Classification Models of Early Stage Parkinson’s Disease Based on Electroencephalography Feature Space

K. Obukhov[1], I. Maluta[1], Yu. Obukhov[2], O. Sushkova[2]

Moscow Institute of Physics and Technology, Moscow, Russia[1]
Kotel’nikov Institute of Radio-engineering and Electronics, RAS, Moscow, Russia[2]

ABSTRACT

The electroencephalographic (EEG) features of Parkinson’s disease (PD) are analyzed in the paper: presence of theta rhythm in low frequency range and disorder of the dominant alpha rhythm of brain activity. Based on these features and the data of 31 patients with clinical diagnosis of 1st stage non-treated PD and 18 control volunteers, the classification model was built. Logistic regression model was used for probability of PD estimation in each of 16 channels. It was shown, that weighted sum of probabilities among channels, where weights refer to classification accuracy AUC in each channel, is a function with high accuracy of classification in accordance with threshold. The model was tested on data of 22 PD patients and 16 normal volunteers. The accuracy of prediction was around 73%. The results of EEG signal analysis, as well as feature extraction techniques and model performance, prove that proposed approach can be applicable to Parkinson’s disease diagnostics on the early stage.

Keywords: Electroencephalography, EEG, Parkinson’s disease, logistic regression, wavelet transform, binary classification.

1. INTRODUCTION AND RELATED WORK

Parkinson's disease (PD) belongs to a wide class of neurodegenerative diseases and is caused by the death of dopaminergic neurons of the brain. Particular attention was paid to the mechanisms of brain plasticity, which serve to compensate functional insufficiency of the degenerating neurons. From this point of view, authors consider the dynamic of neurodegenerative diseases and state the necessity of preclinical diagnostics and preventive therapy development [1]. The main problem of PD diagnostics is the search of disease features at pre-clinical and early clinical stages.

Electroencephalography (EEG) and electromyography (EMG) are applicable methods for brain electrical activity analysis. The decrease of dominant frequency, as well as the power spectral density (PSD) shift, was previously found with the help of EEG and EMG spectral analysis [2]. Disorders of different organism systems, such as movement disorders, vegetative, emotional or physical, are considered as clinical features of PD. It is assumed that such disorders reflect or are coursed by brain electrical activity. This paper proposes an approach for feature extraction and classification of PD in EEG feature space [3]. This model can be applicable to disease risk group identification and screening. The train group consisted of 31 patients with clinical diagnosis of 1st stage non-treated PD and 18 control volunteers. Prediction results were tested on different set of 22 PD patients and 16 normal volunteers.

2. METHODS

Wavelet Morlet transform was used for time-frequency EEG features analysis of early stage PD. Special attention was paid to theta rhythm of EEG (4-6 Hz) and disorder of alpha rhythm (8-12 Hz). Continuous wavelet transform (2) with mother function Morlet (3) was used to process EEG signal x(t) into time-frequency-power density spectrogram (1):

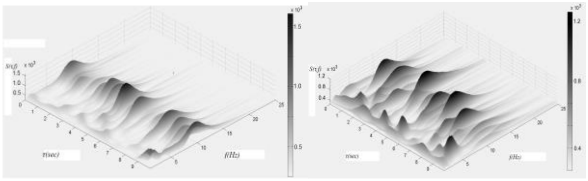
$$S_x = |W(\tau, f)|^2 \qquad (1)$$

$$W(\tau, f) = \frac{1}{\sqrt{f}} \int x(t)\psi*(\frac{t-\tau}{f})dt \qquad (2)$$

$$\psi(\eta) = \frac{1}{\sqrt{\pi F_b}} e^{2i\pi Fc\eta} e^{-\frac{\eta^2}{F_b}} \qquad (3)$$

where τ and f = 1/T are time and frequency of spectrogram, Fb = Fc = 1.

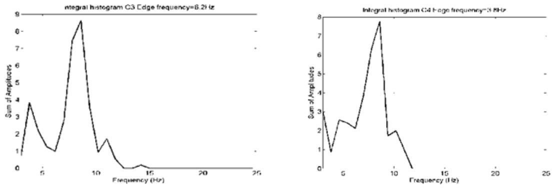
Fig. 1 illustrates the difference of S(τ) in the brain motor zone C3 (according to the standard 10x20 scheme of electrodes layout) for the normal volunteer (a) and (b) the patient at the first PD stage according the qualitative stages of PD described by Hoen-Yahr [4]. It can be seen that patient spectrogram shows high disorder, especially in the 8-12 Hz frequency range. Moreover, a theta rhythm arises in low frequency range. In order to analyze those features it is proposed to consider the time-frequency distribution of spectrogram extrema.



(a) (b)

Fig. 1 Time-frequency power density spectrogram of normal volunteer (a) and of the 1st stage PD patient (b) of the EEG signals in motor cortex zone C3.

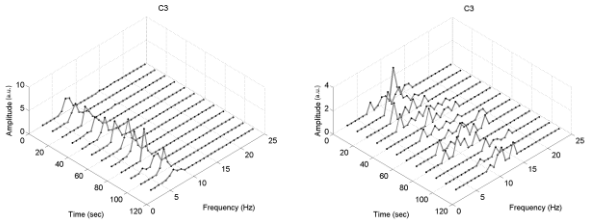
The distribution of extrema in time-frequency buckets (Δf, Δt) is used for further quantitative analysis. Fig. 2 frequency shows histograms asymmetry in 4-5 Hz frequency range of the 1st stage PD patient. The existence of theta rhythm can be found in C3, while there is an absence of such rhythm in C4 electrode. This one-sided development of disease is common on the early stages.



(a) (b)

Fig. 2 Sum of extrema histograms in (0.7 Hz, 180 sec) rectangles for symmetrical C3 (a), and C4 (b) EEG electrodes of 1st stage PD patient

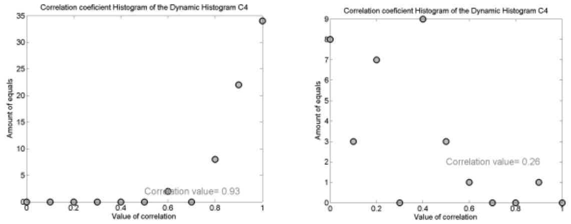
The relation of theta rhythm amplitude to alpha rhythm amplitude can be considered as the feature of PD. Moreover, extrema histograms can be used for quantitative analysis of the dominant alpha rhythm disorder. The dynamical histograms calculated in (0.7 Hz, 10 sec) rectangles show the disorder of electrical activity for the 1st stage PD (Fig. 3).



(a) (b)

Fig. 3. Dynamical histograms of the normal volunteer (a), and for the 1st stage PD patient. Histograms were calculated in (0.7 Hz, 10 sec) rectangles

Such disorder can be evaluated with the correlation matrixes. Thus, the average correlation for normal volunteer would be higher than for PD patient. This evaluation can be done by histograms of correlation values. Fig. 4 shows the difference of such histograms for the normal volunteer and the 2nd stage PD patient.



(a) (b)

Fig. 4 Histograms of correlation values for the normal volunteer (a), and the 2nd stage PD patient (b).

Average value of correlation coefficients as well as its standard deviation can be considered as the features of PD.
Finally, EEG feature space Pi was created:

$$P_i \in \left\{\frac{A_\theta}{A_\alpha}(j), \frac{A_\theta}{A_\alpha}(j^*), r(j), r(j^*), \sigma(j), \sigma(j^*)\right\} \quad (4)$$

Here Aθ/Aα (j) and Aθ/Aα (j*) refer to the ratio of theta rhythm to alpha rhythm in two opposite hemispheres. Indexes j and j* are 8 symmetrical EEG electrodes of standard 10x20 layout:
$j, j^* \in \{FP1FP2, F3F4, C3C4, T3T4, P3P4, T5T6, O1O2\}$.

By j we will refer to the hemisphere with PD and j* as the normal hemisphere. The average and deviation of correlation coefficients are marked as r(j), r(j*) and σ(j), σ(j*) respectively.

In order to aggregate the features into one value, which could indicate the probability of PD, it was proposed to build a logistic regression model. The probability of a particular outcome is linked to the linear prediction function:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i} \quad (5)$$

In this function p_i is the probability of positive outcome for observation i, given $x_{m,i}$ – the feature m in a dataset. The weight of each feature can be computed by maximizing the likelihood function:

$$L = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \quad (6)$$

Receiver Operating Characteristics (ROC) curve can be used for the model efficiency analysis. ROC curve shows the dependence between Sensitivity – True Positive Rate, and Specificity – True Negatives Rate. The Area Under Curve (AUC) can show the quality of the classification model. For random prediction AUC equals to 0.5, while for perfect prediction AUC is 1.

3. CLINICAL RESULTS

Logistic regression model was trained on the EEG data of 31 patients with clinical diagnosis of 1st stage non-treated PD and 18 control volunteers. EEG signals were measured in 8 pairs of electrodes. The table below indicates the AUC results of the models for each pair of electrodes based on the feature space (4) for PD hemisphere and normal hemisphere:

Table 1. AUC values for logistic regression model in 8 pairs of electrodes.

	FP1FP2	F3F4	C3C4
PD hemisphere	0.82	0.79	0.73
Normal hemisphere	0.79	0.67	0.69

T3T4	F7F8	P3P4	T5T6	O1O2
0.69	0.79	0.60	0.68	0.56
0.63	0.60	0.62	0.62	0.68

It can be seen, that the best prediction accuracy is reached in FP1 and FP2 electrodes, while the worst – in O1O2 and P3P4. Moreover, prediction accuracy on the PD hemisphere is slightly higher than on Normal hemisphere. Further the model trained on PD hemisphere was used. In total, 8 models were trained, and the outcome of the models was probability of PD.

The model analysis was also done on the test data of 22 PD patients and 16 normal volunteers. EEG data of this group was not used in model training. Each model predicted probability of PD in every electrode. In order to compose all predictions in one value, it is necessary to develop an aggregation function based on each probability. It was proposed to use function F, which can be calculated according to formula:

$$F = \sum_{j,j^*} |P(j, j^*) - 0.5| * AUC(j, j^*) \quad (7)$$

This function summarizes all probabilities adjusted on 50%. This adjustment is needed, so the contribution is zero for not certain electrodes with 50% probability of PD. Then these adjusted probabilities are weighted with AUC of a particular electrode from Table 1. Thus models with lower accuracy would not have the same weight as good models. In addition, electrodes O1O2 and P3P4 were excluded from this aggregation due to low AUC.

Fig. 5 shows the distribution of F among patients with PD (stage = 1) and normal volunteers (stage = 0).

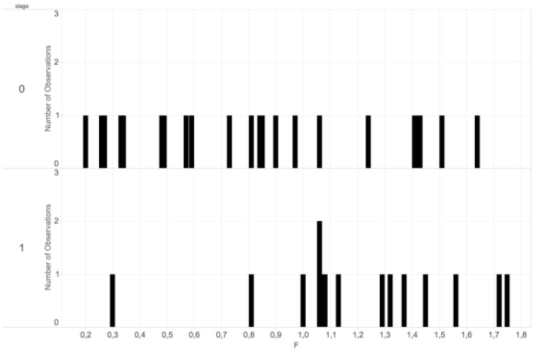


Fig. 5 Distribution of aggregated function F for control (above) and PD patients (below).

Patients with PD have higher values of F, which are mostly concentrated around 1. Although, normal volunteers have low values of F, but with high deviation. In order to make a certain prediction, whether person has PD or not, it is necessary to select a cut off value. In accordance with this cut off value, person would be classified as PD or not. For sure, cut off value should have the optimal accuracy of classification among PD patients and among control volunteers. It is proposed to calculate recall of PD (Recall 1) and control (Recall 0) respectively:

$$Recall_0 = \frac{TP}{TP+FN} \quad (8)$$

$$Recall_1 = \frac{TN}{TN+FP} \quad (9)$$

In the formulas above, TP indicates the number of True Positives (correct prediction of PD), TN refers to True Negative (correct prediction of control), FP refers to False Positives (false prediction of PD) and FN refers to False Negatives (false prediction of control). For each possible cut off value, these recall functions are calculated.

The results are shown on Fig. 6. In case of low threshold, model would assume everyone to have PD, thus recall of PD would be 100%, but recall of control is 0%. On the other hand, if threshold is too high, then model predicts everyone as control, but none as PD. The optimal threshold is around 1.1. In this case, both recalls would be near 73%. Overall accuracy is also present on the chart in circles.

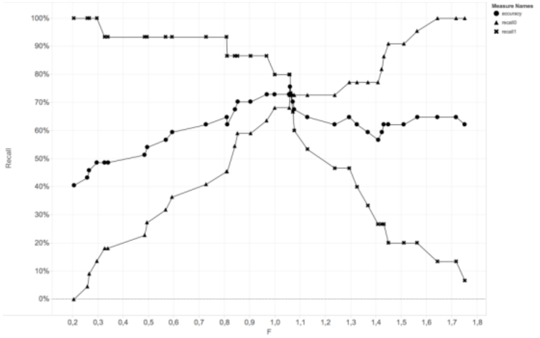


Fig. 6 Recall curve of PD (cross) and control (triangle). Accuracy curve (circles) for various cut off thresholds.

5. CONCLUSIONS

The electroencephalographic (EEG) features of Parkinson’s disease (PD) were analyzed in the paper. It was shown, that low frequency theta rhythm of brain activity as well as disorder of dominant rhythm can be considered as the features of early stage PD. The feature space was created and logistic regression model was built with binary target of disease. The training data consisted of 31 patients with clinical diagnosis of 1st stage non-treated PD and 18 control volunteers. The model was trained for each of 16 channels, and the accuracy AUC was measured in each channel respectively. The probabilities of disease among channels were summarized in the function in accordance with AUC. The resulted function was proven to have high accuracy of prediction on the test data of 22 PD patients and 16 normal volunteers. The accuracy of prediction was around 73%. The future scope of work covers extraction of other potentially important features of PD from EEG data, testing other classification models and aggregate functions in order to increase the accuracy of prediction.

The results of EEG signal analysis, as well as feature extraction techniques and model results, prove that proposed approach can be applicable to Parkinson’s disease diagnostics on the early stage, where the disease is already developed, but the clinical symptoms have not yet appeared

6. ACKNOWLEDGEMENT

The research was supported by the Russian Foundation for Basic Research, project #15-07-07846.

7. REFERENCES

[1] Ugrumov M. V., Khaindrava V. G., Kozina E. A. et al. “Modeling of preclinical and clinical stages of Parkinson’s disease in mice”, **Neuroscience**, 2011. V. 181.

[2] H.W. Berendse, C.J. Stam, “Stage-dependent patterns of disturbed neural synchrony in Parkinson’s disease”. **Parkinsonism and Related Disorders**, 2007. v. 13, Suppl. 3. p. 440-445.

[3] Yu.V. Obukhov, I.A. Maliuta, K.Yu. Obukhov, “Classification of early stage Parkinson’ disease in electroencephalography features space”, **Pattern Recognition and Image Analysis**, 2016, V. 26 Number 4.

[4] M.M. Hoehn, M.D. Yahr, “Parkinsonism: onset, progression and mortality”, **Neurology**, 1967, V. 17, pp. 427-442, PMID 6067254.