Unveiling Parkinson's Disease Features from a Primate Model with Deep Neural Networks

1st Caetano M. Ranieri
ICMC, *USP*
São Carlos, SP, Brazil
cmranieri@usp.br

2nd Renan C. Moioli
IMD, *UFRN*; and
Santos Dumont Institute (ISD)
Natal, RN, Brazil
renan.moioli@imd.ufrn.br

3th Roseli A. F. Romero
ICMC, *USP*
São Carlos, SP, Brazil
rafrance@icmc.usp.br

4th Mariana F. P. de Araújo
CCS, *UFES*
Vitória, ES, Brazil
Santos Dumont Institute (ISD)
Natal, RN, Brazil
mfparaujo@gmail.com

5th Maxwell Barbosa De Santana
ICTA, *UFOPA*
Santarém, PA, Brazil
barbosadesantana@gmail.com

6th Jhielson M. Pimentel
Edinburgh Centre for Robotics, *HWU*
Edinburgh, Scotland, UK
jm210@hw.ac.uk

7rd Patricia A. Vargas
Edinburgh Centre for Robotics, *HWU*
Edinburgh, Scotland, UK
p.a.vargas@hw.ac.uk

***Abstract*—Parkinson's Disease (PD) is a neurodegenerative disorder with increasing prevalence in the world population and is Characterised by motor and cognitive symptoms. Although cortical EEG readings from PD-affected humans have being commonly used to feed different machine learning frameworks, the directly affected areas are concentrated in a group of subcortical nuclei and related areas, the so-called motor loop. As those areas may only be directly accessed through invasive procedures, such as Local Field Potential (LFP) measurements, most data collection must rely on animal models. To the best of our knowledge, no neural networks-based analysis centred on LFP data from the motor loop was reported so far. In this work, we trained and evaluated a set of deep neural networks on a dataset recorded from marmoset monkeys, with LFP readings from healthy and parkinsonian subjects. We analysed each trained neural network with respect to its inputs and representations from intermediate layers. CNN and ConvLSTM classifiers were applied, reaching accuracies up to 99.80%, as well as a CNN-based autoencoder, which has also shown to learn PD-related representations. The results and analysis provided further insights and foster research on the correlates of Parkinson's Disease.**

***Index Terms*—Parkinson's disease, LFP analysis, deep learning, attribution methods, computational neuroscience.**

I. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder and estimates indicate a prevalence between 1 and 2 per 1,000 individuals. Age is the most relevant factor to influence such incidence [1]. The most common symptoms are motor deficits, such as bradykinesia, rigidity, and resting tremors, although cognitive symptoms, especially dementia, may occur in later stages [2]. PD diagnosis is clinical (there is no feasible biomarker), and current treatments provide symptomatic relief, but do not stop, revert, or slow disease progression [3]. In this context, machine learning techniques are being used to characterise the neurophysiological correlates of PD, which can contribute to unveil disease mechanisms as well as non-trivial features that are present on neural data. Ultimately, this may facilitate early diagnoses and support novel therapies.

With few exceptions, human datasets are comprised of non-invasive electroencephalography (EEG) recordings, which capture the neural dynamics from cortex superficial layers. However, the neural circuits directly associated with PD may only be sampled by invasive electrodes, limiting the availability of human studies. Nevertheless, there are consolidated animal models of PD, in which disease symptoms can be elicited by administering neurotoxins [4]. From implanted electrodes, Local Field Potential (LFP) signals are obtained and these have a close relationship with EEG signals [5]. Furthermore, the basic anatomy and structure of the neural circuitry relevant to PD are conserved across most vertebrate species [6], thus supporting the use of such animal models.

In this paper, we designed a set of deep neural networks able to learn explainable features from raw time-domain LFP data with minimum preprocessing. The trained models were evaluated for their ability to classify structured data segments as belonging to healthy or PD animal subjects. Then, we highlighted which properties of the segments contributed most to the networks' classifications. To accomplish that, we used a marmoset monkey database [7] of simultaneous LFP recordings from PD-related brain regions, namely the basal ganglia-thalamus-cortex (BG-T-C) system, known as the *motor loop*). The network architectures include a fully-connected (FC) network, used as a baseline method, a Convolutional Neural Network (CNN), and an hybrid CNN with Long Short-Term Memory (ConvLSTM). We also employ an autoencoder-based unsupervised framework to analyse not only the consistency

This work was funded by the Neuro4PD project-Royal Society and Newton Fund (NAF\R2\180773), and São Paulo Research Foundation (FAPESP), grants 2017/02377-5 and 2018/25902-0. Moioli, Araujo, and Santana acknowledge the support from the brazillian institutions: INCT INCEMAQ of the CNPq/MCTI, FAPERN, CAPES, FINEP, and MEC. This research was carried out using the computational resources from the CeMEAI funded by FAPESP, grant 2013/07375-0. Additional resources were provided by the Robotics Lab within the ECR, and by the Nvidia Grants program.

of the lower-dimension representation (i.e., *embedding*) with respect to the two conditions - healthy or PD - but also the suitability of the learnt features in regular classification.

Our deep learning models trained on LFP readings reached accuracy above 99% and the learnt features resemble those that were previously associated to PD. As the acquisition of LFP data depends on inserting electrodes directly inside the subject's brain, the methods proposed are not directly suitable for diagnosis, but rather to provide an additional framework for better understanding of the underlying mechanisms of the disease. To the best of our knowledge, this paper is the first attempt to apply deep neural networks to better understand PD features extracted from simultaneous multi-region LFP.

II. Related Work

Recent research has shown that deep neural networks may be promising machine learning algorithms for studying biological systems. In [8], classification of Alzheimer's disease was performed based on magnetic resonance brain images fed into a CNN. In [9], the intention to perform certain movements was detected from EEG signals, by generating time-frequency maps through wavelet transforms and feeding them to a CNN.

Regarding PD, in [10], EEG data was collected from 15 patients in early stages of PD, ranging from Hoehn and Yahr (H&Y) [11] stages 1 to 2, and 15 healthy subjects with a similar age profile. They applied the autoregressive Burg and the Wavelet Packet Entropy (WPE) methods to characterise the resulting signals in terms of frequency bands and to identify cortical patterns that may be indicative of PD at its early stages, and found significant differences between the affected patients and the control subjects.

Yuvaraj, Acharya, and Hagiwara [12] provided a machine learning (ML) framework to diagnose the disease in an EEG dataset produced by 20 affected patients with H&Y stage ranging from 1 to 3, though most of them were in stages 2 or 3, and a control group of 20 other subjects with no history of mental illness. They extracted the Higher-Order Spectra (HOS), a well-established technique for feature extraction from biomedical data, and introduced a feature ranking method before applying several classical classifiers, obtaining a state-of-the-art diagnosis with Support Vector Machines (SVM).

An important development in the Brain-Computer Interfaces (BCI) domain was the EEGNet [13], based on the application of a compact CNN in diverse motor tasks. Besides providing a classification framework, the authors explored the interpretation of features by analysing filter outputs, convolutional kernel weights, and single-trial relevance. The SyncNet [14] was another CNN-based deep network capable of handling not only EEG, but also LFP signals from public datasets. When processing EEG data, the framework generated visualisations of the spatial patterns recognised by the network filters with heat maps representing learnt amplitude and phase in different bands of the frequency spectrum. Their work did not describe a visualisation approach for features learnt from LFP data.

In [15], different CNN architectures were employed in order to classify public EEG datasets focused on commands for initiating movement. For visualisation, two types of correlation maps were considered: input-feature unit-output, consisted of bandpass-filtering the input signal to each frequency band of interest and checking the outputs of each unit of the network, and input-perturbation network-prediction, based on perturbations on the network inputs. A paradigm derived from research on video classification was proposed in [16], in which the spatially-coherent readings of the EEG electrodes at a given timestep were represented as a regular 2D image, and stacks of such images were interpreted as sequential frames in a video. The data was collected during a working memory experiment, and different architectures were considered for feature extraction and classification, especially neural networks.

Regarding research aimed at PD diagnosis, [17] presented a thirteen-layer CNN that was applied directly to EEG data from 20 PD patients and 20 healthy subjects from similar age groups for classification. The accuracy obtained was lower than that reported in related work with handcrafted features, though direct comparisons are difficult to make due to the lack of standardised datasets. A different technique was presented on [18], which applied Echo State Networks (ESN) to classify data collected from patients with REM-sleep Behaviour Disorder (RBD), a risk factor for PD, and healthy controls, with promising results. Both papers focused on classification, with few considerations regarding the representations learnt.

Research on learning feature representations from brain signals through unsupervised techniques presented two autoencoder architectures to learn short-time features from EEG data from a public dataset [19]. Each trial was represented as a 2D image whose pixel intensities were related to the power of different EEG frequency bands at the spatial location of each particular electrode in the scalp surface, and channel-wise, in which each EEG electrode was treated as a different channel. The embeddings learnt were fed to fully-connected layers to perform classification tasks, leading to state-of-the-art results in the cross-subject experiments. Another autoencoder-based framework was proposed by Wen and Zhang [20], designed to learn representations related to epilepsy with the so-called AE-CDNN model.

The above-mentioned literature focused on learning representations based on EEG signals from humans by applying different sorts of neural networks, with accurate results in comparison to other approaches. PD-related work was also relied on this modality of data, however work on this subject did not provide an in-depth analysis on the interpratability of the features learnt. Also, LFP data has not been a focus of ML efforts in understanding PD, though we have found research addressing this modality for other purposes. This paper attempts to fulfil those gaps by providing a comparative study via a set of deep networks that learned from a PD-related dataset of marmosets' LFP measurements, in both supervised and unsupervised manners.

III. The Motor Loop

The motor loop of the mammals' brain is formed by the *motor cortex* (M1), the *thalamus* (TH), and the *basal ganglia* (BG), the latter composed of a subset of structures: the *striatum*, which itself includes the *putamen* (PUT) and the *caudate nucleus*, the *globus pallidus*, divided into *pars interna* (GPi) and *pars externa* (GPe), the *subthalamic nucleus* (STN), and the *substantia nigra*, divided into *pars compacta* (SNc) and *pars reticulata* (SNr). McGregor and Nelson [21] provided a discussion about the mechanisms of this loop and presented models to describe it. The most useful model to explain the connections affected by PD is the so-called classic model, illustrated in Fig. 1, which highlights the relationships between the projections of neurons from the SNc to the BG structures, mainly striatum, where dopamine is released.

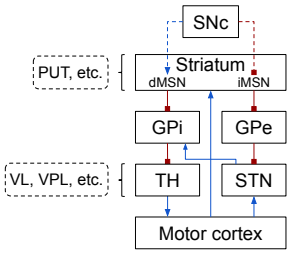


Fig. 1: Excitatory (blue) and inhibitory (red) connections from the circuitry of the motor loop. PD is caused by the loss of neurons of the *substantia nigra pars compacta* (SNc), which weakens the connections represented by the dashed lines. This causes malfunction on both the direct and indirect pathways.

The pathways begin with an excitatory connection from the cortex to the striatum, which projects its output neurons, named *medium spiny neurons* (MSN), to other structures inside the BG. In the direct pathway, the direct MSN (dMSN) inhibits the GPi, which reduces its inhibition to the TH, which then excites the motor cortex. In the indirect pathway, the indirect MSN (iMSN) inhibits the GPe, which reduces its inhibition to the STN, which excites the GPi. Thus resulting on inhibition of the TH and absence of excitatory outputs to the motor cortex. Hence, in summary, the direct pathway excites the cortex (i.e., positive feedback loop), while the indirect pathway inhibits it (i.e., negative feedback loop). PD is characterised by the progressive loss of dopaminergic neurons, especially in the SNc, which causes malfunctions to both pathways.

IV. Methods

This work consists of applying deep neural networks to LFP data collected from marmoset monkeys. We have considered networks for classification, trained to distinguish between healthy and PD-induced individuals, and autoencoders, trained in an unsupervised manner. To explore the representations learnt by each model, we applied attribution methods to segment the input sequences and to look for the most relevant features. All implementations were developed using the TensorFlow/Keras framework. The experiments were performed on a desktop equipped with an Intel Core i7-7700 CPU and a NVidia Titan-V GPU.

A. Datasets

Four adult males and one adult female common marmosets (i.e., *Callithrix jacchus*), weighing 300–550 g, were used in the study performed on [7]. The animals were housed in pairs in a vivarium with a natural light cycle (12/12 hr) and outdoor temperature. All animal procedures followed approved ethics committee protocols (CEUA-AASDAP 08/2011, 11/2011, 02/2015, and 03/2015) strictly in accordance with the NIH Guide for the Care and Use of Laboratory Animals. PD symptoms were elicited in all four male animals with injections of 6-OHDA toxin under deep anesthesia. LFPs were sampled at 1000 Hz and recorded using a 64 multi-channel recording system (Plexon) with fully-awaken animals behaving freely. Electrode coordinates and dopaminergic lesions were verified in all animals.

B. Data Preprocessing

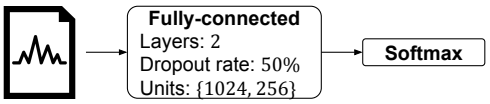
The only healthy individual had recordings from the M1, PUT, GPe, and GPi regions, thus we limited our analysis to those regions. In total, 14 and 16 recording sessions were obtained for the healthy and for the PD conditions, respectively, considering hemispheres independent from each other. Each recording session was segmented in 2-second data segments. As multiple electrodes were recorded for each region, a preprocessing pipeline was required before providing a standardised data structure, as with other approaches in the literature [22]. For each channel, our pipeline began with a low-pass filter (cutoff frequency of 250 Hz), a high-pass filter (cutoff frequency of 0.5 Hz) and a hum notch filter at 60 Hz, 120 Hz, and 180 Hz frequencies. Each signal was then scaled according to a z-score normalisation. The next step was to compute the cross-correlation matrix of each region and discard channels with mean correlation coefficient below the threshold of 0.7. Finally, all channels within a brain region were averaged, which provided a matrix with dimensions 4×2000.

After that, to reduce the amount of noisy or non-meaningful data, we imposed additional criteria to decide whether to keep or discard each resulting instance. An upper threshold of 0.2 was set for the module of the mean of the signal over time at each region, and a lower threshold of 0.1, for the standard deviation. Also, each window was required to show a minimum of 10 peaks.

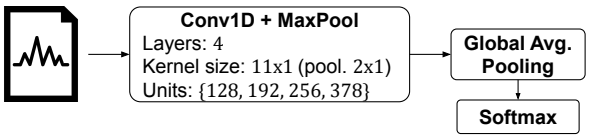
C. Network Architectures

We considered the classification task of distinguishing between healthy and PD-induced individuals and elaborating embedding representations through an autoencoder [23], which could be analysed on its own or coupled with supervised techniques to check its ability to enhance the classification procedure. The different architectures considered are illustrated in Fig. 2. The number of layers and its numbers of neurons were chosen based on literature on EEG classification and

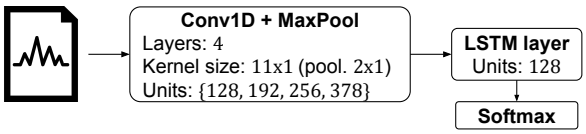
exploratory experiments. The complexity of each model is shown in Fig. 3.



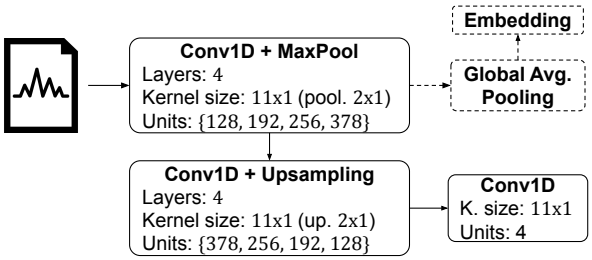
(a) Fully-connected architecture. The two fully-connected layers, both provided with dropout, are followed directly by the softmax layer.



(b) CNN architecture. All convolutional layers within a *Conv1d + MaxPool* block were set to the same kernel size, with one convolutional layer being interspersed with a max-pooling layer. At the top, the features map are processed through global average pooling and fed to a softmax layer.



(c) ConvLSTM architecture. The *Conv1D + MaxPool* block is similar to that of the CNN architecture, however its output is fed to a LSTM layer, whose output is fed to the softmax layer.



(d) Autoencoder architecture. The convolution/upsampling block and the top convolutional layer with the output with the same dimension as the input signal, used for training the autoencoder, is removed and replaced by a global average polling for providing the embeddings.

Fig. 2: Network architectures. The input signals are processed by the intermediate layers, which would be a stack of fully-connected, convolutional, pooling or upsampling, depending on the architecture (icon by Freepic, from www.flaticon.com).

1) Classification Networks: Three different architectures were considered for classification. All of them were endowed with a readout layer made of two neurons, each related to one of the two possible classifications - healthy or PD - and softmax activation function. The baseline, Fig. 2a, was a shallow, fully-connected (FC) neural network, consisted of two intermediate layers with dropout set to 50%. We also considered a 4-layered CNN, Fig. 2b, with the convolutional layers composed of 1-dimensional filters with receptive field

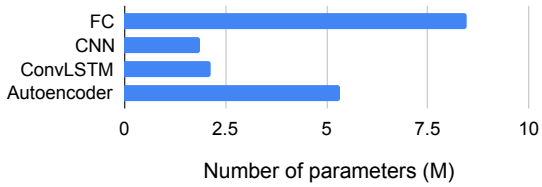


Fig. 3: Complexity of each model, given by the number of parameters, in millions.

of size 11 and interspersed with max-pooling layers with filter size 2, and a ConvLSTM, Fig. 2c, inspired by literature on activity recognition from inertial sensors [24], which consisted of the CNN architecture provided with an additional LSTM layer at the top, right before the softmax layer. The number of units at each layer is depicted in Fig. IV-C1.

2) Autoencoder: The autoencoder, Fig. 2d, reproduced the CNN architecture and endowed it with a reconstruction block. At training time, four convolutional-upsampling pairs were introduced to reverse the encoding produced, followed by a convolutional readout layer to reconstruct the input shape. At test time, the 378×125 encoding following the last max-pooling layer would be processed by global average pooling and turned into a flat feature vector composed of 378 units, which we call *embedding*.

This embedding was employed in two other classification settings. The first consisted of simply feeding the embedding to a fully-connected network, just like the one illustrated in Fig. 2a, and training the fully-connected network regardless of the original CNN that generated the embedding. The second consisted of inserting a softmax layer at the top of the global average pooling layer, resulting in an architecture identical to that of the CNN depicted in Fig. 2b, and fine-tuning all its weights.

D. Attribution Methods

Algorithms to assign a value to the contribution of each input to a given output of a neural network may be called *attribution methods*. A comprehensive summary of different methods was presented on [25]. Formally, given an input $X = [x_1, \ldots, x_N] \in \mathbb{R}^N$ and an output $S(X) = [S_1(X), \ldots, S_C(X)]$, where N is the number of input neurons and C is the number of output neurons, the problem consists of assigning an attribution $R^C = [R_1^C, \ldots, R_N^C] \in \mathbb{R}^N$ of each input feature $x \in X$ with respect to a given output $S_k(X) \in S(X)$. The *Integrated Gradients* method [26], adopted in this work, is based on the gradients obtained through a single backward pass through the network.

Here, we applied the DeepExplain framework [25] to the outputs for computing the attributions of each instance with respect to the input signals and to all intermediate layers. The outcome, in the case of the inputs, may be represented by the example in Fig. 4, which presents the attribution of each timestep of the input channels as colour maps. It is worth mentioning that negative attributions, represented in blue, are

also present, and might be interpreted as evidence *against* the output analysed.

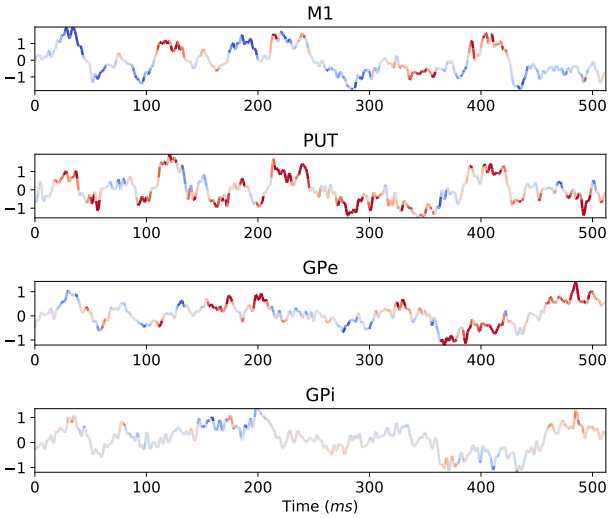


Fig. 4: Example of attributions at the input layer with respect to a given output. Contributions of each timestep are represented in a colour map, with red points corresponding to positive attributions, and the blue points, to negative attributions. In other words, red (blue) points relate to increased (decreased) probability of correct classification.

V. Results and Analysis

After preprocessing, 14 sessions from the healthy condition and 11 sessions from the PD condition were kept. Based on that, the data was split following the rule that segments that belonged to the same recording session would always belong to the same fold. This policy allowed the data to be split into 11 folds, each consisting of one healthy and one PD recording session, preventing the ML algorithms from achieving high accuracy by simply learning session-specific artifacts. The classification networks were trained to optimise the softmax cross-entropy loss, while the autoencoder was trained to optimise the mean squared error (MSE). All neural networks were trained using stochastic gradient descent (SGD), with learning rate 10^{-2} and decay 10^{-4}, for 30 epochs. We have chosen to apply SGD without momentum because this was the most stable training algorithm, generally leading to convergence on both train and test sets. The number of epochs was actually overestimated, since convergence appeared to happen around epoch 10, though it was kept for safety, since the loss remained stable after achieving an optimal set of parameters. The trained models were evaluated with regular evaluation metrics, but they were also analysed with respect to its features, as presented in the next subsections.

A. Performance Evaluation

The classification results are presented in Table I, including the shallow FC network applied to the autoencoder embedding and the pre-trained CNN, which is actually a fine-tuned autoencoder. Accuracy and macro F1-score (i.e., the harmonic mean between macro precision and recall) were close for all models, which suggest an equilibrium between true and false classifications across both classes. The results show that the CNN performed expressively better than the baseline FC network, with a 5.98% accuracy rise from 93.65% to 99.63% and standard deviation an order of magnitude lower. The ConvLSTM presented a perceptible improvement towards the CNN: the error rate, the opposite of accuracy, dropped from 0.37% to 0.20%, with even lower standard deviation. The FC applied to the autoencoder's embedding presented a slight improvement when compared to the baseline FC, despite an increase at the standard deviation, especially regarding the F1-score, which may suggest worse performance at certain circumstances. Pre-training the CNN had little effect on the classification metrics, as the accuracy of the CNN and the pre-trained CNN changed only 0.02%.

TABLE I: Classification metrics for each network architecture, with window size $t = 2,000$ points. Means and standard deviations between all folds.

	Accuracy (%)	F1-score (%)
Fully-connected	93.65 ± 6.03	93.39 ± 6.14
CNN	99.63 ± 0.78	99.61 ± 0.83
ConvLSTM	99.80 ± 0.40	99.79 ± 0.45
AE / FC	95.76 ± 7.93	94.49 ± 10.57
Pre-trained CNN	99.65 ± 0.68	99.63 ± 0.75

The dataset in which we performed the experiments is not public, thus there are no related work to which we can directly compare these results. Also, PD-related LFP data is not readily available for most research on the issue, even considering data from rodents. If compared to EEG datasets, collected under more controlled circumstances, our results would be consistent with the state-of-the-art, in which accuracy of up to 99.62% can be found with HOS features and SVM-RBF classifier [12]. Regarding deep neural networks, the CNN of [17] hit an accuracy of 88.25%, while [18] reported an accuracy around 85% with ESN classifiers.

The autoencoder's embedding went through an additional performance evaluation. Three clustering methods were applied to the feature vector - K-means, agglomerative hierarchical clustering and DBSCAN - and the clusters were evaluated according to entropy-based evaluation metrics [27], which take into account the labels of the instances assigned to each cluster. Those metrics were the Homogeneity of the clusters, according to which each cluster contains only instances of a single class, the completeness, according to which all instances of a given class are assigned to the same cluster, and the V-measure, the harmonic mean between the other two. The results, shown in Table II, give a measurement of whether the features learnt by the autoencoder and grouped by the clustering algorithms, both without considering the annotations, were informative of whether the instance corresponded to a healthy or PD subject.

K-means and agglomerative clustering performed better when the number of clusters was set to $n = 4$. In particular

TABLE II: Entropy-based metrics on clustering methods applied to the autoencoder's embedding. The number associated with the K-means and agglomerative clustering rows refer to the number of clusters n, set as a hyper-parameter of the algorithm.

	Homogeneity (%)	Completeness (%)	V-Measure (%)
K-means_2	53.00 ± 25.84	54.79 ± 24.66	53.63 ± 25.53
K-means_4	86.54 ± 14.19	56.15 ± 9.81	67.91 ± 10.99
Aggl._2	48.84 ± 30.27	51.99 ± 28.07	49.98 ± 29.58
Aggl._4	91.27 ± 14.22	59.85 ± 12.36	72.05 ± 12.87
DBSCAN	69.64 ± 41.65	53.83 ± 27.42	59.42 ± 34.87

considering the proportionally high standard deviations of the other approaches, which indicates that, for some folds, the embedding was not informative with respect to the labels. Even the completeness measurement, which could be expected to be lower when the number of clusters is higher than the number of classes, has actually improved. The homogeneity reached 91.27% with agglomerative clustering, an evidence in favour of the autoencoder's features as discriminative towards detection of PD. The density-based DBSCAN showed intermediate results, though with the highest standard deviations.

B. Feature Analysis

Features learnt by each model were analysed based on the attribution methods (Section IV-D) and spectral analysis. We have considered the input features and the internal representations at the intermediate layers of the convolutional networks.

1) Input Features: The attributions with respect to the input channels (i.e., regions of the motor loop) were used to determine the 1-second segments that show the highest accumulated attributions at each instance (i.e., the highest sum of 1,000 subsequent elements within a given channel of a given input), with the constraint that only segments whose sum of attributions is above a threshold of 1.0 were considered. The power spectral density (PSD) of those segments was computed using the Welsh method [28], and the mean μ_{PSD} of the spectra of each class $\mathcal{C} = \{\text{H}, \text{PD}\}$, where H means "healthy" and PD, "parkinsonian", was considered to calculate the ratio R of Equation 1. The rationale is that a peak at the beta frequency band (13-30 Hz) is a relevant marker of PD brain signals [29].

$$R = \frac{\mu_{\text{PSD}}(\mathcal{C} = \text{PD})}{\mu_{\text{PSD}}(\mathcal{C} = \text{H})} \tag{1}$$

Results for each model are presented in Fig. 5, alongside a baseline spectrum corresponding to random segments of each instance. The beta frequency peak can be clearly seen in random segments, and was enhanced on all models except for the autoencoder without fine-tuning. The autoencoder situation was expected, since the gradients were not updated with respect to the inputs of the network, but only to the encoding produced after the convolutions. The M1 and GPi ratios were close to zero because few segments of PD individuals with relevant attributions were present in the analysis.

As expected, the CNN and pre-trained CNN elicited high attributions to segments with similar spectral densities. The

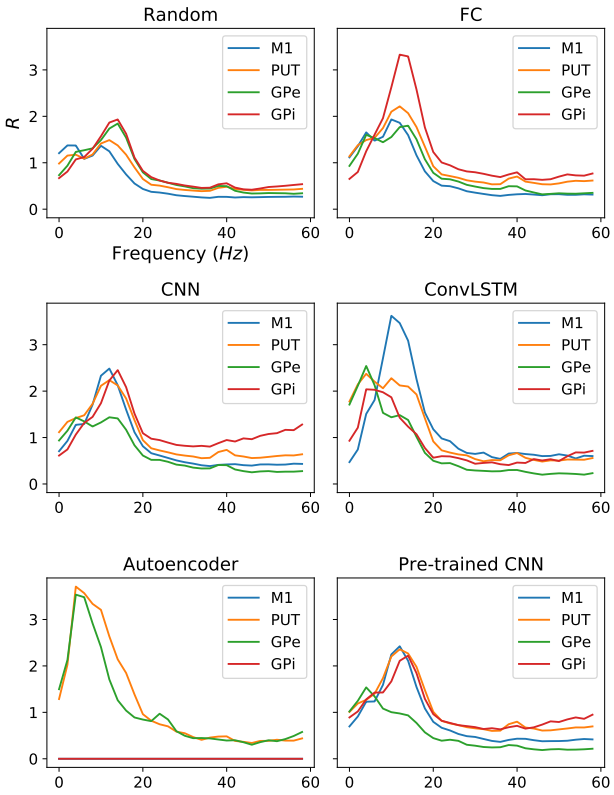


Fig. 5: Ratio between the mean PD and healthy PSD of the 1-second snippets with the highest accumulated attribution per input segment, above a given threshold of 1.0.

FC network was also consistent with the literature, with even a more acute beta peak in GPi. In the ConvLSTM spectrum, this peak was very high in M1, though less prominent in GPe and GPi. These differences in spectra show that each model make predictions based on different input features, however all of them were in consonance with previous PD literature.

To verify the contribution of each region for the models' performance, we evaluated the total number of regions with at least one 1-second segment whose sum of attributions was above the threshold of 1.0, across all folds. In Fig. 6, this evaluation is shown in terms of the proportion of segments above such threshold with respect to the total number of segments within each given region.

This evaluation suggests that the PUT and GPe regions were generally more relevant for recognising the healthy condition for all models, and also for recognising the PD condition for the ConvLSTM and the autoencoder-based network. Therefore, the particularly high frequencies for the GPi spectrum at the baseline FC and for the M1 spectrum at the ConvLSTM, previously shown in Fig. 5, does not imply that those regions have given the highest contributions for the classifications.

2) Intermediate Convolutional Layers: We also evaluated the features at the intermediate convolutional layers. Given the internal representation that followed each max-pooling layer, our analysis computed the spectral power at the delta (1-3 Hz),

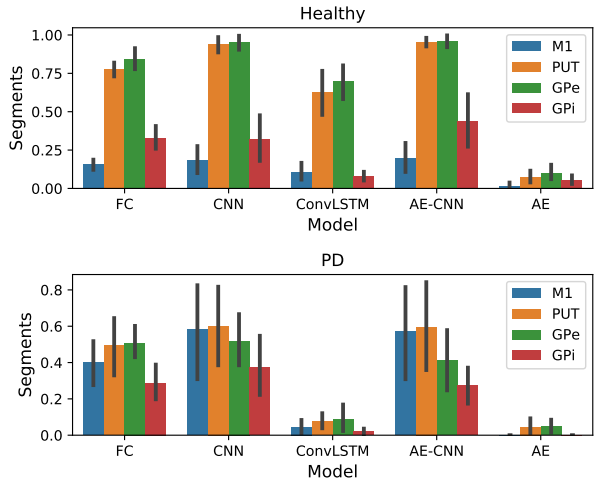


Fig. 6: Proportion of segments above threshold for each classification model (mean between all folds). In the graph, the pretrained CNN was named AE-CNN.

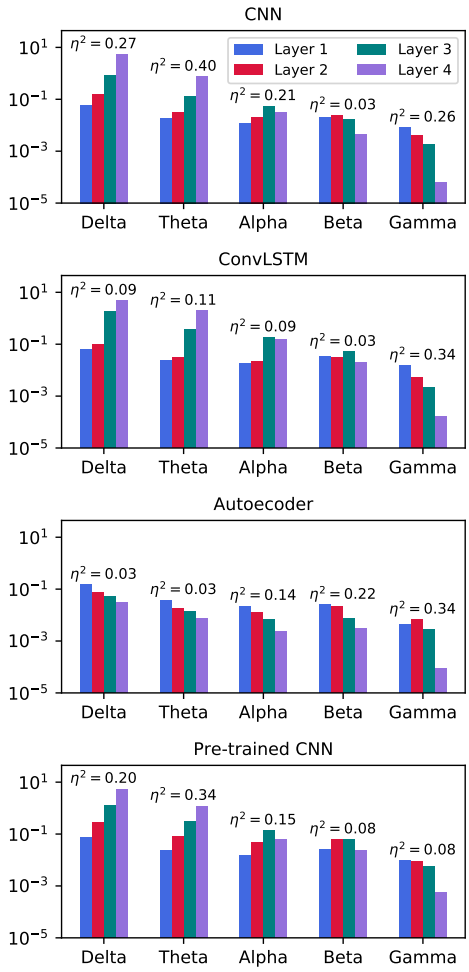


Fig. 7: Average frequency power bands over the spectrum of each max-pooling layer and η^2 effect size over all pairs of layers with summed attributions above a given threshold of 0.5 for the pre-trained CNN or 0.7 for the other models.

theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz) and gamma (30-100 Hz) bands of the LFP. We considered feature maps whose sum of attributions is above a threshold of 0.5 for the pre-trained CNN and 0.7 for the other models. The average spectral power of those representations is shown in Fig. 7.

Differences on the features learnt at each layer were also verified at a given power band. As the number of samples is massive in all of the considered cases, the outcome of a significance test would provide a very low p-value even if the effect of the significance detected was only trivial [30]. In fact, we got $p \approx 0.00$ for all ANOVA tests applied. Hence, in order to understand the effect size of this statistical significance, we measured the η^2 measure [31], also reported in Fig. 7. A small effect size is determined by $\eta^2 \in [0.01, 0.09]$, a medium one, by $\eta^2 \in [0.09, 0.25]$, and a large one, by $\eta^2 > 0.25$.

Except for the autoencoder, the evaluations of all models shared most of its properties. Regarding the sub-alpha waves (i.e., delta and theta), CNN, ConvLSTM, and pre-trained CNN produced feature maps with higher amplitudes the deeper the layer was, with medium to large effect sizes. This pattern started to reverse at the alpha band, with layer 4 producing less power at such frequency interval than layer 3. At the beta band, the pattern was less uniform across models, though relevant (i.e., large effect size for CNN and ConvLSTM). At the gamma frequency, the tendency of the lower bands was reverted, with first layers producing less of those waves. The high effect size was possibly due to the lower resolution at layer 4, which penalises spectral analysis of higher frequencies.

The autoencoder model was considerably different than other models, as one may expect due to its different, unsupervised optimisation strategy. It produced the same pattern at all sub-gamma frequency bands, with a higher prevalence of those frequencies the lower down was the layer. We highlight that only small effect sizes were detected at the delta and theta bands, and medium effects, at the alpha and beta ones. The layers were less specialised regarding the gamma waves. In common with the other models, the autoencoder has also shown a sharp drop of gamma waves at layer 4.

VI. Conclusion and Future Work

In this paper we proposed a deep framework to extract features related to Parkinson's Disease (PD) from Local Field Potential (LFP) brain signals of a marmoset monkey dataset. Different neural networks were applied as machine learning techniques, both as classifiers and autoencoders, and results were reported in terms of accuracy and properties of the representations learnt by each model.

The deep networks presented classification metrics higher than the shallow networks, with accuracy up to 99.80% for the ConvLSTM model. The autoencoder embedding has shown to be informative of the PD-related features, with clustering approaches reaching homogeneity up to 91.27%, and higher

classification metrics when fed to a fully-connected network, in comparison to the raw input (e.g., 95.76% accuracy, against 93.65%). Pre-training the CNN, on the other hand, had little effect compared to training from scratch.

Even though the convolutional networks extract features in the time domain, the input segments with higher attributions presented an enhanced peak at the beta frequency range of the average spectrum of the PD individuals when compared to the healthy ones. Regarding the intermediate representations of the convolutional layers, we have analysed the average power spectra at five frequency bands of feature maps with the highest attributions. Although LFP readings are not a feasible source of data for diagnosing PD, the proposed methods and analysis may contribute for a better understanding of the mechanisms underlying Parkinson's disease.

Future work includes the use of the same deep learning approach to simulated data originated from computational models of PD. This will assist on the validation of artificial models of the motor loop, apart form enhancing our current understanding of the PD neurophisiology. We will also embed such models into a robot, given rise to a neurorobotics model which could simulate the symptoms of this disease and provide a platform to perform preliminary experiments on proposed new therapies. A better understanding of the BG-T-C circuitry might give further insights on related systems regarding decision-making, homeostasis and learning [32]–[35], which are of particular interest to the field of robotics.

REFERENCES

[1] O. B. Tysnes and A. Storstein, "Epidemiology of Parkinson's disease," *Journal of Neural Transmission*, vol. 124, no. 8, pp. 901–905, 8 2017.

[2] J. J. Gaare, G. O. Skeie, C. Tzoulis, J. P. Larsen, and O.-B. Tysnes, "Familial aggregation of Parkinson's disease may affect progression of motor symptoms and dementia," *Movement Disorders*, vol. 32, no. 2, pp. 241–245, 2 2017.

[3] B. S. Connolly and A. E. Lang, "Pharmacological treatment of Parkinson disease: A review," *JAMA - Journal of the American Medical Association*, vol. 311, no. 16, pp. 1670–1683, 2014.

[4] H. Kita and T. Kita, "Cortical stimulation evokes abnormal responses in the dopamine-depleted rat basal ganglia," *Journal of Neuroscience*, vol. 31, no. 28, pp. 10 311–10 322, 2011.

[5] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents–EEG, ECoG, LFP and spikes." *Nature reviews. Neuroscience*, vol. 13, no. 6, pp. 407–20, 5 2012.

[6] J. B. Koprich, L. V. Kalia, and J. M. Brotchie, "Animal models of α-synucleinopathy for Parkinson disease drug development," *Nature Reviews Neuroscience*, vol. 18, no. 9, pp. 515–529, 8 2017.

[7] M. B. Santana, P. Halje, H. Simplício, U. Richter, M. A. M. Freire, P. Petersson, R. Fuentes, and M. A. Nicolelis, "Spinal cord stimulation alleviates motor deficits in a primate model of Parkinson Disease," *Neuron*, vol. 84, no. 4, pp. 716–722, 11 2014.

[8] R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based alzheimer's disease classification from magnetic resonance brain images," *Cognitive Systems Research*, vol. 57, pp. 147–159, 2019.

[9] N. Mammone, C. Ieracitano, and F. C. Morabito, "A deep cnn approach to decode motor preparation of upper limbs from time–frequency maps of eeg signals at source level," *NN*, vol. 124, pp. 357–372, 2020.

[10] C. X. Han, J. Wang, G. S. Yi, and Y. Q. Che, "Investigation of EEG abnormalities in the early stage of Parkinson's disease," *Cognitive Neurodynamics*, vol. 7, no. 4, pp. 351–359, 8 2013.

[11] Y. J. Zhao, H. L. Wee, Y.-H. Chan, S. H. Seah, W. L. Au, P. N. Lau, E. C. Pica, S. C. Li, N. Luo, and L. C. Tan, "Progression of Parkinson's disease as evaluated by Hoehn and Yahr stage transition times," *Movement Disorders*, vol. 25, no. 6, pp. 710–716, 4 2010.

[12] R. Yuvaraj, U. Rajendra 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, 8 2018.

[13] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, 10 2018.

[14] Y. Li, m. Murias, s. Major, g. Dawson, K. Dzirasa, L. Carin, and D. E. Carlson, "Targeting EEG/LFP synchrony with neural nets," in *NIPS*, Long Beach, CA, EUA, 2017, pp. 4620–4630.

[15] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization." *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[16] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in *4th ICLR*, San Juan, Puerto Rico, 2016, pp. 1–15.

[17] S. L. Oh, Y. Hagiwara, U. Raghavendra, R. Yuvaraj, N. Arunkumar, M. Murugappan, and U. R. Acharya, "A deep learning approach for Parkinson's disease diagnosis from EEG signals," *Neural Computing and Applications*, pp. 1–7, 2018.

[18] G. Ruffini, D. Ibañez, M. Castellano, S. Dunne, and A. Soria-Frisch, "EEG-driven RNN classification for prognosis of neurodegeneration in at-risk patients," in *ICANN*. Springer, 2016, pp. 306–313.

[19] Y. Yao, J. Plested, and T. Gedeon, "Deep feature learning and visualization for EEG recording using autoencoders," in *ICONIP-LNCS-vol. 11307*, 2018.

[20] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of EEG signals," *IEEE Access*, vol. 6, pp. 25 399–25 410, 2018.

[21] M. M. McGregor and A. B. Nelson, "Circuit Mechanisms of Parkinson's Disease," pp. 1042–1056, 3 2019.

[22] 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, pp. 1–19, 6 2015.

[23] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," in *NIPS)*, 2018.

[24] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 1 2016.

[25] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for Deep Neural Networks," in *6th ICLR*. OpenReview.net, 2018, pp. 1–16.

[26] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *34th ICML*, vol. 7. IMLS, 2017, pp. 5109–5118.

[27] A. Rosenberg and J. Hirschberg, "V-Measure: A conditional entropy-based external cluster evaluation measure," in *Proc. of the EMNLP-CoNLL*. Association for Computational Linguistics, 2007, pp. 410–420.

[28] K. D. Rao, M. Swamy, K. D. Rao, and M. Swamy, "Spectral analysis of signals," in *Digital Signal Processing*. Springer, 2018, pp. 721–751.

[29] G. Tinkhauser, A. Pogosyan, H. Tan, D. M. Herz, A. A. Kühn, and P. Brown, "Beta burst dynamics in Parkinson's disease off and on dopaminergic medication," *Brain*, no. 11, 2017.

[30] B. Lantz, "The large sample size fallacy," *Scandinavian Journal of Caring Sciences*, vol. 27, no. 2, pp. 487–492, 6 2013.

[31] T. R. Levine and C. R. Hullett, "Eta squared, partial eta squared, and misreporting of effect size in communication research," *Human Communication Research*, vol. 28, no. 4, pp. 612–625, 2002.

[32] P. Vargas, R. Moioli, F. Von Zuben, and P. Husbands, "Homeostasis and evolution together dealing with novelties and managing disruptions," *International Journal of Intelligent Computing and Cybernetics*, vol. 2-3, pp. 435–454, 2009.

[33] R. C. Moioli, P. A. Vargas, and P. Husbands, "A multiple hormone approach to the homeostatic control of conflicting behaviours in an autonomous mobile robot," in *IEEE CEC*, 2009, pp. 47–54.

[34] M. Keysermann and P. Vargas, "Towards autonomous robots via an incremental clustering and associative learning architecture." *Cognitive Computation*, vol. 7-4, pp. 414–433, 2015.

[35] C. Rizzi, C. G. Johnson, F. Fabris, and P. A. Vargas, "A situation-aware fear learning (safel) model for robots," *Neurocomputing*, vol. 221, pp. 32 – 47, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216310529