Parkinson Disease Early Detection using EEG Channels Cross-Correlation

Gabriel Silva¹, Marco Alves¹, Rui Cunha¹, Bruno C. Bispo² and Pedro M. Rodrigues³*

¹Universidade Católica Portuguesa, CBQF - Centro de Biotecnologia e Química Fina - Laboratório Associado, Escola Superior de Biotecnologia, Rua Diogo Botelho 1327, 4169-005, Porto, Portugal.
²Department of Electrical and Electronic Engineering, Federal University of Santa Catarina, 88040-370, Florianópolis, SC, Brazil.
*Corresponding author’s email: prodrigues@porto.ucp.pt

ABSTRACT

Background: Nowadays, Parkinson’s Disease (PD) is the second most common neurodegenerative disease worldwide and has no known cure. Researchers are focused on delaying the disease progression and, to this end, it is necessary to detect PD as early as possible.

Methods: This work proposes a method to detect early stages of PD using the cross-correlation function between selected electroencephalogram (EEG) channels, which is used to find time delays at scalp level in order to serve as a metric for distinguishing study groups. Genetic algorithms and a binary classifier (X-ROC) were used to preselect the best features combinations to feed a neural network for data classification.

Results: This novel approach achieved an overall accuracy of 92.66%.

Conclusion: The applied method detects relevant early PD activity at scalp level in right temporal, frontal, parieto-occipital and occipital regions with high accuracy.

Keywords: Parkinson’s Disease, cross-correlation, electroencephalogram, classification, method.

INTRODUCTION:

Parkinson’s disease (PD) is one of the most common neurodegenerative disorders [1] and it is believed that it affects 0.3% of the global population [2]. PD normally occurs in people aged from 50 to 60 and is not common before 50’s. When it appears before age 21, genetic factors are presumed to be the cause [3].

The origin of PD remains unknown until today, although it is believed that the interaction between environmental and genetic factors may explain its appearance [4]. The lack of a cure for PD makes it important to research a method for early stages detection in order to delay the disease progression and its symptoms, thereby improving the patient’s quality of life [5].

Patients diagnosed with PD are known by the rest tremor but also present other symptoms, such as: muscular rigidity, bradykinesia or postural instability [6]. Moreover, they present an accumulation of α-synuclein, which leads to the formation of Lewy bodies and a progressive loss of dopaminergic neurons in the substantia nigra [7–9]. In order to describe how the symptoms of PD progress, the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Hoehn and Yahr scale are commonly used [10].

Although these aforementioned physiologic patterns are not unique to PD, they can be used together to diagnose Parkinson [2]. Currently, there are some techniques to diagnose PD: cognitive impairment tests, techniques based on biological markers and techniques based on physiological signals analysis.

Cognitive impairment tests are clinical tests that indicate how well the brain and nervous system are working [11]. As examples, one can mention the Montreal Cognitive Assessment (MoCA), Hopkins Verbal Learning Test-Revised (HVLT-R), Symbol Digit Modalities Test (SDMT) and Mini-Mental State Examination (MMSE) [11,12].

The biological markers can be detected through non-motor symptoms, cerebral fluid tests and imaging modalities as the Functional Magnetic Resonance Imaging (fMRI) and Single-Photon Emission Computed Tomography (SPECT). These biological markers can be used to predict if a person will develop PD, control the progression of the disease and analyses the effects of the medication prescribed to a diagnosed patient [13].

With regard to the physiological signals, the Magnetoencephalography (MEG) and Electroencephalography (EEG) are used to record electromagnetic signals from brain activity. Both techniques are apparently complementary to study brain electrical activity but MEG recording offers some important advantages over EEG as, for example, the resistant properties of extra-cerebral tissues do not affect the magnetic fields and the magnetic recording does not depend on any reference point [14].

But MEG also has limitations, including the need for an armored room to correctly register the signals. The armored room involves a high cost and probably because of this, its use has not been very generalized [14]. Furthermore, because of its portability, low cost and capacity of representing the brain electrical activity intensity with a great temporal resolution over time [15], the EEG is the most common exam in medical clinics [13].

In recent years, computational methods have been developed to assist in early-stage PD diagnose. In [16], MEG signals were transformed into finite symbol sequences and the sequences complexity was measured through the so-called Lempel-Ziv complexity (LZC). A forward stepwise linear discriminant analysis (LDA) was applied to investigate the best combination of the LZC results, achieving an accuracy of 84.21% using the results from the right parental and temporal cortical areas.

In [17], an independent component analysis (ICA) was applied to resting-state fMRIs in order to create basal ganglia networks (BGN) and the functional connectivity of these BGNs was characterized by an average parameter. A connectivity threshold for optimal group separation was tested and an
accuracy of 85% was obtained.

The authors in [18] combined MRI extracted features and Positron Emission Tomography (PET) to detect nigrosome 1 asymmetries using 3T equipment with an accuracy level of 94.6%.

In [19], several biological markers, namely, olfactory loss score, rapid eye movement sleep behavior disorder (RBD) score, biomarkers from cerebrospinal fluid (CSF), and striatal binding ratios (SBR) from SPECT imaging, were combine to diagnose early-stage PD. Several classifiers were tested and the best performance was an accuracy of 96.40% achieved using a support vector machine (SVM).

The work presented in [20], various cognitive impairment tests, like MoCA, HVLT-R and SDMT, were used to access the present clinical ability in order to detect the Mild Cognitive Impairment (MCI) in early PD stage, reaching an overall accuracy of 67.4%.

In [10], the authors utilized a set of MDS-UPDRS features (Cognitive impairment, Hallucination and Psychosis, depressed mood, Sleep Problems, Handwriting, Speech, Tremor, Gait and several others) and the Hoehn and Yahr scale, achieving an accuracy of 97.46% by using an AdaBoost-based classifier.

The authors in [21] proposed a method for early-stage PD diagnose from EEG where it was demonstrated that control subjects showed reliable EEG habituation to novel events (auditory oddball paradigm) while PD patients did not. Then, some spectral features and seven temporal factors were extracted from the EEG signals through a principal component analysis (PCA). Using an SVM as a classifier, an accuracy of 82% was achieved, representing the best result so far for the considered dataset.

In [22], an convolutional neural network (CNN) was proposed to learn higher level features from EEG signals, achieving an accuracy of 88.25% in detecting early PD stages. In the study presented in [23], imaging features, like gray matter and white matter voxel distribution, were extracted using images from Magnetic Resonance Imaging (MRI) in order to create a brain regions network model, where each region is represented by a feature vector. Random Forests (RF) algorithm and SVM were used to feature selection and classification, respectively, achieving an accuracy of 93%.

The authors in [24] reported an accuracy of 85.4% when evaluating possible differences in midbrain T1w/T2w ratio between controls and PD patients by using both voxel-based and region-of-interest approaches in normalized space. The study presented in [25] used volatile biomarkers to differentiate, for the first time, study groups with non-medicated patients and an accuracy level of 81% was obtained.

This work proposes a new approach to characterize early-stage Parkinson’s disease through the EEG channel signals. The metric is based on the time delay shifts between the recording EEG channel signals, which are estimated by means of the cross-correlation function. This work is organized as follows: in Section II, the materials and methods are described in detail; in Section III, the results obtained by the proposed metric are presented and discussed; finally, Section IV concludes the paper emphasizing its main contributions.

### MATERIALS AND METHODS

#### The Dataset:

This work used an EEG database from the patient repository for EEG data + computational tools (PRED+CT) [21]. This database is composed of 56 EEG signals, where 28 ones correspond to patients diagnosed with early-stage PD (patients’ group) and other 28 correspond to individuals that are not affected by PD (control group). The EEG signals were sampled at 500 Hz and recorded in a state of total relaxation from 67 electrodes placed in the scalp according to the international 10-20 EEG system [21]. Along with the EEG signals, the database provides the MMSE score for each volunteer. The MMSE is a cognitive impairment test used to evaluate dementia in individuals. It is built around a series of questions focused on the memory, orientation and ability of writing and reading [26]. As dementia is related to PD and other neurodegenerative diseases, this mental examination is then used as a diagnostic tool. The MMSE average score and the average age, as well as their standard deviations, of each study group are described in Table 1.

#### Cross-Correlation:

The Cross-correlation is a measure of the linear relationship between two different signals as a function of time delay. It is generally used to quantify the similarity between signals and to detect delays between them [27].

Due to aforementioned capability, the proposed method uses the cross-correlation to estimate the time delay between EEG signals of different channels. The cross-correlation function between the real-valued signals x(n) and y(n) is defined as [27]

\[
\rho_{xy}(m) = \begin{cases} 
\sum_{n=0}^{N-m-1} x(n + m)y(n), & m \geq 0 \\
-\hat{\rho}_{xy}(m), & m < 0
\end{cases}
\]  

(1)

while its unbiased version is defined as [28]

\[
\hat{\rho}_{xy,\text{unbiased}}(m) = \frac{1}{N-|m|}\rho_{xy}(m),
\]  

(2)

where m is the time delay (commonly called lag) and N is the signals length. Moreover, the cross-correlation coefficient between x(n) and y(n) is given by [28]

\[
\rho_{xy} = \frac{\hat{\rho}_{xy}(0) - \mu_x \mu_y}{\sigma_x \sigma_y},
\]  

(3)

where \(\mu_x\) and \(\mu_y\) are the mean values of the signals x(n) and y(n), respectively, and \(\sigma_x\) and \(\sigma_y\) are their standard deviations. Note that \(-1 \leq \rho_{xy} \leq 1\), where \(\rho_{xy} = 1\) indicates total positive linear correlation, \(\rho_{xy} = 0\) indicates no linear correlation, and \(\rho_{xy} = -1\) indicates total negative linear correlation [28].

<table>
<thead>
<tr>
<th>Table 1: Dataset information.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Volunteers</td>
</tr>
<tr>
<td>Age Average</td>
</tr>
<tr>
<td>MMSE Score Average</td>
</tr>
</tbody>
</table>
Method:

The proposed method can be divided into four steps as illustrated in Figure 1: data preprocessing, data processing, pattern extraction, organization and selection, and classification. All steps were performed using MATLAB 2018b® and are detailed below.

Data pre-processing:

In this work, the EEG signals were split in segments of 5 s (2500 samples). All segments were filtered using a digital band-pass filter with cut-off frequencies of 2 and 40 Hz, in order to remove signal artifacts and the electrical noise. Then, the amplitude of each segment was normalized between 0 and 1 in order to improve the correlation computation and the algorithm performance.

Data processing:

The signal processing includes the following stages: (1) selection of pairs of channels based on their cross-correlation coefficients; (2) computation of unbiased cross-correlation functions magnitude between all segments of all selected channels pairs; (3) application of smoothing filter; (4) detection of maxima and minima values; (5) measurements of the time interval between the maximum and minimum values.

The goal of stage (1) is to decrease the algorithm running time and to find pairs of channels that exhibit some similarity. Because it is a neurodegenerative disease, PD causes the death of neurons and can maybe result in delays in the transmission of electrical activity over the brain [29]. Then, since the PD is at an early stage, the EEG signals are expected to be slightly delayed among each other and, therefore, have some similarity. Authors experiments have shown that a good choice for this degree of similarity between channels is that they have a cross-correlation coefficient in the range of 0.55 and 0.65. So, in this method stage, pairs of channels that present $0.55 \leq \rho_{xy} \leq 0.65$ are selected.

In stage (2), for each selected pair of channels, the unbiased cross-correlation functions between all their segments are calculated. It should be emphasized that unbiased cross-correlation functions are calculated using one segment of each channel, and not two segments of the same channel. Then, the magnitude of the cross-correlation functions is computed.

In stage (3), the magnitudes of the unbiased cross-correlation functions were smoothed using a Savitzky-Golay filter with order 50. This intends to improve the detection of maximum and minimum values of each cross-correlation function magnitude performed in stage (4). Finally, in stage (5), the time interval between the maximum and minimum values of each cross-correlation function magnitude is computed.

Pattern extraction, organization and selection:

In this work, the measured time intervals were used to perform global and channel analyses. The global analysis aims to detect early stages of PD while the channel analysis aims to find evidence of PD's activity in each electrode, that is, throughout the scalp. For both purposes, the measured time intervals were associated with the segments and channels used in calculating the corresponding cross-correlation function. Thus, each time interval was associated with two different channels and with one segment of each channel.

Figure 1. Methodology diagram
For the global analysis, the measured time intervals were organized per subject in a matrix where the rows represent channels and the columns represent the values of the time intervals associated with the channel. For the channel analysis, on the other hand, averages of the time intervals associated with each segment of each channel were performed. For the \( k \)-th segment of the \( l \)-th channel, one average was calculated using the time intervals measured from the cross-correlation function between this segment and the 1st segments of the channels that formed par with the \( l \)-th channel. Another average was calculated using the time intervals measured from the cross-correlation function between this segment and the 2nd segments of the channels that formed par with the \( l \)-th channel. And so on. These averages were organized per subject in a matrix where the rows represent channels and the columns represent their averages.

With the aiming of reducing the amount of the data to be used in the analyses, a genetic algorithm with entropy criterion and an X-ROC binary classifier were applied. In the global analysis, the genetic algorithm and the X-ROC classifier selected the combination of time intervals and the number of channels, respectively, that lead to the best differentiation between groups. As a result, 29-time intervals from 15 channels were selected per subject. In the segment analysis, on the other hand, the genetic algorithm selected the best combination of 100-time intervals that leads to the best differentiation between groups. The resulting data matrices are illustrated in Figure 2.

**Classification:**

For data classification, a feed-forward artificial neural network (fANN) with a hyperbolic tangent sigmoid transfer function was used. The Levenberg-Marquardt (LM) algorithm was used for training the fANN. The number of neurons within the hidden layer \((NCH)\) was given by \(\sqrt{NCI + NCO + 4}\), where \(NCI\) and \(NCO\) are the number of the input layer and output layer neurons, respectively [30].

In the global analysis, the matrix information for fANN training, testing and validation were split in ratios of 60%, 20%, 20%, respectively. Each phase had an input/output pair that ensured an unbiased generalization. As illustrated in Figure 3, in order to avoid overfitting caused by the spread of subject information over all fANN classification groups, different channels correspond to individual fANN input/output pairs.

**Figure 2.** Data matrices for global (a) and channel analysis (b).
In the channel analysis, the classification was performed for each data matrix, that is, for each channel. As each subject information was restricted to single line in the matrix, the aforementioned concern was not necessary and the subjects were randomly selected to meet the same chosen ratio for training, testing and validation, as illustrated in the Figure 4 for a single channel.

RESULTS

In the global analysis, the proposed method achieved an accuracy of 92.66%. For comparison purposes, the results of the state-of-the-art methods for PD detection in early stages are summarized in the Table 2.

Table 2: Comparison of the results between the proposed method and the state-of-the-art methods for the global analysis.

<table>
<thead>
<tr>
<th>Exam Type</th>
<th>Authors</th>
<th>Overall Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biomarkers</td>
<td>[25]</td>
<td>81.00</td>
</tr>
<tr>
<td>MEG</td>
<td>[16]</td>
<td>84.21</td>
</tr>
<tr>
<td>MRI</td>
<td>[24]</td>
<td>85.40</td>
</tr>
<tr>
<td></td>
<td>[23]</td>
<td>93.00</td>
</tr>
<tr>
<td></td>
<td>[18]</td>
<td>94.60</td>
</tr>
<tr>
<td>fMRI</td>
<td>[17]</td>
<td>85.00</td>
</tr>
<tr>
<td>EEG</td>
<td>[21]</td>
<td>82.00</td>
</tr>
<tr>
<td>Present Study</td>
<td>[22]</td>
<td>92.66</td>
</tr>
<tr>
<td>Multimodal Features</td>
<td>[19]</td>
<td>96.40</td>
</tr>
</tbody>
</table>

DISCUSSION

From Table 2 it can be noticed that the proposed method does not over-come the best results known until today. However, among the EEG-based methods, the proposed method stands out for achieving the best classification result, outperforming the state-of-art EEG-based method by 4.41%. This performance becomes even more relevant considering that EEG is a low-cost exam widely available in medical clinics. As intended, the channel analysis provided an overview of PD activity on the scalp, representing the discrimination accuracy of study groups per channel. The resulting topographic map is shown in Figure 5. As can be observed, the TP8 channel presented the best result, achieving an accuracy of 89.95% and outperforming the state-of-art method proposed in [21] by 7.95%. In addition, several other channels (F3, Pz, P3, P7, Oz, CP6, T8, FC6, F4, AF3, P1, PO7, PO8, FC4, and F6) outperformed the accuracy of 82% obtained in [21]. From Fig 5, it can be also concluded that the electrodes with higher accuracy level to differentiate the study groups can be found in right temporal region (T8, FT8, CP6 e TP8), frontal region (AF3, F3, F4, F6, FC6, FC4), parieto-occipital region (PO8, Pz, P1, P3, P7, PO7), and occipital region (Oz). These identified regions were already been reported in previous works as the main region of PD early activity. Using MEG activity, the study presented in [16] found differences that were statistically significant for several channels placed at parietal, occipital and temporal scalp regions. Through EEG activity, the authors in
[31] have used the wavelet packet entropy to confirm that the most significant differences appear outside of the central scalp region. And in [21], using the same EEG dataset as this work, the authors reported major differences in frontal and mid-frontal regions, as well as parietal regions.

CONCLUSIONS

Parkinson’s Disease (PD) is the second most common neurodegenerative disease worldwide and has no known cure. In recent years, researchers have been focusing their efforts to develop automatic methods for detecting PD in early stages in order to prevent as much as possible the disease symptoms, which unfortunately diminish patient's quality of living. For that reason, this work proposed a novel approach to characterize early stage PD activity over the EEG channels, which is a low-cost exam widely available in medical clinics. The proposed method selects pairs of channels that exhibit some similarity. Experiments have shown that a good choice is to select pairs of channels that have a cross-correlation coefficient in the range of 0.55 and 0.65. For each selected pair of channels, the magnitudes of unbiased cross-correlation functions between their segments are calculated and the time interval between the resulting maximum and minimum values are computed. The measure data was used for global and channel analyses. With the aim of reducing the amount of the data, a genetic algorithm and an X-ROC binary classifier were applied to select the combination of time intervals and the number of channels, respectively, in former. And the genetic algorithm was also used to select the best combination of time intervals in the latter. A feed-forward artificial neural network was used as classifier in both cases. With regard to the global analysis, the proposed method achieved an accuracy of 92.66% and outperforming the state-of-art EEG-based method by 7.95%. In addition, several other channels, namely F3, Pz, P3, P7, Oz, CP6, T8, FC6, F4, AF3, P1, P07, P08, FC4, and F6, also outperformed the state-of-art method.

ACKNOWLEDGMENTS

This work was supported by National Funds from FCT - Fundação para a Ciência e a Tecnologia through project UID/Multi/50016/2019.

REFERENCES


