therefore, each exam needs to be reshaped to a squared matrix to fulfill our purposes (notice the number of rows r may differ from each test, since a person may take longer than another to perform the exam). After rescaling, each matrix is then normalized to be handled as a gray-scale image. Figs. 4 and 5 illustrate some drawings and their transformed versions into time series-based images. One can observe the different patterns among the test images, as well as different patterns among the same drawings of healthy and PD patients.

Datasets "Exam 1", "Exam 2", "Exam 5" and "Exam 6" contain 76 images each, being 56 from PD patients and 20 from the control group. The datasets "Exam 3" and "Exam 4" contain 304 images each, being 224 from PD patients and 80 from the control group. The difference in the number of images concerns the fact that spirals and meanders are drawn four times per exam. Additionally, for each dataset, we created two versions varying the image size: (i) datasets with images of 64×64 pixels, and (ii) datasets with images of 128×128 pixels. The idea is to evaluate the influence of the image resolution during the experiments.

Fig. 6 displays the handwritten dynamics of a patient and a healthy individual (control group) concerning the six exams. One can realize the difference concerning the images from the patient (first row) and the healthy individual (second row). Since we modeled the handwritten dynamics (time series) as images, the problem now becomes to learn texture-oriented features, since they encode the tremors during the exam.

Additionally, despite the time series depicted in Fig. 6 showed to be quite important to distinguish between control individuals and patients, one can observe a considerable difference among the exams, which means they are important to capture distinct information, as further detailed in Section 3.3. We showed their combination is a powerful tool to enhance the automatic diagnosis concerning both the control and patient group.

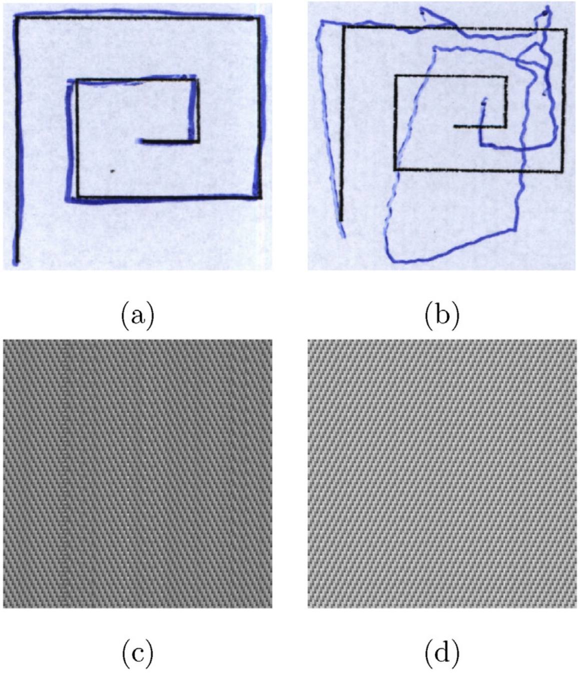


Fig. 4. Meander images from: (a) control and (b) PD patient, and their respective time series-based images in (c) and (d).

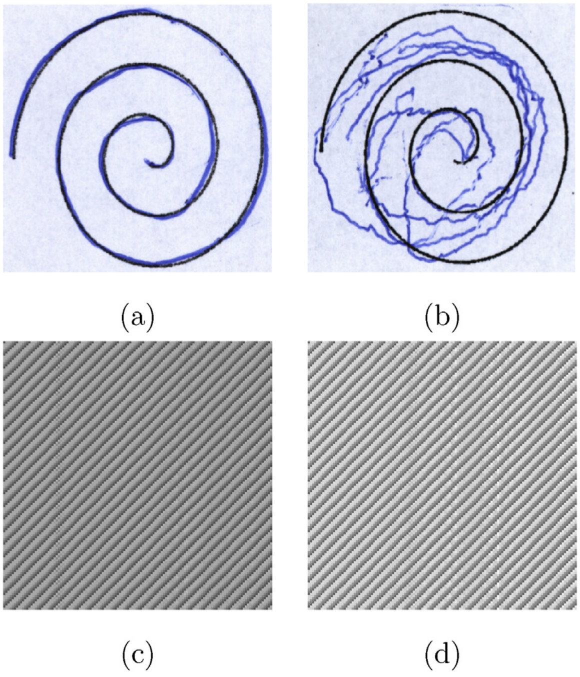


Fig. 5. Spiral images from: (a) control and (b) PD patient, and their respective time series-based images in (c) and (d).

2.3. Assessment through convolutional neural networks

In this section, we explain the proposed approach to assess the automatic diagnosis of Parkinson's disease using convolutional neural networks. The experiments were divided into two rounds: (i) single-assessment and (ii) combined-assessment. In the former

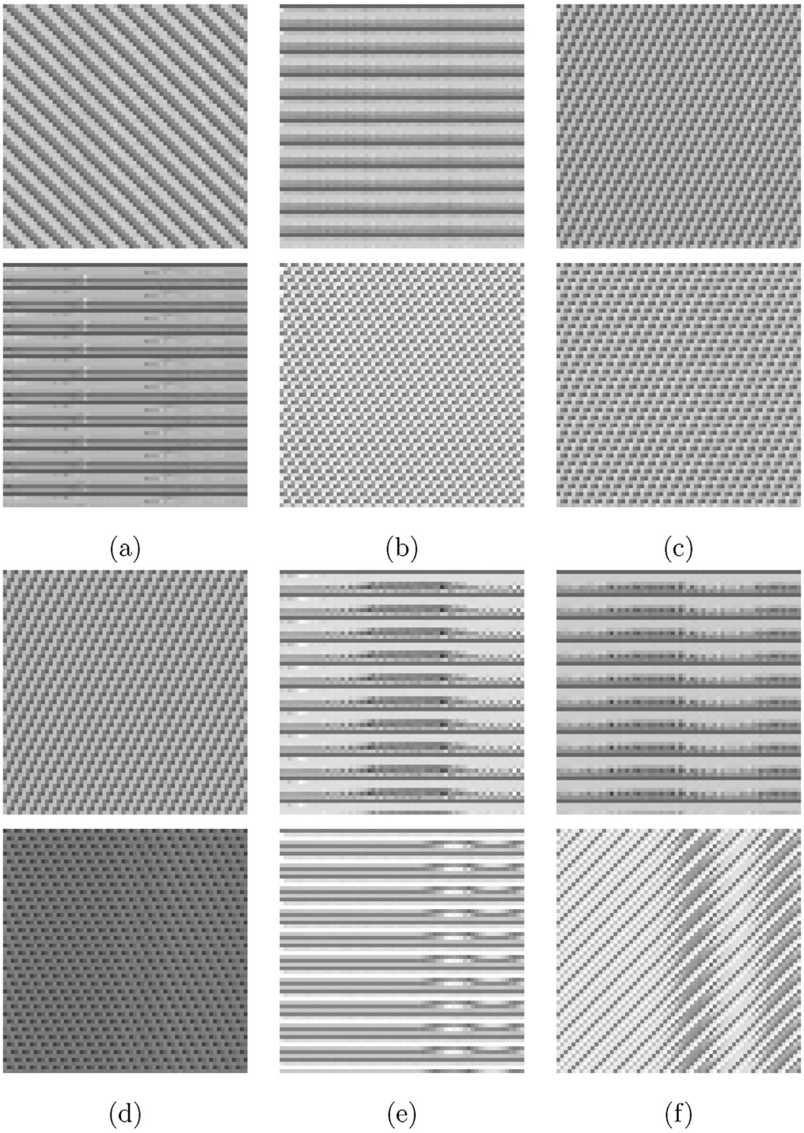


Fig. 6. Time series-based images pattern considering a patient (first and third row) and a healthy individual (second and fourth row): (a) "Exam 1" (circle in the paper), (b) "Exam 2" (circle in the air), (c) "Exam 3" (spiral), (d) "Exam 4" (meander), (e) "Exam 5" (diadochokinesis with the right hand) and (f) "Exam 6" (diadochokinesis with the left hand).

experiment, we analyzed the robustness of the features learned through CNNs considering each exam individually. Regarding the second experiment, we proposed to combine the output of each CNN (i.e., each one trained over a specific exam) using majority voting to obtain the final result. Since we have six different exams, it is quite reasonable to assume that each one encodes/captures different information related to handwriting skills. Fig. 7 depicts the aforementioned proposed combination step to assess the robustness of CNNs when modeling the problem of PD recognition using time series-based images.

3. Experiments

In this section, we present the experimental setup and the results obtained using the proposed methodology.

3.1. Experimental setup

In this work, we used CNNs to classify the time-series-based images drawn by the control group and PD patients. In order to provide more conclusive results, we also considered standard classifiers trained over the data to serve as baselines for comparison purposes: (i) Optimum-path forest (OPF) [8,9], (ii) Support vector machines (SVM) [28] and (iii) Näive-Bayes [29]. Notice SVM parameters have been optimized through a grid-search procedure within the ranges $C \in [2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}]$ and $\sigma \in [2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}]$, in which C and σ stand for the SVM soft-margin parameter and the radial basis function (RBF) kernel variance values, respectively. Additionally, CNN parameters were hand-tuned, i.e., the parameters were empirically chosen.

As one can observe, the images may look discriminative by texture. In order to serve as a baseline experiment, we used two texture descriptors to feed the supervised techniques addressed in this work: the gray level co-occurrence matrices (GLCM) [30], and the local binary patterns (LBP) [31]. GLCM stands for the distribution of co-occurring pixels values concerning a given offset. For our purpose, we used the energy, entropy, contrast, homogeneity and correlation features computed over the matrices built upon the angles θ as 0°, 45°, 90° and 135°. Regarding LBP, it recaps the local structure in an image by comparing each pixel with its neighbors. In

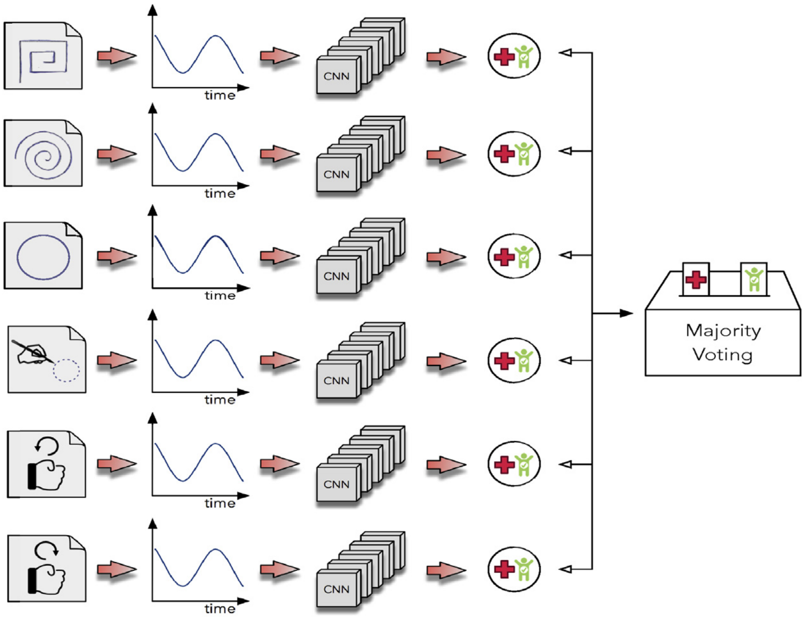


Fig. 7. Proposed combination approach to evaluate CNNs in the context of PD identification.

other words, it labels the pixels by thresholding the neighborhood of each pixel and transforming them into binary numbers. Considering our approach, we used the naïve LBP with 8 neighbors and radius equals to 1.

To provide consistent experiments, we partitioned each dataset considering individuals and not just images. Therefore, when a given individual is selected to compose the training set, all the twelve images of that individual are used for training purposes in all six exams. Notice the same methodology is applied to the testing set. We considered using 50% of the samples to compose the training set, and the remaining 50% to be part of the testing phase. Such percentages were empirically chosen, being reasonable to have the same proportion of samples for both sets. Very often, medical-dependent applications suffer from the lack of data, and it is quite usual to find works that make use of larger training sets to alleviate this problem. In this work, we decided to keep a fair distribution among the sets to evaluate the proposed approach. Also, we assessed the robustness of the experiments over two versions of the datasets: (i) the first one uses images of 64 × 64 pixels, and the second one (ii) applies images of 128 × 128 pixels.

Aiming at evaluating the experiments using statistical analysis, we randomly generated 10 different training and testing sets. The statistical evaluation was carried out using the Wilcoxon signed-rank test with a significance of 0.05 [32]. As mentioned earlier, the experiments were divided into two steps: (i) the first one concerns the single assessment, where CNN, OPF, SVM and Näive-Bayes were evaluated on each exam individually (Section 3.2), and the second step (ii) consists the combined assessment, which considers the output of each classifier over the individual exam in a majority voting-based schema to produce the final results (Section 3.3).

Regarding the source code, we used the well-known Caffe library[3] [33], which is developed under GPGPU (General-Purpose computing on Graphics Processor Units) platform, thus providing more efficient implementations concerning CNNs. Concerning OPF and SVM, we used the LibOPF [34] and libSVM [35], respectively, and for Näive-Bayes we used our implementation. Additionally, we considered two different CNN architectures to provide a more in-depth experimental analysis:

1 ImageNet: composed of 5 convolution layers, 5 pooling layers and 2 normalization layers. It is also constituted by 5 ReLU layers among the convolution ones, 2 inner product layers, 2 dropout layers, 1 softmax loss layer and 1 accuracy layer for testing purposes. The first convolutional layer uses a kernel size of 11 × 11 with stride of 4, and the second convolutional layer employs a kernel size of 5 × 5 with stride of 5 pixels. The next convolutional layers use kernels of size 3 × 3 with stride of 1 pixel.
2 Cifar-10: a quick version is used, composed of 3 convolution layers and 3 pooling layers. It is also constituted by 3 ReLU layers among the convolution ones, 2 inner product layers, 1 softmax loss layer and 1 accuracy layer for testing intentions. All convolutional layers employ kernels of size 5 × 5 with stride of 1.

Since the images used in the experiments are domain-specific, we did not employ transfer learning. In addition, we used 10, 000 training iterations with mini-batches of size 16 concerning CNN experiments.

3.2. Single-assessment

This section aims at presenting the experimental results concerning the CNN-based Parkinson's disease identification over the individual exams. As mentioned earlier in Section 3.1, we compared two distinct CNN architectures and three baseline approaches. Tables 2 and 3 present the average results regarding the single-assessment results. The most accurate results, according to Wilcoxon signed-rank test, are in bold. Table 2 presents the overall accuracy, while Table 3 presents the recognition rates per class, i.e., the sensitivity (true positive) and specificity (true negative) values. Regarding the overall recognition rates, we employed an accuracy measure proposed by Papa et al. [8] that considers unbalanced datasets. Notice we did not employ the well-known receiver operating characteristic (ROC) curve since most of the techniques used in this paper do not produce output probabilities.

One can observe that CNN-based features obtained the most accurate results for all experiments, being SVM similar to CNN-ImageNet concerning "Exam 4" dataset with 128 × 128 images. The results over the images with higher resolution were slightly more accurate, despite the main difference when working with images with different resolution concerns CNN-Cifar10 architec-

[3] http://caffe.berkeleyvision.org.

Table 2
Average overall accuracy over the test set considering the six exams, different image resolutions and classification/feature extractor techniques.

Classifier	Accuracy(%) – 64 × 64					
	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
CNN-ImageNet	**67.75 ± 3.86**	68.04 ± 9.02	**78.02 ± 2.48**	**80.15 ± 2.91**	**72.56 ± 4.72**	**74.23 ± 5.60**
CNN-Cifar10	**68.04 ± 4.17**	**71.62 ± 4.04**	73.77 ± 5.20	**80.13 ± 2.54**	70.93 ± 0.77	66.34 ± 9.98
OPF-Raw	57.26 ± 4.71	59.16 ± 2.31	71.79 ± 1.50	71.67 ± 3.41	61.04 ± 1.71	**74.23 ± 5.60**
SVM-Raw	58.69 ± 3.76	61.44 ± 2.00	74.61 ± 2.50	77.83 ± 5.34	55.49 ± 8.46	66.34 ± 9.98
Bayes-Raw	54.82 ± 4.15	58.11 ± 1.22	73.45 ± 1.76	73.54 ± 3.53	61.40 ± 4.12	55.64 ± 4.81
OPF-GLCM	50.08 ± 7.62	62.40 ± 8.98	68.76 ± 2.40	74.99 ± 2.98	54.38 ± 5.01	54.96 ± 7.22
SVM-GLCM	53.18 ± 6.61	51.01 ± 4.61	71.70 ± 4.42	76.25 ± 3.80	52.00 ± 5.77	54.57 ± 7.13
Bayes-GLCM	58.00 ± 9.82	51.82 ± 5.75	69.82 ± 4.62	72.63 ± 3.02	58.93 ± 9.89	55.39 ± 6.41
OPF-LBP	56.20 ± 6.94	56.17 ± 8.36	60.61 ± 3.93	58.24 ± 3.12	64.33 ± 4.99	59.43 ± 6.15
SVM-LBP	49.40 ± 0.98	50.28 ± 4.55	64.09 ± 7.10	62.29 ± 6.90	61.98 ± 9.89	59.46 ± 7.65
Bayes-LBP	52.24 ± 6.69	58.04 ± 2.96	64.46 ± 4.80	64.43 ± 3.41	64.58 ± 5.58	69.64 ± 8.04
Classifier	Accuracy(%) – 128 × 128					
	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
CNN-ImageNet	**68.04 ± 2.96**	**73.41 ± 3.66**	**78.26 ± 1.97**	**80.75 ± 2.08**	**73.59 ± 3.57**	**76.32 ± 5.18**
CNN-Cifar10	55.46 ± 3.25	61.98 ± 8.52	52.10 ± 19.09	60.94 ± 14.12	52.05 ± 2.81	50.97 ± 2.17
OPF-Raw	58.09 ± 3.13	61.79 ± 2.50	72.83 ± 2.20	76.28 ± 2.91	59.58 ± 3.96	52.86 ± 4.87
SVM-Raw	58.39 ± 7.40	64.61 ± 2.50	77.17 ± 4.00	**80.74 ± 3.22**	58.21 ± 8.61	56.86 ± 6.39
Bayes-Raw	57.55 ± 3.69	63.45 ± 1.76	72.11 ± 3.33	74.38 ± 4.80	62.11 ± 3.47	51.79 ± 3.73
OPF-GLCM	46.40 ± 6.47	57.48 ± 7.23	66.59 ± 2.31	75.97 ± 2.34	55.65 ± 4.14	58.64 ± 3.49
SVM-GLCM	57.11 ± 5.23	53.10 ± 4.85	71.07 ± 3.48	79.46 ± 3.50	52.06 ± 6.06	58.79 ± 7.51
Bayes-GLCM	54.42 ± 5.79	53.04 ± 5.30	68.33 ± 3.81	74.15 ± 3.04	56.83 ± 8.66	54.54 ± 6.78
OPF-LBP	54.63 ± 5.95	55.71 ± 5.97	61.49 ± 4.00	66.09 ± 4.18	64.33 ± 9.35	56.68 ± 6.43
SVM-LBP	50.42 ± 3.60	48.65 ± 4.03	60.51 ± 6.96	65.65 ± 5.46	61.54 ± 8.84	63.32 ± 10.60
Bayes-LBP	60.05 ± 5.20	52.13 ± 5.73	63.65 ± 3.50	63.81 ± 4.13	57.01 ± 5.67	68.50 ± 8.11

Table 3
Average class accuracy over the test set considering the six exams, different image resolutions and classification/feature extractor techniques

Classifier	Accuracy(%) – 64 × 64					
	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
CNN-ImageNet	81.19 (**54.31**)	83.04 (53.04)	89.23 (**67.20**)	**88.91** (71.38)	94.10 (51.03)	85.57 (**62.98**)
CNN-Cifar10	83.68 (**54.20**)	83.12 (**60.12**)	**90.69** (56.86)	85.68 (**74.60**)	85.44 (**56.43**)	73.43 (59.24)
OPF-Raw	79.29 (35.24)	80.49 (37.84)	90.39 (53.20)	85.71 (57.62)	85.71 (36.36)	79.29 (32.00)
SVM-Raw	87.76 (29.52)	87.86 (35.02)	**91.59** (57.64)	**89.46** (66.19)	**96.43** (14.55)	83.57 (48.00)
Bayes-Raw	78.21 (31.43)	79.61 (36.63)	89.73 (57.18)	86.61 (60.48)	86.43 (36.36)	85.00 (34.00)
OPF-GLCM	81.06 (19.11)	85.71 (39.10)	82.05 (55.47)	85.43 (64.52)	77.85 (30.94)	73.93 (36.00)
SVM-GLCM	89.99 (16.37)	89.29 (12.74)	89.12 (54.30)	86.07 (66.42)	90.35 (13.65)	**92.15** (17.00)
Bayes-GLCM	79.64 (36.38)	79.99 (23.66)	86.06 (53.58)	85.26 (60.01)	87.86 (30.02)	81.79 (29.00)
OPF-LBP	77.86 (34.57)	81.42 (30.93)	83.83 (37.39)	79.10 (37.40)	73.23 (55.44)	87.86 (31.00)
SVM-LBP	**96.07** (2.73)	**89.64** (10.92)	88.66 (39.53)	88.38 (36.19)	92.14 (31.83)	83.92 (35.00)
Bayes-LBP	73.56 (30.93)	86.07 (30.03)	81.06 (47.85)	81.24 (47.63)	86.44 (42.76)	89.29 (50.00)
Classifier	Accuracy(%) - **128x128**					
	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
CNN-ImageNet	81.30 (**54.78**)	82.33 (**64.48**)	88.19 (**68.36**)	**90.99** (**70.51**)	94.77 (**52.41**)	92.52 (**60.13**)
CNN-Cifar10	89.74 (21.18)	84.65 (39.31)	84.17 (20.03)	72.81 (49.06)	**98.60** (5.50)	**99.30** (2.64)
OPF-Raw	74.29 (41.90)	82.69 (40.90)	82.32 (63.36)	88.95 (63.62)	86.63 (32.73)	85.71 (20.00)
SVM-Raw	88.21 (28.57)	**88.46** (40.81)	**91.74** (62.62)	**91.46** (**69.99**)	96.43 (20.00)	95.71 (22.00)
Bayes-Raw	76.07 (39.05)	84.05 (42.87)	87.50 (56.73)	86.49 (62.30)	87.86 (36.36)	83.57 (20.00)
OPF-GLCM	76.44 (16.38)	76.78 (38.21)	84.35 (48.82)	86.69 (65.23)	78.56 (32.75)	79.28 (38.00)
SVM-GLCM	87.86 (26.37)	**88.94** (17.29)	89.30 (52.86)	88.94 (69.99)	93.21 (10.92)	88.57 (29.00)
Bayes-GLCM	74.28 (34.56)	76.07 (30.02)	83.56 (53.10)	85.44 (62.84)	88.21 (25.48)	81.06 (28.00)
OPF-LBP	81.08 (28.20)	71.44 (40.02)	83.93 (39.04)	85.27 (46.91)	83.21 (45.47)	80.33 (33.00)
SVM-LBP	**93.58** (7.28)	**88.20** (9.10)	86.25 (34.76)	84.65 (46.67)	90.36 (32.74)	94.64 (32.00)
Bayes-LBP	74.64 (45.47)	76.09 (28.21)	82.76 (44.53)	79.99 (47.60)	88.58 (25.48)	95.00 (42.00)

ture, which has been considerably affected by images with higher resolution. Since we used a "quick version" of Cifar10 architecture, i.e., a shallower network, images with higher resolution may require more neurons, and thus a deeper neural network. Therefore, we can conclude that the CNN-Cifar10 network may have overfitted the data. Although the images depicted in Fig. 6 appear to be texture-oriented, they do not reflect the whole dataset, which can explain the results using GLCM and LBP feature extraction algorithms. Interestingly, the raw data achieved similar results to the texture-based ones, which means the real problem is the identification of early-stage-affected patients, as discussed later on Section 3.4.

Additionally, "Exam 4" (i.e., meanders) has the better discriminative ability since it obtained the best recognition rates concerning the different resolutions and architectures. A finer statistical evaluation using Wilcoxon test pointed out that both "Exam 3" and "Exam 4" are similar to each other concerning 64 × 64 images, but "Exam 4" is the single best approach concerning 128 × 128 images. However, a closer look at the recognition rates shows us a quite similar behavior among the different image resolution experiments.

Table 4
Average overall accuracy over the test set considering the combined-assessment approach.

Classifier	Accuracy(%) – **Voting**	
	64 × 64	128 × 128
CNN-ImageNet	**93.42 ± 3.17**	**95.74 ± 1.60**
CNN-Cifar10	90.28 ± 6.09	76.96 ± 21.60
OPF-Raw	79.44 ± 2.67	82.44 ± 4.54
SVM-Raw	74.80 ± 9.32	70.56 ± 3, 56
Bayes-Raw	85.51 ± 1.98	83.45 ± 3.48
OPF-GLCM	**93.01 ± 3.64**	**95.21 ± 2.13**
SVM-GLCM	79.23 ± 8.23	**95.21 ± 2.13**
Bayes-GLCM	**93.05 ± 3.97**	93.48 ± 2.78
OPF-LBP	90.21 ± 2.56	91.14 ± 4.56
SVM-LBP	69.98 ± 14.43	75.03 ± 9.57
Bayes-LBP	91.14 ± 2.68	88.17 ± 4.59

Table 5
Average class accuracy over the test set considering the combined-assessment approach.

Classifier	Accuracy(%) – **Voting**	
	64 × 64	128 × 128
CNN-ImageNet	97.84 (**89.00**)	97.48 (**94.01**)
CNN-Cifar10	98.56 (82.00)	98.92 (55.00)
OPF-Raw	**100.00** (58.87)	**100.00** (64.33)
SVM-Raw	**100.00** (49.59)	**100.00** (41.11)
Bayes-Raw	**100.00** (71.01)	**100.00** (66.90)
OPF-GLCM	98.02 (88.00)	98.92 (91.50)
SVM-GLCM	94.97 (63.50)	98.92 (91.50)
Bayes-GLCM	99.10 (87.00)	99.46 (87.50)
OPF-LBP	98.92 (81.50)	99.28 (83.00)
SVM-LBP	99.46 (40.50)	98.56 (51.50)
Bayes-LBP	99.28 (83.00)	97.84 (78.50)

Therefore, without loss of generality, we can assume both "Exam 3" and "Exam 4" are similar to each other when we consider the overall (global) accuracy.

Table 3 presents the average results per class using the following format $x(y)$, where x and y stand for the accuracy concerning the patients and the control group, respectively. Since our dataset is not balanced, it is quite useful to provide the recognition rates per class. Although we have more control people than PD patients, a considerable number of healthy individuals were classified as patients, since the dataset comprises PD patients with exams quite alike to the ones performed by healthy individuals. Texture-based and raw data information achieved reasonable recognition rates concerning the patients group, but their global accuracy were considerably affected due to the effectiveness over the control group. As mentioned earlier, we assume two types of problems: (i) first, the dataset is not balanced (less control group), and (ii) we have a number of early-stage patients, which means they behave similarly to healthy individuals. We observed a very few control individuals that have similar drawings to a patient's one in his/her advanced state of disease. Probably, the individual was affected by another disease, he/her was nervous during the exam, or even the exam was labeled incorrectly.

The best results concerning 64 × 64 images were obtained by SVM over "Exam 5" for patient identification (96.43%), and the more accurate recognition rates considering the control group were obtained by CNN-Cifar10 over "Exam 4". Regarding 128 × 128 images, CNN-Cifar10 obtained 99.30% of recognition rate concerning patients, and CNN-ImageNet obtained 70.51% of accuracy considering the control group. Meanders ("Exam 4") appear to be the most important test to identify healthy people, and "Exam 5" and "Exam 6" (right- and left-wrist movements) seemed to be the best ones to recognize PD patients. Although we can observe pretty much different exams (Figs. 4 and 5) between patients and control group, the problem gets worse when we have patients at the very early stage of the disease, which may have quite close exams. However, the left- and right-wrist movements ("Exam 5" and "Exam 6") seems to detect some subtle disorders in the motor system, thus obtaining better results.

3.3. Combined-assessment

In this section, we present the results concerning the proposed approach that considers the combination of all decisions made by the classifiers trained on each exam (Fig. 7). We also tried the harmonic-weighted voting, since the number of exams is small for combination purposes (i.e., eight), the results were pretty much the same as the standard majority voting. Therefore, we opted to show the results concerning the latter approach only.

Table 4 presents the overall (global) accuracy, being the best results in bold according to Wilcoxon statistical test. One can observe that the proposed approach improved the results presented in Table 3, confirming our hypothesis that different exams encode/model different handwritten dynamic properties. Also, CNN-ImageNet obtained the best results so far, being consistently more accurate than the compared approaches. Notice that GLCM-based features obtained results similar to the ones achieved by CNN-ImageNet when we use the ensemble of classifiers. However, as one can observe in Table 5, CNNs are more consistent with the different exams, i.e., they figured as the best learners for all exams, which does not happen when dealing with the texture descriptors.

Table 5 presents the results for each class using the very same format employed in the previous section (sensitivity and specificity), i.e., the number in parenthesis stands for the mean accuracy concerning the control group. Once again, the ensemble of CNNs provided a considerable enhancement considering the recognition rates for both patients and control group. Also, the baseline classifiers can benefit from such process, with their accuracies increased for both patients and healthy people recognition. We can conclude that the best trade-off between patient and control group recognition was obtained using CNN-ImageNet with 128 × 128 images. In this case, higher resolution images played a significant role, despite the individual experiments have highlighted that different image sizes are not so crucial to the classification process.

3.4. Early stage detection

The main challenge regarding PD concerns its detection at the early stages, where the symptoms are almost imperceptible. To evaluate the robustness of the proposed approach when identifying patients with early-stage Parkinson's Disease, we have manually selected eight patients with very similar traces to healthy individuals. Fig. 8 presents some images of the selected individuals. Notice that these patients can perform the tasks nearly to a healthy individual. Notice we considered only spirals and meanders in this section since they provided the best results.

Fig. 9 displays some time series extracted from the exams concerning Fig. 8. One can observe that the signals extracted from those exams present an oscillatory behavior, i.e., the pen can capture subtle movements during the exam, which can not be observed in the handwritten exams using visual inspection. The time series depicted in Fig. 9 are pretty much similar to those presented in Fig. 3, which shows the signals extracted from an advanced-stage patient.

For each selected patient, we computed the accuracy of both meander and spiral images considering the very same datasets generated in the previous experiments. Table 6 presents the accuracy concerning these selected images. One can draw two

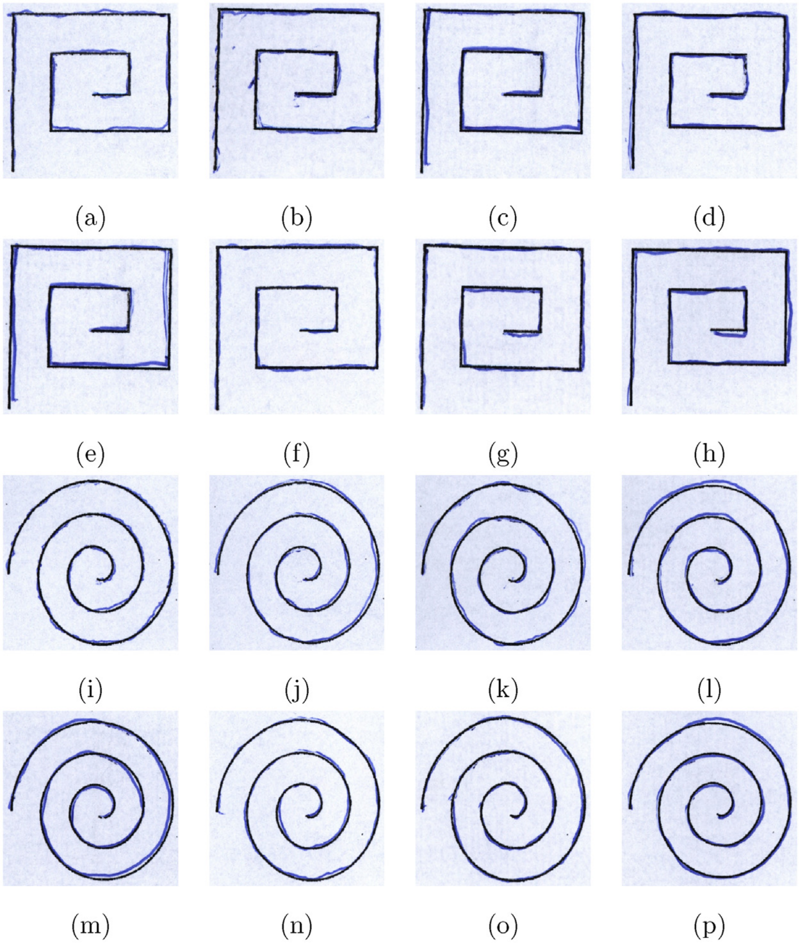


Fig. 8. Examples of meander (first and second row) and spiral (third and fourth row) images obtained by some patients in the early stages of the disease.

Table 6
Average accuracy over the early stage (selected) images.

Classifier	Exam 3	Exam 4
Accuracy (%) 64 × 64		
CNN-Cifar10	**95.83 ± 7.04**	**94.01 ± 6.72**
CNN-ImageNet	**96.35 ± 8.08**	**94.01 ± 6.23**
Bayes-Raw	52.08 ± 46.09	51.04 ± 48.57
OPF-Raw	51.04 ± 46.63	52.20 ± 45.49
SVM-Raw	50.00 ± 47.53	50.00 ± 47.91
Bayes-GLCM	52.08 ± 43.35	51.04 ± 50.28
OPF-GLCM	47.91 ± 44.93	50.00 ± 53.45
SVM-GLCM	53.15 ± 49.17	47.91 ± 49.75
Bayes-LBP	51.04 ± 47.02	55.20 ± 42.47
OPF-LBP	48.95 ± 47.02	54.16 ± 39.08
SVM-LBP	52.08 ± 47.50	51.04 ± 50.48
Accuracy (%) 128 × 128		
CNN-Cifar10	**95.83 ± 7.04**	**94.01 ± 6.72**
CNN-ImageNet	**96.35 ± 8.08**	**94.01 ± 6.23**
Bayes-Raw	47.91 ± 44.04	51.04 ± 46.83
OPF-Raw	48.95 ± 48.57	50.00 ± 45.74
SVM-Raw	50.00 ± 51.63	47.91 ± 49.76
Bayes-GLCM	54.16 ± 45.42	52.08 ± 51.31
OPF-GLCM	53.12 ± 48.37	53.12 ± 50.38
SVM-GLCM	52.08 ± 51.51	53.12 ± 47.96
Bayes-LBP	42.70 ± 40.44	45.83 ± 37.26
OPF-LBP	43.74 ± 44.09	43.75 ± 39.27
SVM-LBP	46.85 ± 44.52	51.04 ± 42.82

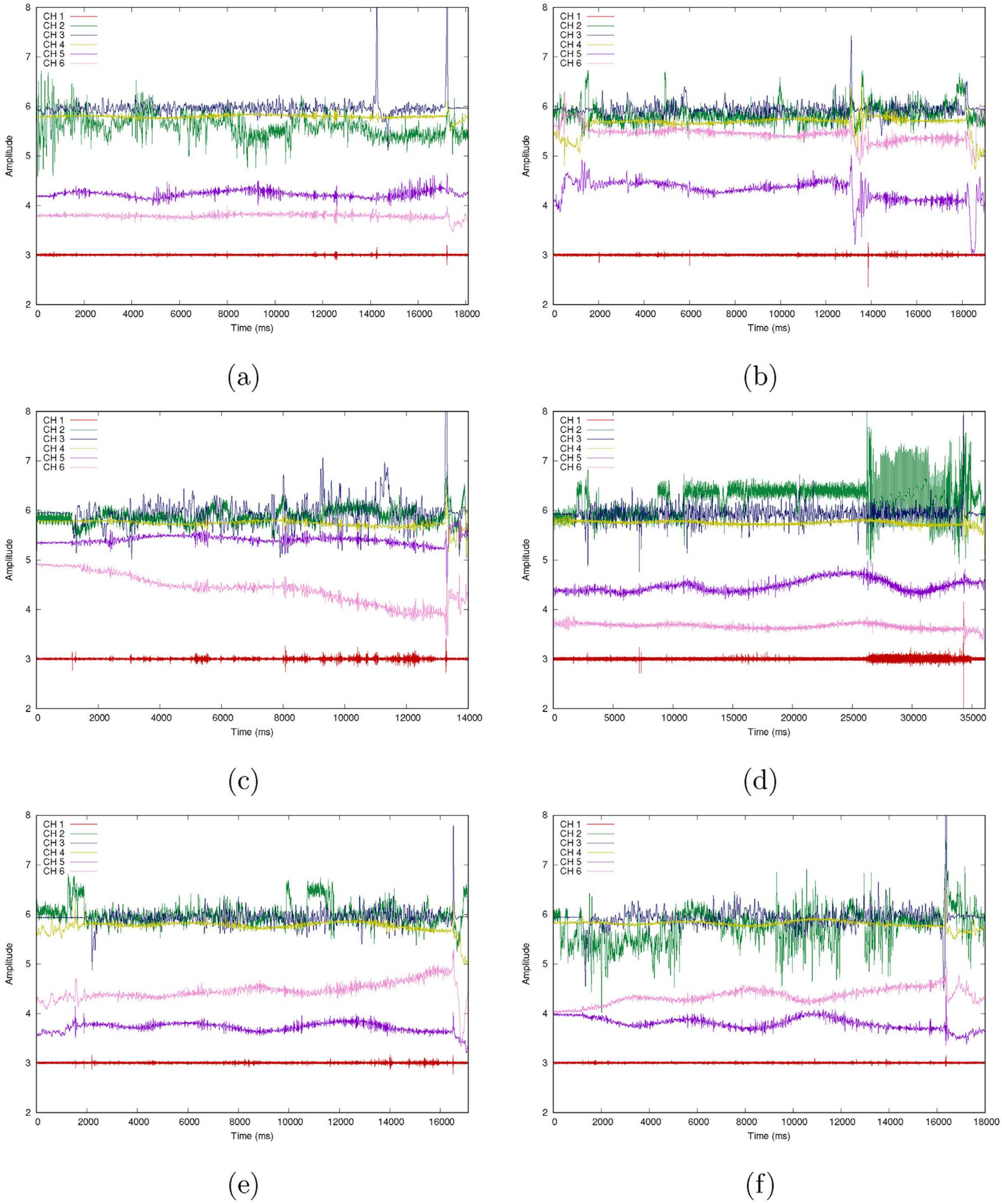
main conclusions: (i) first, CNN-based features (Cifar10 and ImageNet) presented the highest accuracy rates compared to other approaches, and (ii) the proposed approach is robust enough to detect early-stage PD patients since they obtained quite good recognition rates (above 94%). The high standard deviation values concerning raw and texture-based features indicate they can either recognize or miss the majority of early-stage PD patients.

3.5. Discussion

In this section, we present a discussion about the results obtained in this work concerning the ones obtained previously by our research group. Pereira et al. [21] proposed handcrafted features based on the images extracted from the exams, and they achieved 65.88% and 66.36% of recognition rates concerning the spirals and meanders, respectively. The main problem was related to the identification of control group, which degraded the global accuracy.

Later on, Pereira et al. [23] obtained accuracies of around 84.42% and 83.77% concerning meanders and spirals, respectively, using features learned from Convolutional Neural Networks. However, they considered the signals generated from the smartpen as time-series images, instead of data from the drawings. Similarly, but now using a CNN fine-tuned using Bat Algorithm, the same group of authors achieved around 84.35% of recognition rates considering meanders [22].

The approach proposed in this paper, which combines data from six different exams using features learned from CNNs over time-series images, obtained an accuracy nearly to 93.50%, thus outperforming by far all the aforementioned previous works. We showed that mapping handwritten dynamics to time-series images for further feature learning using Convolutional Neural Networks



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 9. Time series concerning the patient group: (a) Fig. 8a, (b) Fig. 8c, (c) Fig. 8e, (d) Fig. 8i, (e) Fig. 8k, and (f) Fig. 8m.

is a very promising approach, mainly when one considers different exams for combination purposes.

A closer look at the single-assessment results (Section 3.2) reveals some interesting conclusions. First, convolutional neural networks showed to be pretty much suitable to learn features from the time-series data. As expected, the baseline approaches that worked on the raw data did not obtain satisfactory results. They may not be able to capture the spatial information about texture neither temporal data. Similar behavior can be observed for GLCM features, which may highlight that mapping signals to images requires a more robust classifier than just texture or any other information. CNNs can capture information from different levels, even texture, but outputs a high-dimensional feature vector at the end of the process.

Regarding the combined assessment presented in Section 3.3, one can observe that the difference among the techniques concerning the recognition rates has decreased considerably, which means that information from different sources can alleviate the gap of knowledge from one single source, as well as it can dramatically improve the recognition of individual classifiers that performed poorly.

4. Conclusions

In this paper, we deal with the problem of computer-assisted Parkinson's disease identification using Convolutional Neural Networks. We propose to map signals extracted from handwriting dynamics into images that can be further used to feed a Convolutional Neural Network. The main contributions of this paper are: (i) to employ a deep learning-oriented approach to aid Parkinson's Disease diagnosis, (ii) to design a signal-based dataset composed of features related to handwritten dynamics, and (iii) to propose an ensemble of CNNs to better distinguish PD patients from control group.

The experimental section comprised different CNN architectures, as well as images with different resolutions and distinct training set sizes. The results obtained by CNNs were compared against the raw data and texture descriptors classified using the traditional pattern recognition techniques. These results show to be very promising, since CNNs were able to learn essential information to differentiate PD patients from healthy individuals, thus obtaining promising results. The ensemble of CNNs was able to capture different information from each exam, thus providing considerably better results.

In regard to future works, we intend to combine the original image obtained through the exam together with the time-series-based version. Also, we plan to apply Auto-encoders right after CNNs to reduce the dimensionality of the feature space.

Acknowledgements

The authors would like to thank FAPESP grants #2010/15566-1, #2013/07375-0, #2014/16250-9, #2014/12236-1, #2015/25739-4, and #2016/19403-6, Capes, and CNPq grants #306166/2014-3 and #307066/2017-7.

References

[1] Maki BE, McIlroy WE. Change-in-support balance reactions in older persons: an emerging research area of clinical importance. Neurol Clin 2005;23(3):751–83.
[2] Marchetti GF, Whitney SL. Older adults and balance dysfunction. Neurol Clin 2005;23(3):785–805.
[3] Zhao YJ, Tan LCS, Lau PN, Au WL, Li SC, Luo N. Factors affecting health-related quality of life amongst Asian patients with Parkinson's disease. Eur J Neurol 2008;15(7):737–42.
[4] Prashanth R, Roy SD, Mandal PK, Ghosh S. High accuracy classification of Parkinson's disease through shape analysis and surface fitting in 123i-ioflupane SPECT imaging. IEEE J Biomed Health Inf 2016;PP(99):1.
[5] Fahn S, Oakes D, Shoulson I, Kieburtz K, Rudolph A, Lang A, et al. Levodopa and the progression of Parkinson's disease. N Engl J Med 2004;351(24):2498–508.
[6] Das R. A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Syst Appl 2010;37(2):1568–72.
[7] Spadotto AA, Guido RC, Papa JP, Falcão AX. Parkinson's disease identification through optimum-path forest. International conference of the IEEE engineering in medicine and biology society 2010:6087–90.
[8] Papa JP, Falcão AX, Suzuki CTN. Supervised pattern classification based on optimum-path forest. Int J Imaging Syst Technol 2009;19(2):120–31.
[9] Papa JP, Falcão AX, Albuquerque VHC, Tavares JMRS. Efficient supervised optimum-path forest classification for large datasets. Pattern Recogn 2012;45(1):512–20.
[10] Gharehchopogh FS, Mohammadi P. Article: a case study of Parkinsons disease diagnosis using artificial neural networks. Int J Comput Appl 2013;73(19):1–6.
[11] Spadotto AA, Guido RC, Carnevali RF, Pagnin AF, Papa JP, Falcão AX. Improving Parkinson's disease identification through evolutionary-based feature selection. International Conference of the IEEE Engineering in Medicine and Biology Society 2010:7857–60.
[12] Memedi M, Nyholm D, Johansson A, Pålhagen S, Willows T, Widner H, et al. Validity and responsiveness of at-home touch screen assessments in advanced Parkinson's disease. IEEE J Biomed Health Inf 2015;19(6):1829–34.
[13] 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(3):508–16.
[14] Taleb C, Khachab M, Mokbel C, Likforman-Sulem L. Feature selection for an improved Parkinson's disease identification based on handwriting. In: 1st international workshop on arabic script analysis and recognition. ASAR; 2017. p. 52–6.
[15] Lones MA, Smith SL, Alty JE, Lacy SE, Possin KL, Jamieson DRS, et al. Evolving classifiers to recognize the movement characteristics of Parkinson's disease patients. IEEE Trans Evolut Comput 2014;18(4):559–76.
[16] Pan S, Iplikci S, Warwick K, Aziz TZ. Parkinson's disease tremor classification, a comparison between support vector machines and neural networks. Expert Syst Appl 2012;19:10764–71.
[17] Peker M, Sen B, Delen D. Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm. J Healthc Eng 2015;6(3):281–302.
[18] Hariharan M, Polat K, Sindhu R. A new hybrid intelligent system for accurate detection of Parkinson's disease. Comput Methods Prog Biomed 2014;11(3):904–13.
[19] 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.
[20] Sadikov A, Groznik V, Možina M, Žabkar J, Nyholm D, Memedi M, et al. Feasibility of spirography features for objective assessment of motor function in Parkinson's disease. Artif Intell Med 2017, http://dx.doi.org/10.1016/j.artmed.2017.03.011.
[21] Pereira CR, Pereira DR, Silva FA, Masieiro JP, Weber SAT, Hook C, et al. A new computer vision-based approach to aid the diagnosis of Parkinson's disease. Comput Methods Prog Biomed 2016;136:79–88.
[22] Pereira CR, Pereira DR, Papa JP, Rosa GH, Yang X-S. Convolutional neural networks applied for Parkinson's disease identification. Springer International Publishing; 2016. p. 377–90.
[23] Pereira CR, Weber SAT, Hook C, Rosa GH, Papa JP. Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics. Proceedings of the SIBGRAPI 2016 – conference on graphics, patterns and images 2016:340–6.
[24] Afonso LCS, Pereira CR, Weber SAT, Hook C, Papa JP. Parkinson's disease identification through deep optimum-path forest clustering. 30th SIBGRAPI Conference on Graphics, Patterns and Images 2017:163–9.
[25] Peuker D, Scharfenberg G, Hook C. Feature selection for the detection of fine motor movement disorders in Parkinson's patients. In: Advanced research conference, ARC '11. Shaker Verlag; 2011.
[26] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989;1(4):541–51.
[27] Gemmert WAV, Teulings HL, Stelmach GE. Parkinsonian patients reduce their stroke size with increased processing demands. Brain Cogn 2001;47(3):504–12.
[28] Cortes C, Vapnik V. Support vector networks. Mach Learn 1995;20:273–97.
[29] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York, USA: Wiley-Interscience; 2000.
[30] Haralick RM, Shanmugam K, et al. Textural features for image classification. IEEE Trans. Syst. Man Cybern 1973;6:610–21.
[31] Ojala T, Pietikä inen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recogn 1996;29(1):51–9.
[32] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull 1945;1(6):80–3.
[33] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: convolutional architecture for fast feature embedding; 2014 arXiv:1408.5093.
[34] Papa JP, Suzuki CTN, Falcão AX. LibOPF: a library for the design of optimum-path forest classifiers, software version 2.1; 2014. Available from: http://www.ic.unicamp.br/afalcao/libopf/index.html.
[35] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol 2011;2(3):1–27.