Budget-based classification of Parkinson's disease from resting state EEG

Ilkka Suuronen, Antti Airola, Tapio Pahikkala, Mika Murtojärvi, Valtteri Kaasinen, Henry Railo

***Abstract*— Early detection is vital for future neuroprotective treatments of Parkinson's disease (PD). Resting state electroencephalographic (EEG) recording has shown potential as a cost-effective means to aid in detection of neurological disorders such as PD. In this study, we investigated how the number and placement of electrodes affects classifying PD patients and healthy controls using machine learning based on EEG sample entropy. We used a custom budget-based search algorithm for selecting optimized sets of channels for classification, and iterated over variable channel budgets to investigate changes in classification performance. Our data consisted of 60-channel EEG collected at three different recording sites, each of which included observations collected both eyes open (total N = 178) and eyes closed (total N = 131). Our results with the data recorded eyes open demonstrated reasonable classification performance (ACC = .76; AUC = .76) with only 5 channels placed far away from each other, the selected regions including right-frontal, left-temporal and midline-occipital sites. Comparison to randomly selected subsets of channels indicated improved classifier performance only with relatively small channel-budgets. The results with the data recorded eyes closed demonstrated consistently worse classification performance (when compared to eyes open data), and classifier performance improved more steadily as a function of number of channels. In summary, our results suggest that a small subset of electrodes of an EEG recording can suffice for detecting PD with a classification performance on par with a full set of electrodes. Furthermore our results demonstrate that separately collected EEG data sets can be used for pooled machine learning based PD detection with reasonable classification performance.**

***Index Terms*— Parkinson's disease, electroencephalography, sample entropy, group-wise feature selection, classification**

I. Introduction

Parkinson's disease (PD) is a progressive degenerative neurological disorder whose cardinal motor signs are rest tremor, slowness of movement (bradykinesia), and muscular rigidity. The characteristic neuropathological feature of PD is the formation of abnormal alpha-synuclein aggregations, and the degeneration of dopaminergic cells in the midbrain region substantia nigra pars compacta. Although multiple biomarkers of PD have been proposed [1], the diagnosis of PD is primarily based on motor symptoms. This is problematic because by the time the motor symptoms emerge, the neuropathological changes of PD are already widespread [2], and a large proportion of dopaminergic cells has been lost [3]. An effective neuroprotection for PD will require earlier disease detection, and procedures that aid in the identification of PD from early biomedical signals are therefore needed. One of the proposed approaches is to classify PD based on electroencephalographic (EEG) recordings. EEG measures electrophysiological activity of large neuronal populations from electrodes attached to the scalp. It is low-cost, easy to measure, and used as a standard clinical assessment procedure in hospitals.

PD is associated with changes to the EEG spectrum [4]–[8], and researchers have tested whether PD can be classified based on features extracted form the EEG. Studies which have classified PD patients and neurologically healthy age-matched controls based on linear features calculated from different frequency bands have yielded approximately 75–82% accuracy [9], [10]. Building on the assumption that instead of reflecting linear stochastic processes, EEG is characterized by non-linear dynamics [11], studies suggest that non-linear methods may provide valuable additional information for classification [12]–[14]. In studies employing non-linear measures for classification, accuracy has been around 80–95% [12], [13], [15]. For instance, Lainscsek et al. [12] showed that nine patients and nine age-matched controls could be classified almost perfectly with merely one second of EEG when delay differential equations were used. Liu et al. [13] studied 17 PD patients and 25 healthy controls using resting eyes-closed EEG, and reported over 90% classification accuracy based on discrete wavelet transformation and sample entropy. Following the promising results obtained in prior investigations [13], [15] using signal complexity based features, we based our feature extraction procedure on a complexity measure and a related algorithm known as sample entropy.

Studies often report which channels or channel clusters yield highest classifier performance (e.g., [8], [10], [12], [16]), but previous studies have not systematically ex-

This work was supported in part by the Academy of Finland (A.A, grant 340140 and 345805, T.P., grants 313266, 311273, 340182 and 345804 and H.R, grant 308533)

I. Suuronen is with the Department of Computing at University of Turku, 20500, Turku, Finland (e-mail: ilksuu@utu.fi)

A. Airola is with the Department of Computing at University of Turku, 20500, Turku, Finland (e-mail: ajairo@utu.fi)

T. Pahikkala is with the Department of Computing at University of Turku, 20500, Turku, Finland (e-mail: aatapa@utu.fi)

M. Murtojärvi is with the Department of Computing at University of Turku, 20500, Turku, Finland (e-mail: mianmu@utu.fi)

V. Kaasinen is with the Turku University Hospital, 20521 Turku, and Department of Clinical Neurosciences, University of Turku, 20500, Turku, Finland (e-mail: valtteri.kaasinen@utu.fi)

H. Railo is with the Department of Psychology and Speech-Language Pathology at University of Turku, 20500, Turku, Finland (e-mail: henry.railo@utu.fi)

amined how classification performance is affected by the number of EEG channels. From the practical point-of-view, it would be a considerable advantage if the classification of PD from EEG could be performed with minimal number of channels. Firstly, this would make the EEG set-up fast. Secondly, high-density EEG montages (e.g., 64-channel EEG) used in research are rarely used in hospitals. Previous research suggests that small number of channels is sufficient for classification. For example, Lainscsek et al. [12] used clusters of 2–6 local neighbouring channels in their analysis, and Liu et al. used 10-channel EEG with electrodes placed evenly across the head. However, systematic analysis of how classification performance varies as the function of number of channels is lacking. Although few channels may enable adequate classification, it could be that additional channels further improve the performance of the classifier.

A related question is which electrode locations yield the highest classification performance. On the one hand, recording montages that cover the whole head could work best because they allow sampling of wide-ranging sources of neural activity. On the other hand, pathological features of EEG could predominantly influence activity in specific EEG channels (e.g., frontal electrode locations over the motor cortex). If certain scalp locations provide more information for classification of PD than others, this information could be used to optimize EEG recording montages for PD classification. Finally, due to the high correlation of EEG signals across electrodes, it could be that the selection of specific, "optimal" electrode locations does not significantly improve classifier performance.

Given that the number of different channel combinations can be very large (e.g., there are $1.51 * 10^{11}$ different 10-channel combinations with 64 electrodes), finding optimal combinations, and determining how the number of selected channels influences classification is difficult. To examine how classification performance changes as a function of number of EEG channels, we implemented a novel channel-wise feature selection method which first fits a model to the features extracted from the single best channel, and then augments the model further with features from additional performance-maximizing channels, until a preset budget for channels in the model is depleted. By iterating over variable budgets, we were able to observe at which number of channels the increase in classification performance becomes negligible, and which channels provide most value for classification. To examine how much additional information the budget-based search algorithm provides, we compare its performance to randomly selected channel subsets.

While previous studies report high classification accuracy, the studies have typically relied on a relatively small sample, and all of the data has been collected at one laboratory (except for [8]). Given that small sample sizes are known to bias classifier performance [17], [18], we combined data from three different laboratories to reach a reasonable sample size. Combining data collected at different laboratories is also arguably a more realistic and likely more difficult situation for classification (e.g., because of systematic differences in EEG recording equipment, procedure and environment). The budget-based classification procedure was run separately for eyes-closed and eyes-open resting state EEG data, which are both standard clinical EEG protocols [19]. Most previous studies which classify PD based on resting state EEG have used eyes closed recordings [10], [13], [15], [20], [21].

II. Method

A. EEG data

The analysis was based on pooled resting state EEG data collected at University of Turku [22], University of New Mexico, and University of Iowa (see, [8]). Data collected at each site were approved by the local Institutional Review Board. Each data set contained data collected both with the subjects' eyes open (henceforth EO) and closed (henceforth EC). The Turku and New Mexico datasets contain both EO and EC data from each subject (except that one patient is missing the EC in the Turku data). In the Iowa data, the EC condition has less participants than the EO condition. The total number of PD patients is 89 (EO condition) and 68 (EC condition), and 89 (EO condition) and 63 (EC condition) control participants. Because four PD patients and two control participants had an irregular number of channels, they were excluded from further analysis.

Figure 1 shows the power spectrum of the EEG data in different conditions across all three datasets. The comparison of EEG power spectral density between PD and control groups displays the often reported pattern that PD patients display stronger EEG activity around 5–10 Hz. Demographical, clinical and neuropsychological data describing these datasets are presented in Tables I-IV.

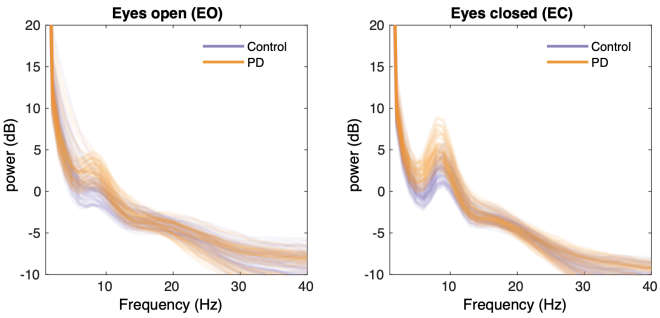


Fig. 1: Comparison of EEG power spectrum in PD (orange lines) and control (blue lines) groups in the eyes open (left panel) and eyes closed (right panel) conditions. Lines represent different channels (average across participants/patients).

In the Turku dataset, the patients were asked to go through a voluntary medication break before the test session to counteract possible intervening effects of medication. Thirteen patients underwent the medication break, and did not take their morning medication during the day of the study (at least 12 h break in medication). These

Turku (Eyes open and closed)	Control	PD
N	20	20
Sex (male/female)	8/12	9/11
Age, years	67.8 (6.2)	69.8 (7.2)
levodopa eq. daily dose, mg	n/a	663.2 (509.1)
Disease duration, years	n/a	6.4 (4.9)
Mini-mental state score	28.2 (1.5)	27.8 (1.8)
Beck's depression inventory	5.0 (3.0)	8.4 (6.2)
Montreal cognitive assessment	n/a	n/a
Geriatric depression scale	n/a	n/a
MDS-UPDRS, motor	5.1 (3.5)	28.9 (16.4)

TABLE I: Demographic, clinical and neuropsychological data. Turku dataset. Reported as mean (SD).

New Mexico (Eyes open and closed)	Control	PD
N	28	28
Sex (male/female)	18/10	18/10
Age, years	69.2 (9.0)	69.7 (8.4)
levodopa eq. daily dose, mg	n/a	703.6 (432.7)
Disease duration, years	n/a	5.5 (4.1)
Mini-mental state score	n/a	28.6 (1.0)
Beck's depression inventory	n/a	7.6 (5.1)
Montreal cognitive assessment	n/a	n/a
Geriatric depression scale	n/a	n/a
MDS-UPDRS, motor	n/a	24.8 (8.6)

TABLE II: Demographic, clinical and neuropsychological data. New Mexico dataset. Reported as mean (SD).

patients can therefore be considered being in OFF phase. In the New Mexico dataset, 14 patients were OFF medication (the New Mexico data we analyzed here contain each participant's first recording session). All patients were on medication in the Iowa dataset.

In all datasets, EEG was originally sampled at 500 Hz frequency using 64 channels. All data recording protocols were approved by the local Ethics Committee and followed the Declaration of Helsinki. During data procerssing, the EEG was resampled to 250 Hz, and then run through an automated PREP preprocessing pipeline [23]. During preprocessing, 50 or 60 Hz line noise was removed, bad channels were interpolated, and EEG was referenced to robust average reference. We used PREP preprocessing pipeline because it follows standardized processing steps, and doesn't include time consuming computational steps such as independent component analysis. Because the three datasets were collected using different electrode arrangements, we removed channels that were not included in all three datasets. This left us with 60 channel EEG data. EEG preprocessing was performed in Matlab 2016b

Iowa (Eyes open)	Control	PD
N	41	41
Sex (male/female)	23/18	28/13
Age, years	71.3 (7.4)	66.1 (8.0)
levodopa eq. daily dose, mg	n/a	887.6 (414.0)
Disease duration, years	n/a	5.4 (3.9)
Mini-mental state score	n/a	n/a
Beck's depression inventory	n/a	n/a
Montreal cognitive assessment	26.6 (1.7)	25.8 (3.2)
Geriatric depression scale	0.7 (1.0)	3.3 (2.8)
MDS-UPDRS, motor	n/a	13.6 (7.11)

TABLE III: Demographic, clinical and neuropsychological data. Iowa eyes open dataset. Reported as mean (SD).

Iowa (Eyes closed)	Control	PD
N	15	20
Sex (male/female)	7/8	12/8
Age, years	70.8 (5.1)	67.2 (7.8)
levodopa eq. daily dose, mg	n/a	947.2 (526.0)
Disease duration, years	n/a	5.5 (3.0)
Mini-mental state score	n/a	n/a
Beck's depression inventory	n/a	n/a
Montreal cognitive assessment	26.4 (1.6)	23.5 (3.6)
Geriatric depression scale	1.1 (1.5)	4.4 (3.8)
MDS-UPDRS, motor	n/a	14.8 (4.6)

TABLE IV: Demographic, clinical and neuropsychological data. Iowa eyes closed dataset. Reported as mean (SD).

using EEGlab plugin [24].

B. Outline of the analysis

The primary objective of this analysis was to investigate the performance of a machine learning model in correctly classifying subjects as PD patients or healthy control participants when using a limited selection of channels of an EEG recording. To this end, we employed a custom greedy budget-based and grouped feature selection algorithm [25] to include or exclude groups of features into the model in such a manner that all features in a given group pertain to the same channel. Because of the practical interchangeability between the two concepts, and for the sake of simplicity, such groups of features pertaining to the same channel will be henceforth referred to as "channels". Performing the feature selection procedure in such a channel-wise manner allowed us to observe the effect of gradually introducing channel-specific information into the model on the classification performance, ultimately yielding insight into a favourable number and placement of electrodes for recording EEG. Further, we performed the same analyses for a simple baseline method based on random selection of channels.

Preprocessing, as described in the previous chapter, was followed by a feature extraction procedure consisting of separating a number of distinct wave-form components from each channel-specific signal using band-pass filtering, and computing a measure of signal complexity called sample entropy [26] for each such component. The resulting feature vectors were then used for training and evaluating the test performance of a logistic regression classifier model using nested 10-fold cross-validation to perform the budget-based grouped feature selection procedure.

K-fold cross-validation is a conventional method for evaluating a machine learning model's generalization ability to previously unseen data by partitioning the training data into k distinct subsets called *folds*. Each of the folds is in turn used for testing, while the remaining $k-1$ folds are pooled and used for training the model. This procedure results in k fold-specific test performance estimates, from which a total test performance estimate can be computed by averaging. Nested k-fold cross-validation involves a further partitioning of the training data for the purpose of model selection which can involve hyperparameter tuning or feature selection. The rationale for employing a

nested cross-validation algorithm instead of a regular, "flat" one when performing model selection is that the latter procedure has been reported to result in a highly biased performance estimate whereas using a nested cross-validation reduces this bias, resulting in a more realistic estimate of the model's generalization capability. [27]

By using the "inner" loop of the nested cross-validation procedure to perform budget-based grouped feature selection, we were able to observe the feature selection process on the level of the EEG channels, as constricted by the budget variable. Furthermore, by gradually increasing the budget, we were able to observe the effect of further augmenting the model with additional channels with respect to the classifier performance. The primary metrics used for reporting the classifier performance were area under the receiver operating characteristic curve (henceforth AUC), which can be interpreted as the probability that the model ranks a random "positive" case higher than a random "negative" case[1], and classification accuracy defined as the ratio of correctly classified cases to both correctly and incorrectly classified cases. In order to determine the statistical significance of the model's performance, we performed permutation tests [29] on both datasets at full budget and observed p-values of < 0.0001 for both performance metrics. In addition to the aforementioned metrics, sensitivity, specificity, positive predictive values and negative predictive values were reported for certain budgets. The analysis was performed using Python 3 with NumPy [30], Scikit-learn [31], Scipy [32], Neurokit2 [33] and Xarray [34] as external libraries.

C. Feature extraction

Feature extraction procedure consisted of applying a sixth order forward-backward Butterworth band-pass filter to each of the 60 channel-specific signals of the EEG recording for each subject to separate the conventional delta, theta, low alpha, high alpha and beta components. The respective low and high cutoff frequencies were defined as $[0.5, 4.0]$, $[4.0, 8.0]$, $[8.0, 10.0]$, $[10.0, 13.0]$ and $[13.0, 30.0]$ in Hz. As the final step of feature extraction stage, sample entropy was computed for each such signal component resulting in total 300 numeric features[2] per subject. As there was considerable variance in lengths of the EEG recordings between sites, only the first $15,000$ data points corresponding with the first 30 seconds of the signal were used for feature extraction.

Sample entropy (henceforth SampEn) is a measure of complexity used for quantifying the unpredictability of a physiological time series, based on estimating the probability that two matching subsequences of a given length m are still similar at length m+1. It is closely related to another complexity measure known as approximate entropy (ApEn), but designed to eliminate a bias toward regularity that originates from ApEn comparing each subsequence with itself.

SampEn is defined by a series of N data points $X = x_1, x_2, ..., x_N$, an embedding dimension m and a tolerance r. The embedding dimension, also known as order, is used for constructing a set of template vectors $u_m(i) = [x_i, x_{i+1}, ...x_{i+m-1}]$ for each i such that $1 \leq i \leq N - m$ in the embedding space $\mathbf{R}^m$. Another such set $u_{m+1}(i)$ of template vectors is constructed in the embedding space $\mathbf{R}^{m+1}$. Also, a distance function $d(u(i), u(j))$ is defined for template vectors $u(i), u(j)$ as the Chebyshev distance (or some other distance metric) in their respective embedding spaces.

Given a time series $x(n)$, a distance metric d, an embedding dimension m and a tolerance r, SampEn can be defined as:

$$SampEn = log\frac{B}{A} \tag{1}$$

where

$$A = \sum_{i=1}^{N-m} \sum_{j=1, j\neq i}^{N-m} \theta(r - d(u_{m+1}(i), u_{m+1}(j))) \tag{2}$$

$$B = \sum_{i=1}^{N-m} \sum_{j=1, j\neq i}^{N-m} \theta(r - d(u_{m}(i), u_{m}(j))) \tag{3}$$

and where θ is the Heaviside step function.

In the present analysis, SampEn was computed with embedding dimension of 2 and tolerance of $0.2 * \sigma$, where σ is the standard deviation of the sample.

D. Budget-based grouped feature selection

To find locally optimal[3] subsets of channels for training the model with, we employed a custom budget-based and grouped feature search algorithm. Given a budget B, the algorithm selects in total B channels to include in the model. New channels are selected in a greedy fashion with respect to the validation performance, so that each new channel selected by the algorithm is the one the inclusion of which in the model has the highest positive effect (or the least negative effect) on the validation performance. The selected channel is then removed from the set of available channels, the cost of the model is increased and the selection procedure continues until the cost is equal to the budget, at which point the selection procedure terminates. After the feature selection procedure is completed, a final model is trained with the selected channels using the full training data (i.e. the $k - 1$ folds) and its performance is evaluated against the test fold. Validation AUC was used as the primary optimization criterion in the feature selection procedure. As a simple and computationally efficient baseline, we consider also a channel selection method that simply selects the channels

[1]For an introduction to ROC-curves and computing AUC, see e.g. Fawcett 2006 [28]

[2]60 channels * 5 frequency band specific features per channel.

[3]By "locally optimal" subset of channels we refer to such a subset of channels that has been built up by optimal individual choices at each stage of the greedy search procedure.

in a random order. For pseudocode representation of the greedy channel selection procedure, see Algorithm 1. To assess the effectivity of the feature selection algorithm, we compare it with a random grouped feature selection algorithm. For a given budget B, the random grouped feature selection algorithm iteratively selects B channels at random instead of based on validation performance. This procedure was repeated 100 times, and an average performance was computed for each budget in order to establish a baseline performance.

Algorithm 1: Budget-based greedy channel selection

Input : C = Classifier, X = Training data, Y = Target variable, B = Budget, A = Set of available channels

Output: Locally optimal model M given budget B

Initialize the set for selected channels: $S \leftarrow \emptyset$
Initialize model cost variable: $c \leftarrow 1$
while $c \leq B$ **do**
  $s \leftarrow argmax_{a \in A}(Score(C(X_{S \cup a}, Y))$
  $S \leftarrow S \cup s$
  $A \leftarrow A \setminus \{s\}$
  $c \leftarrow c + 1$
end
$M \leftarrow C(X_S, Y)$

III. Results

The development of main performance metrics (accuracy, and AUC) as functions of budget can be inspected in Figure 2. The results demonstrate a difference between the datasets with respect to the performance, the EO dataset generally outperforming the EC dataset. The EO dataset has the minimum values for both metrics at a budget of one channel (accuracy: 0.66, AUC: 0.71) and the maximum values for both metrics at a budget of five channels (accuracy: 0.76, AUC: 0.76). The EC dataset likewise has the minimum values for both metrics at a budget of one channel (accuracy: 0.52, AUC: 0.69), a maximum value for AUC at a budget of 59 channels (0.78) and a maximum value for classification accuracy at a budget of 56 channels (0.72). For selected budgets, average receiver operating characteristic curves are depicted in Figure 3, in addition to which average specificity, sensitivity, positive predictive value and negative predictive value are displayed in table V.[4] Most immediately notable about these metrics is the imbalance between the high average sensitivity and the low average specificity in the case of the EC data set, especially at low budgets, while the metrics are relatively well balanced in the case of the better performing EO dataset.

As shown in Figure 2, in the case of the EO data, the greedy search outperforms the random selection at low to medium budgets. This is in line with our prior results with greedy selection under a budget (e.g. [25]). Namely, the first steps of greedy selection tend to be good, because there is still a lot of room for improvement. However, after the channels providing the best improvement have already been selected, the remaining channels can not be distinguished from each other anymore due to noise, and the greedy selection starts to resemble the random one. Indeed, at medium to high budgets, greedy selection performs like random with bad luck, since it is even outperformed by the random search averaged over several runs. This is even more clear with the EC data, on which the greedy search only sporadically reaches the average performance of the random selection.

[4]Missing negative predictive values for certain budgets are caused by lack of negative predictions in one or more folds.

Budget	Sensitivity	Specificity	Pos. pred. value	Neg. pred. value
EO 5	0.80	0.71	0.74	0.81
EC 5	0.91	0.26	0.58	n/a
EO 10	0.75	0.69	0.70	0.77
EC 10	0.92	0.33	0.61	n/a
EO 20	0.75	0.69	0.70	0.77
EC 20	0.85	0.46	0.65	0.76
EO 30	0.73	0.67	0.67	0.75
EC 30	0.82	0.52	0.67	0.76
EO 60	0.75	0.70	0.70	0.77
EC 60	0.77	0.60	0.69	0.73

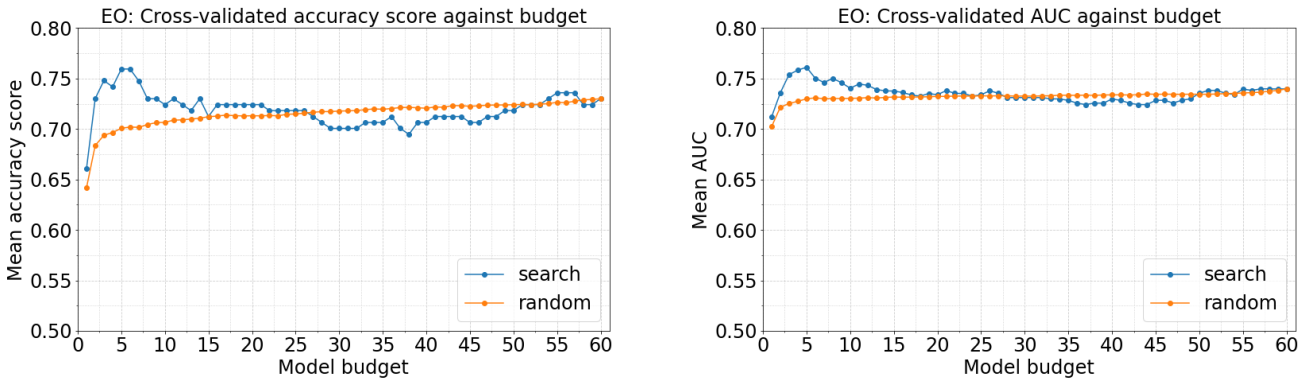
TABLE V: Sensitivity, specificity, positive predictive values and negative predictive values for budgets of 5, 10, 20, 30 and 60 for both EO and EC datasets.

The above result suggests that optimized channel combinations improve classification when the classification is performed with limited number of electrodes (in the EO condition). Figure 4 shows, for six different sized channel budgets, the frequency a given channel was selected into the model. With EO data and a five channel budget, the search algorithm selected channels located in right-frontal, left-temporal and midline-occipital sites. With higher number of channel budgets, frontal channels over the motor cortex produced highest increase in classifier performance in the EO condition. In contrast, selected locations in the EC condition appear more random, with some concentration on the right parietal sites, consistent with the observation that the search algorithm was outperformed by average random feature selection.

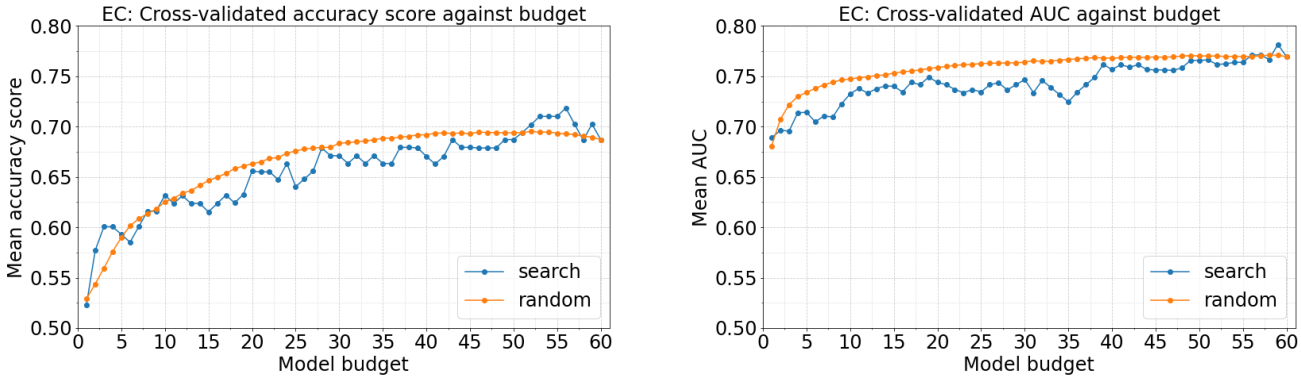
IV. Discussion

Our results indicate that a pooled data set of EEG recordings, comprising of data collected at three different laboratories, can be classified reasonably well (AUC $>$ 0.7) using complexity-based features and a linear classifier based on a small number (5) of channels. This is consistent with previous studies showing high accuracy classification with 1–10 electrodes ([8], [12], [13]). Our results further show that when classification is performed with eyes open (EO) resting state EEG data, classifier performance can be optimized by selecting features form a limited number of different channels. In contrast, when classification was based on eyes closed (EC) data, random feature selection yielded better performance, and classifier performance

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3235040



(a)



(b)

Fig. 2: A plot depicting classification accuracy and AUC as a function of budget for the EO (a) and EC (b) datasets. Greedy search algorithm and random selection.

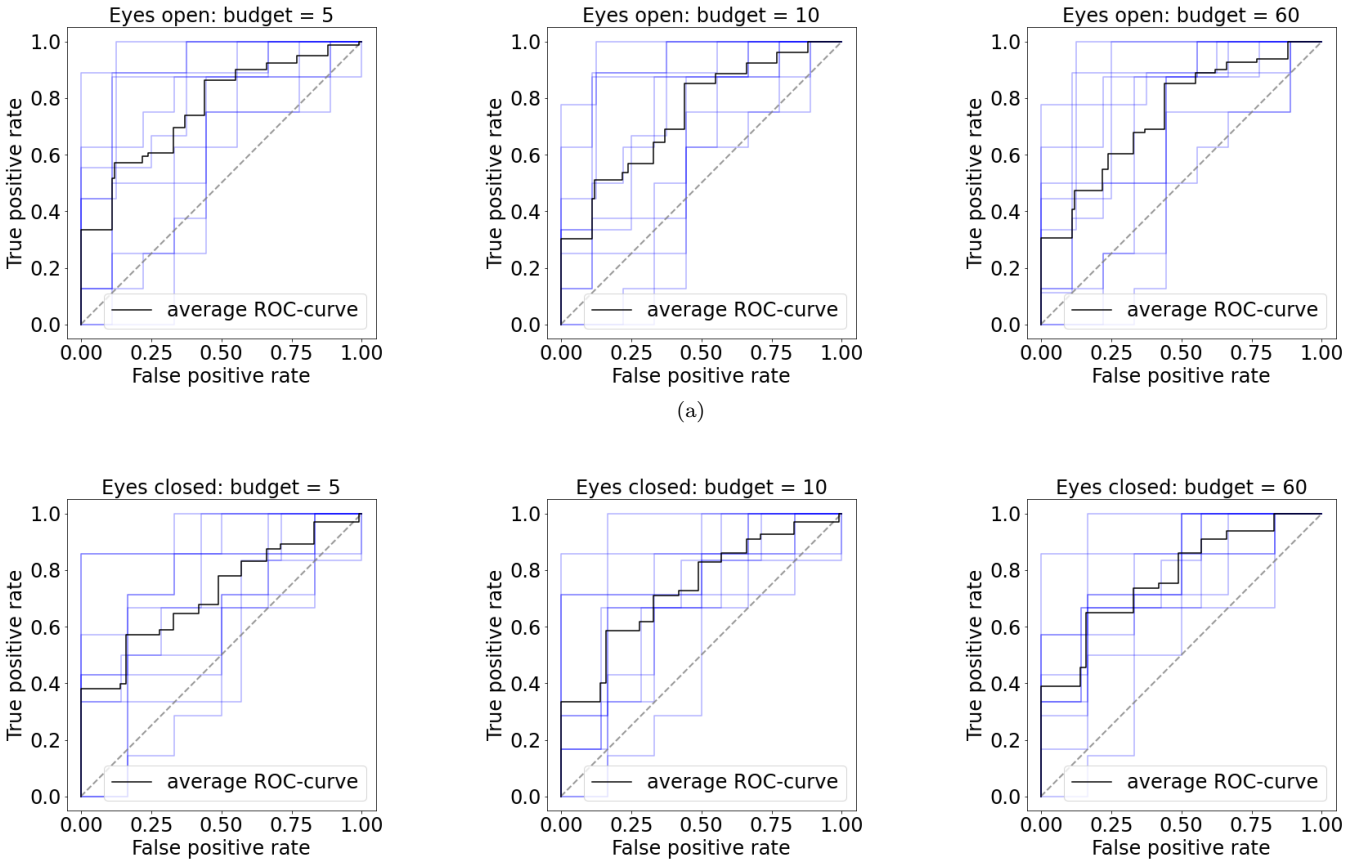
improved as a function of channels (at least with these features, and a linear classifier).

Our budget-based search algorithm indicated that with EO data, the best classification performance was obtained with only five channels. With the five channel budget, channels that were placed far away from each other on scalp were selected into the classifier, possibly because this enabled sampling from somewhat independent EEG activity sources. With larger budgets, the search algorithm most often selected channels over the motor cortex. This suggests that EEG activity in the motor cortical areas may have contributed to the classifier performance. This result suggests that (at least when entropy-based features and eyes-open resting state data), EEG recording montages where electrodes are placed on these sites may yield best classification of PD from healthy controls.

The performance of our classifier is modest when compared to previously reported EEG complexity-based classifiers [13], [15]. Previous studies are typically based on clearly smaller number of patients, and the data have been collected in a single site. Because small sample sizes are likely to lead to biased classifiers, the present results suggest, in contrast to previous research, that EEG complexity-based classifiers alone are not sufficient for accurate classification. It should be noted, however, that our findings might not generalize well to other feature generalization or classification algorithms than those as reported in this paper. Features reflecting coherence between different EEG channels [35], or methods that characterize the holistic shape of the EEG power spectrum [8] or stimulus-evoked activation [9] may further improve the classifier.

Our results suggest that classification may work better based on eyes open than eyes closed resting state EEG data. This may indicate that sample entropy features or the classification algorithm used in the present study was ill-suited for eyes closed data. Because there are substantial differences in EEG dynamics between eyes closed and open conditions [15], [16], [36] — modulating, for instance, motor-related neural activity [37]–[39] — different features might be needed to accurately classify PD based on eyes closed EEG. Previously, Liu et al. [13] showed that PD could be accurately predicted from eyes closed resting state EEG data when sample entropy was used for feature extraction. While this is at odds with the present results, direct comparison of the studies is difficult. Sample size was smaller in the Liu et al. [13] study, and for example, the article do not report how severe PD the patients in their sample had. If the sample consisted of PD patients with e.g. severe rest tremor, high classification

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3235040



(a)

(b)

Fig. 3: Cross-validated average receiver operating characteristic curve for EO (a) and EC (b) data sets with varying budgets. Fold-specific receiver operating characteristic curves displayed in blue.

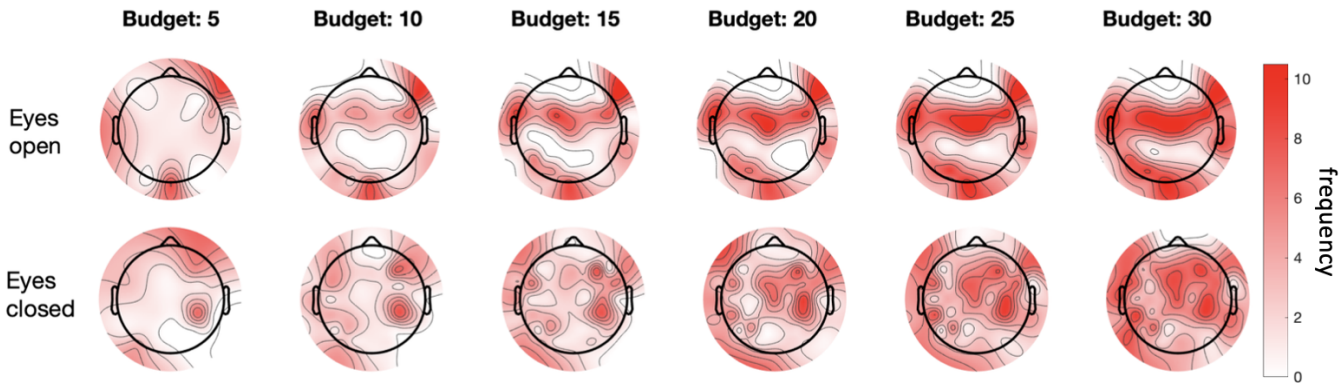


Fig. 4: Heatmaps depicting the number of times specific channels were selected into a model between the 10 CV-iterations for different budgets.

accuracy is relatively trivial as movement affects EEG. That said, many previous papers have reported relatively good classification performance with eyes closed resting state data [10], [13], [15], [20], [21]. It should also be taken into account that the smaller sample size associated with the EC data set may have contributed to the difference between the respective classification performance estimates of the two data sets.

Our analysis is based on early or intermediate PD patients. If classifiers are to help in the diagnosis in future, naturally the classifiers should work with early stage PD patients. Because our patients were all on medication, we cannot rule out confounds related to medication. Possibly classifiers focusing on unmedicated, or strictly on patients in ON or OFF-phase could work better. Gomez et al. [15] report that early non-medicated PD patients could be clas-

sified from healthy controls with high accuracy based on eyes closed resting state magnetoencephalography (MEG) with Lempel-Ziv complexity.

V. Acknowledgements

We thank James Cavanagh and Nandakumar Narayanan laboratories for sharing their data.

References

[1] N. Titova, M. A. Qamar, and K. R. Chaudhuri, "Biomarkers of Parkinson's Disease: An Introduction," in *Parkinson's Disease*, ser. International Review of Neurobiology, K. P. Bhatia, K. R. Chaudhuri, and M. Stamelou, Eds., 2017, vol. 132, pp. 183–196.

[2] C. Gaig and E. Tolosa, "When does Parkinson's disease begin?" *Movement Disorders*, vol. 24, no. S2, pp. S656–S664, 2009.

[3] J. M. Fearnley and A. J. Lees, "Ageing and parkinson's disease: Substantia nigra regional selectivity," *Brain*, vol. 114, no. 5, pp. 2283–2301, 1991.

[4] J. L. Bosboom, D. Stoffers, C. J. Stam, B. W. van Dijk, J. Verbunt, H. W. Berendse, and E. C. Wolters, "Resting state oscillatory brain dynamics in Parkinson's disease: An MEG study," *Clinical Neurophysiology*, vol. 25, no. 4, pp. 187–193, 2006.

[5] M. Y. Neufeld, R. Inzelberg, and A. D. Korczyn, "EEG in demented and non-demented parkinsonian patients," *Acta Neurologica Scandinavica*, vol. 78, no. 1, pp. 1–5, 1988.

[6] R. Soikkeli, J. Partanen, H. Soininen, A. Pääkkönen, and P. Riekkinen, "Slowing of EEG in Parkinson's disease," *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 3, pp. 159–165, 1991.

[7] H. Tanaka, T. Koenig, R. D. Pascual-Marqui, K. Hirata, K. Kochi, and D. Lehmann, "Event-related potential and EEG Parkinson's disease without and with dementia," *Dementia and Geriatric Cognitive Disorders*, vol. 11, no. 1, pp. 39–45, 2000.

[8] M. F. Anjum, S. Dasgupta, R. Mudumbai, A. Singh, J. F. Cavanagh, and N. S. Narayanan, "Linear predictive coding distinguishes spectral eeg features of parkinson's disease," *Parkinsonism & Related Disorders*, vol. 79, pp. 79–85, 2020.

[9] J. F. Cavanagh, P. Kumar, A. A. Mueller, S. P. Richardson, and A. Mueen, "Diminished EEG habituation to novel events effectively classifies Parkinson's patients," *Clinical Neurophysiology*, vol. 129, no. 2, pp. 409–418, 2018.

[10] M. Chaturvedi, F. Hatz, U. Gschwandtner, J. G. Bogaarts, A. Meyer, P. Fuhr, and V. Roth, "Quantitative EEG (QEEG) measures differentiate Parkinson's disease (PD) patients from healthy controls (HC)," *Frontiers in Aging Neuroscience*, vol. 9, p. 3, 2017.

[11] T. M. McKenna, T. A. McMullen, and M. F. Shlesinger, "The brain as a dynamic physical system," *Neuroscience*, vol. 60, no. 3, pp. 587–605, 1994.

[12] C. Lainscsek, M. E. Hernandez, J. Weyhenmeyer, T. J. Sejnowski, and H. Poizner, "Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals," *Frontiers in Neurology*, vol. 4, p. 200, 2013.

[13] G. Liu, Y. Zhang, Z. Hu, X. Du, W. Wu, C. Xu, X. Wang, and S. Li, "Complexity Analysis of Electroencephalogram Dynamics in Patients with Parkinson's Disease," *Parkinson's Disease*, vol. 2017, 2017.

[14] C. J. Stam, B. Jelles, H. A. M. Achtereekte, S. A. R. B. Rombouts, J. P. J. Slaets, and R. W. M. Keunen, "Investigation of EEG non-linearity in dementia and Parkinson's disease," *Electroencephalography and Clinical Neurophysiology*, vol. 95, no. 5, pp. 309–317, 1995.

[15] C. Gómez, K. T. Olde Dubbelink, C. J. Stam, D. Abásolo, H. W. Berendse, and R. Hornero, "Complexity analysis of resting-state MEG activity in early-stage Parkinson's disease patients," *Annals of Biomedical Engineering*, vol. 39, no. 12, p. 2935, 2011.

[16] J. Gómez-Ramírez, S. Freedman, D. Mateos, J. L. Pérez Velázquez, and T. Valiante, "Exploring the alpha desynchronization hypothesis in resting state networks with intracranial electroencephalography and wiring cost estimates," *Scientific Reports*, 2017.

[17] A. H. Neuhaus and F. C. Popescu, "Sample size, model robustness, and classification accuracy in diagnostic multivariate neuroimaging analyses," *Biological psychiatry*, vol. 84, no. 11, pp. e81–e82, 2018.

[18] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of Neuroscience Methods*, vol. 250, pp. 126–136, 2015.

[19] S. R. Sinha, L. R. Sullivan, D. Sabau, D. S. J. Orta, K. E. Dombrowski, J. J. Halford, A. J. Hani, F. W. Drislane, and M. M. Stecker, "American clinical neurophysiology society guideline 1: minimum technical requirements for performing clinical electroencephalography," *The Neurodiagnostic Journal*, vol. 56, no. 4, pp. 235–244, 2016.

[20] S. Vanneste, J.-J. Song, and D. De Ridder, "Thalamocortical dysrhythmia detected by machine learning," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[21] R. Yuvaraj, U. R. Acharya, and Y. Hagiwara, "A novel parkinson's disease diagnosis index using higher-order spectra features in eeg signals," *Neural Computing and Applications*, vol. 30, no. 4, pp. 1225–1235, 2018.

[22] H. Railo, N. Nokelainen, S. Savolainen, and V. Kaasinen, "Deficits in monitoring self-produced speech in Parkinson's disease," *Clinical Neurophysiology*, vol. 131, no. 9, pp. 2140–2147, 2020.

[23] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K. M. Su, and K. A. Robbins, "The PREP pipeline: Standardized preprocessing for large-scale EEG analysis," *Frontiers in Neuroinformatics*, vol. 9, p. 16, 2015.

[24] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[25] M. Murtojärvi, A. S. Halkola, A. Airola, T. D. Laajala, T. Mirtti, T. Aittokallio, and T. Pahikkala, "Cost-effective survival prediction for patients with advanced prostate cancer using clinical trial and real-world hospital registry datasets," *International Journal of Medical Informatics*, 2020.

[26] J. S. Richman, D. E. Lake, and J. R. Moorman, "Sample entropy," in *Numerical Computer Methods, Part E*, ser. Methods in Enzymology. Academic Press, 2004, vol. 384, pp. 172–184.

[27] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, 2006.

[28] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[29] P. Golland, F. Liang, S. Mukherjee, and D. Panchenko, "Permutation tests for classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005.

[30] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: A structure for efficient numerical computation," *Computing in Science and Engineering*, 2011.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.

[32] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[33] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1–8, 2021.

[34] S. Hoyer and J. Hamman, "xarray: N-D labeled arrays and datasets in Python," *In revision, J. Open Res. Software*, 2017.

[35] P. Silberstein, A. Pogosyan, A. A. Kühn, G. Hotton, S. Tisch, A. Kupsch, P. Dowsey-Limousin, M. I. Hariz, and P. Brown,

"Cortico-cortical coupling in parkinson's disease and its modulation by therapy," *Brain*, vol. 128, no. 6, pp. 1277–1291, 2005.

[36] R. J. Barry, A. R. Clarke, S. J. Johnstone, C. A. Magee, and J. A. Rushby, "EEG differences between eyes-closed and eyes-open resting conditions," *Clinical Neurophysiology*, 2007.

[37] K. H. Chen and Y. Z. Huang, "The change of motor cortical excitability between eyes open and closed conditions," *NeuroReport*, 2018.

[38] S. Rimbert, R. Al-Chwa, M. Zaepffel, and L. Bougrain, "Electroencephalographic modulations during an open-or closed-eyes motor task," *PeerJ*, 2018.

[39] J. Wei, T. Chen, C. Li, G. Liu, J. Qiu, and D. Wei, "Eyes-open and eyes-closed resting states with opposite brain activity in sensorimotor and occipital regions: Multidimensional evidences from machine learning perspective," *Frontiers in Human Neuroscience*, 2018.