IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 July 2022, accepted 15 August 2022, date of publication 26 August 2022, date of current version 30 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3201843

RESEARCH ARTICLE

BERT Learns From Electroencephalograms About Parkinson's Disease: Transformer-Based Models for Aid Diagnosis

ALBERTO NOGALES[1], **ÁLVARO J. GARCÍA-TEJEDOR**[1], **ANA M. MAITÍN**[1], **ANTONIO PÉREZ-MORALES**[1], **MARÍA DOLORES DEL CASTILLO**[2], **AND JUAN PABLO ROMERO**[3,4]

[1]CEIEC, Universidad Francisco de Vitoria, 28223 Madrid, Spain
[2]Neural and Cognitive Engineering Group, Centre for Automation and Robotics, Spanish National Research Council, 28500 Madrid, Spain
[3]Facultad de Ciencias Experimentales, Universidad Francisco de Vitoria, 28223 Madrid, Spain
[4]Brain Damage Unit, Hospital Beata María Ana, 28007 Madrid, Spain

Corresponding author: Álvaro J. García-Tejedor (a.gtejedor@ceiec.es)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Hospital 12 de Octubre Committee in Madrid, under Application No. 20/515 and performed in line with the Helsinki Declaration.

ABSTRACT Medicine is a complex field with highly trained specialists with extensive knowledge that continuously needs updating. Among them all, those who study the brain can perform complex tasks due to the structure of this organ. There are neurological diseases such as degenerative ones whose diagnoses are essential in very early stages. Parkinson's disease is one of them, usually having a confirmed diagnosis when it is already very developed. Some physicians have proposed using electroencephalograms as a non-invasive method for a prompt diagnosis. The problem with these tests is that data analysis relies on the clinical eye of a very experienced professional, which entails situations that escape human perception. This research proposes the use of deep learning techniques in combination with electroencephalograms to develop a non-invasive method for Parkinson's disease diagnosis. These models have demonstrated their good performance in managing massive amounts of data. Our main contribution is to apply models from the field of Natural Language Processing, particularly an adaptation of BERT models, for being the last milestone in the area. This model choice is due to the similarity between texts and electroencephalograms that can be processed as data sequences. Results show that the best model uses electroencephalograms of 64 channels from people without resting states and finger-tapping tasks. In terms of metrics, the model has values around 86%.

INDEX TERMS Deep learning, transformers, BERT, neurology, parkinson's disease, electro-encephalography.

I. INTRODUCTION

The brain is estimated to be formed by order of 1011 neurons plus its connections. Its functioning needs the interaction of neurons of different areas organized in networks [56]. The brain's primary functions are the analysis of afferent stimuli, also called sensitive information and the production of motor and cognitive responses or efferences. Between afference and efference, there is an analysis of information that can be altered in pathological states. The interactions between neurons are mediated by molecules called neurotransmitters that allow them to reach different membrane states that produce their activation and depolarization, thus creating an electric stimulus that neurophysiological techniques can register.

Electroencephalography, invented by Hans Berger, is a method for recording superficial brain waves as electroencephalograms (EEGs) [21]. EEGs are a particular type of data called time series. We define them as sets of repeated observations of a single unit or individual at regular intervals over many instances [55]. EEGs can be recorded using

The associate editor coordinating the review of this manuscript and approving it for publication was Md Kafiul Islam.

electrodes placed around the scalp to measure the cerebral cortex electric changes, called superficial EEG. There is also the possibility of implanting the electrodes directly on the brain cortex during surgery, called deep EEG. Still, its use is rare and specific for certain situations, such as epilepsy surgery.

An excellent temporal resolution characterizes superficial EEG. This characteristic supposes that the cortical electric changes are recorded in real-time but have a poor spatial resolution. This problem is because the recorded electric changes may be influenced by different local electric sources making it very difficult to locate the exact anatomic origin of the changes.

Traditionally EEG has been visually analyzed, limiting its applications for functional cortical paroxystic impairments like epilepsy. Visual analysis of EEGs is a potentially flawed process that does not allow the detection of complex patterns. This fact restricts its use in other diseases, such as neurodegenerative diseases with subcortical involvement. Thanks to the recent advances in quantitative EEG analysis, it is possible to characterize different neurological disorders involving subcortical structures and circuits, such as Parkinson's Disease (PD).

PD is a neurodegenerative and progressive disease of the central nervous system. First described by James Parkinson in 1817, its principal characteristic is a progressive deterioration of the neurons in the brain's central area of the substantia nigra. The associated neurodegeneration produces a decrease in dopamine secretion that leads to the appearance of motor and non-motor symptoms that reflect the involvement of different non-dopaminergic pathways [36]. PD, also known as agitant paralysis, is the second most common neurodegenerative disease after Alzheimer's. There are many possibilities that by the year 2040, there will be around 17 million affected, which makes it the fastest-growing neurological illness in the world [46].

According to the updated Movement Disorder Society (MDS) Criteria [43], the diagnosis of this disease is mainly clinic and based on the identification of the cardinal motor manifestations of the disease, rest tremor, bradykinesia, and rigidity. The problem is that these symptoms are evident only when the neurodegeneration of the basal ganglia has reached up to 80%. So, the treatment of the disease remains symptomatic [14], being levodopa the gold standard treatment for the symptoms since 1961 [52].

The patient's response to levodopa is one of the criteria used to confirm the diagnosis of PD. Some patients must reach high doses of levodopa (up to 1 gram) to confirm or rule out the diagnosis by this therapeutic test. This method could lead to a delay in the diagnosis besides the side effects of high levodopa doses. In some cases, the average time to diagnose PD could reach two years [13]. It is crucial to make an early diagnosis to look for preventive treatments.

Although alterations in EEGs are possible in PD patients, according to Yoo *et al.* [60], they have not been entirely justified. As dopaminergic deficit is the principal hallmark explaining the functional changes in this disease, previous works demonstrate the changes produced by this medication [48], so the precise moment when the EEG is registered may be a crucial factor for the analysis. So, the primary motivation of this paper is to find a non-invasive diagnostic method that will prevent patients from taking large amounts of medicine that could be harmful to their health. In this respect, the most useful would be raw EEGs (without transformation that could lead to information losses) as they are the default physiological brain signal. Also, visual recordings of PD must not allow physicians to make diagnoses, so we need to apply Artificial Intelligence techniques. These techniques nowadays allow the processing of large amounts of data. In particular, deep learning techniques are models that comprise multiple processing layers to learn data representations with various abstraction levels [12]. The use of such analysis applied to EEGs implies its use as an early diagnosis method with an impact on disease characterization and management.

In this paper, deep learning techniques from the Natural Language Processing (NLP) research area have been applied to build a model for characterizing Parkinson's EEG changes in different states of dopaminergic stimulation. These changes are compared with those from controls and the obtained differences. Then, this information could give hints for early diagnosis of the disease. Considering the nature of the data (texts and EEGs) can be processed as data sequences. Within all the NLP models, we use Bidirectional Encoder Representations from Transformers (BERT) as the last breakthrough in the area [16]. The paper's main contribution is the application of this model that will lead to obtaining a non-invasive method to help clinicians diagnose PD. As far as we know, this is the first time BERT has been adapted to an EEG classification task for diagnosis. This fact is endorsed by Maitin *et al.* [34], a review of machine learning techniques for PD classification.

The main benefit of applying deep learning techniques for diagnosing PD using EEGs is that there are no evident brain structural alterations as may be the case of epilepsy, and the functional changes such as motor performance depend on the dopaminergic stimulation. Thus, the cortical activity may vary depending on the degree of degeneration. The external dopamine administration makes it quite challenging to differentiate from healthy subjects depending on the patient's functional state.

The rest of the paper is structured as follows. Section 2 summarizes the state of art related to computer science models and PD diagnosis. Section 3 describes the dataset used in the research and defines the methods used. Section 4 discusses the results obtained during the study. Finally, section 5 gives some conclusions and suggests some future works.

II. RELATED WORK

There are many studies of EEGs with classical machine learning techniques. A work that uses EEGs from Alzheimer's patients can be found in Podgorolec. It applies subspace methods and its version of decision trees. In another case, Sohaib *et al.* use different machine learning algorithms to classify brain activity changes related to emotions from EEGs. Reference [27] show a comparison of algorithms like Artificial Neural Networks (ANN), Naïve Bayesian, K-Nearest Neighbors (KNNs), Support Vector Machine (SVM), and K-Means for the recognition of epileptic seizures. Some previous methods, plus tree bagging or random forest, have been applied in classifying brain states related to activities such as reading or playing video games [31]. Finally, Wang *et al.* present a use case in measuring sleep quality with KNN, SVM, and discriminative Graph regularized Extreme Learning Machine (GELM). The study concludes that the gamma band is the most relevant for sleep quality assessment.

As can be seen, all the previous papers solve EEG classification tasks related to different cases: brain activities based on emotions, brain states when performing activities, sleep quality, Alzheimer and epileptic seizures. In our study, the classification task aims to discriminate between EEGs of healthy people a PD patients. Another difference is the usage of DP techniques against classical ML methods.

Some works also use classical techniques for classifying PD, sometimes using EEGs. For example, Altay and Alatas [3] evaluates different algorithms modeling the task of PD diagnosis as a multi-objective problem using several characteristics of voice recordings. In [58], EEGs alongside PET images obtain neurophysiological biomarkers using measures like reliability or coherence. These biomarkers let to discriminate between healthy and PD patients and give a level of affection based on the Unified Parkinson's Disease Rating Scale (UPDRS). Classification of Parkinson's severity into five different groups is approached in [11]. This work uses SVM and K-Nearest Neighbors. Another work is [19], classifying EEGs according to three levels of cognition by applying the Boruta algorithm for feature extraction and random forest for the classification. Finally, Vaneste *et al.* use SVM for classifying Parkinson's EEGs to search for spectral equivalence between various neurological (PD between them) and neuropsychiatric disorders with Thalamocortical dysrhythmia. If we compare the previous work with ours, some of them perform the same task, classifying between PD and healthy, but none apply DL techniques.

Several works have these characteristics in the case of DL using EEGs since these techniques appeared a few years ago. For example, it has an application in movement recognition. Reference [61] use Long Short Term Memory (LSTM) networks with attention modules to classify left and right-hand movements based on EEGs. In [40], EEGs with neural networks identify movements that let a user control a LEGO robot. Refernce [47] apply a deep learning model called Convolutional Neural Network (CNN) to classify EEG changes related to motor tasks like moving hands or feet. The first visual object classifier driven by EEGs [51], uses a hybrid CNN and Recurrent Neural Networks (RNNs) model that discriminates 40 class images. CNNs are also applied by Achayra *et al.* to detect epileptic seizures in EEGs automatically. In [59], another hybrid model with CNN and RNN classifies affective mental states. Also, in [38], a particular RNN called LSTM, alongside a neural network classifier, is used to discriminate normal, pre-seizure, and seizure states. In [62], EEG-based emotion recognition uses a simple deep learning model, a CNN model, an LSTM model, and a hybrid model of the previous two. The diagnosis of REM Behavior Disorder (RBD), a sleep disorder commonly associated with PD, is studied using CNNs and RNNs using spectrograms of the EEGs [45]. Finally, Gemein *et al.* [20] evaluate classic methods like SVM vs. Temporal CNN to classify pathological and non-pathological EEGs. In the previous works, different EEG tasks have been achieved: image discrimination, epileptic seizures or epileptic states detection, and emotion or movement recognition. The models used are typical architectures applied to EEGs like CNNs and RNNs. In our work, we are focused on discriminating between EEGs of PD patients and healthy people, and our main contribution is demonstrating that complex NLP techniques like BERT can be used with EEGs.

Some research can be highlighted in the particular case of PD and deep learning. Reference [41] built a CNN classifier for aided diagnosis to analyze images of handwritten figures. Also, Eskofier *et al.* [17] studied PD with CNN trained with pictures of drawings, but in this case, focused on the detection of bradykinesia. Another work is by Camps *et al.* (2017), where a typical alteration of PD, Freezing Of Gait (FOG), is detected using CNNs in data collected with a wrist-worn accelerometer. Ogawa and Yang stand out for using voice recording and CNN for a PD classification. Another approach that uses CNN is [39] processing EEGs as images obtaining an accuracy of around 88% in the discrimination between Parkinson's and normal EEGs. Few previous deep learning reports applied to PD studies mainly used CNNs and RNNs. Some use clinical data as different features and the PD rating scale; some use neurophysiological signals as the Rapid Eye Movements neurophysiological registers. Most of them use neuroimaging as Magnetic Resonance Image (MRI) or Single Photon Emission Computed Tomography (SPECT) imaging, as described in [1], [28], [49], and [57], respectively. As far as we know, there are no papers where EEGs of PD have directly been used with BERT models as described in our paper.

In this paper, inspired by the language representation model BERT, we developed a neural model to process and classify EEGs diagnosing if a patient suffers from PD or not. The main novelty of this work is the direct use of EEGs (for being a non-invasive technique) to diagnose PD with BERT models.

III. RESOURCES AND METHODS

The following subsections describe the resources used in this work and the techniques applied. First, a brief description of the EEGs and their collected data. Secondly, a formal definition of the deep learning models that have been applied.

A. A DATASET OF EEGS WITH PD PATIENTS AND CONTROLS

The data in this research corresponds to some EEG tests on patients with PD and healthy people. EEGs are collected by electrodes positioned along the scalp that measure the brain's electrical activity. The information, arranged in channels, is the difference in potential between a reference electrode and the active one. Also, different systems can be considered depending on the positions of the electrodes.

In the present study, EEGs used 64 channels and the 10-20 system. This system means that electrodes are spaced between 10% and 20% of the total distance between some particular skull points. Another critical parameter is the frequency which means the number of measures taken in one second. In this case, the frequency is 512 Hz which is one measure every 1.9531 milliseconds.

1) PARTICIPANTS

Eighty patients were recruited in the movement disorders clinic of Hospital Beata María Ana in Madrid from March 2018 to February 2022. 24 age and gender-matched controls were also recruited among relatives and companions of the patients. All the patients had been diagnosed with PD according to London Brain Bank criteria (mean time from onset years), with Hoehn and Yahr (HY) scale (range I-III).

Exclusion criteria included patients using advanced therapies (apomorphine pump/duodenal dopamine infusion) for PD, epilepsy history, or structural alterations in previous imaging studies. Montreal Cognitive Assessment (MoCA) score <25, Nazem *et al.*, poor response to levodopa or suspicion of atypical parkinsonism, any other neurological disease, or severe comorbidity. Inclusion criteria for patients with PD were to be over 18 years of age; idiopathic PD diagnosed according to London brain bank criteria Hughes *et al.* (1992), stage <III Hoehn-Yahr, not having noticeable motor fluctuations, and clinical stability (not having changed the anti-dopaminergic medication in the last 30 days or anti-depressives during the previous 90 days).

CEIC Fuenlabrada Hospital, Madrid, Spain, approved the research protocol. All subjects gave written informed consent following the Declaration of Helsinki.

2) INTERVENTION

EEG comprised 64 electrodes placed according to the 10-20 system. Resting EEG activity was recorded over one minute; every subject was comfortably seated with their hands on their laps, relaxed jaw, and eyes open, looking at a white wall. Immediately afterward, each patient has to tap the thumb with the index finger of the left hand (left finger tapping) continuously for five intervals of 30 seconds. Finally, the patient repeated the former task with the right hand (right finger tapping). Healthy controls EEG were also recorded in the same conditions.

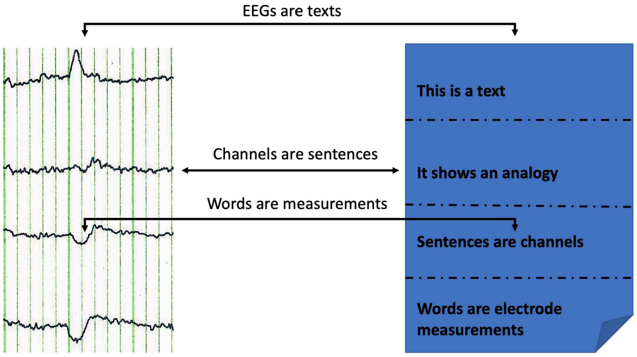

FIGURE 1. EEG and text analogy where a channel corresponds to a sentence and a word to a measure.

3) DATA COLLECTION MATERIALS

actiCHamp amplifier (Brain Vision LLC, NC, USA) was used to amplify and digitize the EEG data at a sampling frequency of 512 Hz. The EEG data were stored in a PC running Windows 7 (Microsoft Corporation, Washington, USA). EEG activity was recorded from 64 positions (channels) with active Ag/AgCl scalp electrodes (actiCAP electrodes, Brain Vision LLC, NC, USA). The ground and reference electrodes corresponded to AFz and FCz, respectively. EEG acquisition was carried out by NeuroRT Studio software (Mensia Technologies SA, Paris, France).

4) DATASET SUMMARY

The dataset was automatically extracted from the EEGs. In total, it consists of 80 Parkinson patients (48 males; age: 63,89 ± 9,21 years; disease duration: 7,21 ± 4,54; stage of Hoehn-Yahr: 2.99 ± 1.35) and 24 healthy patients (19 males; age: 58,12 ± 6,91) that serve as control. From each EEG was extracted both finger tapping tasks of about 30 seconds of duration and one test of about 1 minute from the resting state. So, each patient has 3 EEGs. Summing them all up makes a total of 240 different tests for patients and 72 different tests for controls in the dataset with a total duration of 12,480 seconds. Although that amount of data seems small to train a BERT model, some papers have demonstrated its good performance with small datasets. For example, Barz and Denzler [8] obtains accuracies of over 80% with datasets of 10 samples per class. Also, Elze-Can [18] obtains good metrics, near 80% accuracy, with a dataset of around 100 instances per class.

B. NLP TECHNIQUES TO CLASSIFY EEGS

Every channel of an EEG is a sequence of values measuring potential differences at each point of the process. NLP state-of-the-art neural models can process sequences efficiently to generate different outputs. These models can even attend to other parts of an input sequence to produce the desired result ([6], [16], and [54]).

This paper considers a parallelism between EEG and texts. An EEG channel is a sequence of measurements like a

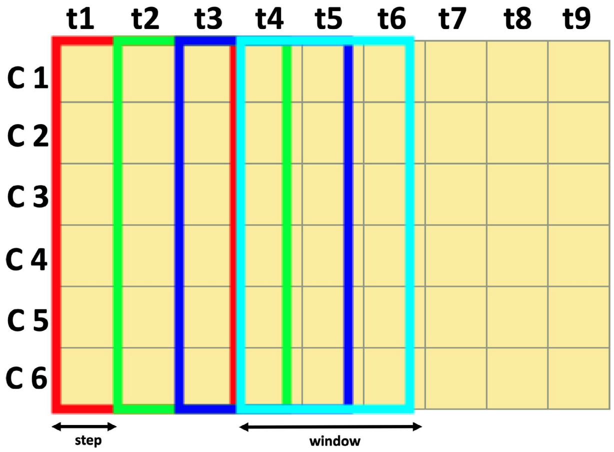

FIGURE 2. Sliding window to process an EEG of 6 channels with four windows.

sentence is a sequence of words. Then, suppose the meaning of a word in a sentence depends on the previous and subsequent ones. In that case, a measurement in an EEG can be understood by analyzing the previous and following ones, which describe the brain's activity in a particular moment. Furthermore, an EEG can be considered like a text formed by a set of sentences (the different channels) that the models implemented in this work will process as a whole. Then, the model will consider the values of all channels for a particular moment as the minimum input data unit. Figure 2 shows an example of this analogy between EEGs and texts. This analogy assumes that EEG values have a local context like words in a sentence. It is considered that, in an EEG, a specific sequence of values is more likely to be observed than others, just like in a particular sequence of words. The decision about using NLP techniques is based on this analogy but the strategy used to preprocess the EEGs is quite different and will be explained later.

1) BERT MODEL

Among all the NLP models, we have decided to use BERT as its performance has been obtaining excellent results recently. This model uses stacked Transformers, a revolution in the field in 2017, [54]. Transformers have encoder-decoder architectures, a model aiming to reduce the input data into a small piece containing the most relevant info (encoder) and then upsample it until the output data is obtained (decoder). In Transformers, the encoder comprises six identical layers with two sublayers: a self-attention layer and a feed-forward layer. The encoder seeks to code a specific word (EEG measure in our case) of the input data while considering other relevant ones. The decoder has a similar architecture but also implements a multi-head attention sublayer connected to the output of the encoder.

BERT is implemented based on this architecture but using only de encoder part. It is considered a multilayer bidirectional Transformer encoder formed by six stacked Transformer encoders. Input data goes through an embedding layer and a positional encoding layer. The former transforms each EEG measure into an n-dimensional vector. The latter provides the positions of each element in the input data. As has been said before, each encoder has two sublayers: self-attention and feed-forward, which also receive information from a residual layer. This layer aims to introduce information from previous states that could be lost during the data processing [16]. BERT's latest Transformer connects to a simple neural network classifier with several hidden layers and a bicategorical output layer using a SoftMax activation function. SoftMax will let BERT discriminate between Parkinson's patients and healthy people [16].

C. EVALUATION METRICS

The four metrics used to evaluate the models are accuracy, specificity, sensibility, and precision. Accuracy is the number of correct predictions divided by the total number of performed predictions. The interpretation serves as a guide to measuring the performance of the approaches. Specificity measures the ratio between the number of true negatives (healthy people diagnosed as healthy people) and the total of those predicted as true negatives and false positives (healthy people diagnosed as Parkinson's patients). This metric avoids healthy people taking the medication when they do not need it. Precision measures the ratio between the number of true positives (Parkinson's patients diagnosed correctly) and the total of those predicted as true positives and false positives, which is interesting in terms of economic costs. Sensitivity is the same as precision but considers false negatives (Parkinson's patients diagnosed as healthy) instead of false positives, which is very useful to avoid undiagnosed patients.

IV. RESULTS AND DISCUSSION

A. DATA PREPROCESSING

As texts, EEGs have the particularity that a value in a specific moment needs to consider the previous values to be understood. In our case, EEGs must be evaluated using what is happening in all the channels at given moments. This approach determines how EEGs get into the neural models. A sliding window mechanism uses all the channels at the same time and splits each EEG into different small pieces. The use of small data has the advantage of reducing the input data and allowing a more populated dataset with small instances. This sliding window has two parameters to decide how to create the instances. The first parameter is called the step and controls how much the start of a window is shifted concerning an instant of the EEG, which is the beginning of a previous window. The second parameter is the width and controls the number of values between the window's start and end. In the present work, these parameters have the following values: step comprises 95% of the data and width of 256 instances. In this way, we go through the EEG employing windows with an overlapping of 5% to maintain its continuity. Fig. 3 describes this paragraph. C1 to C6 denote six channels, and t1 to t9 are nine timestamps corresponding to the win-

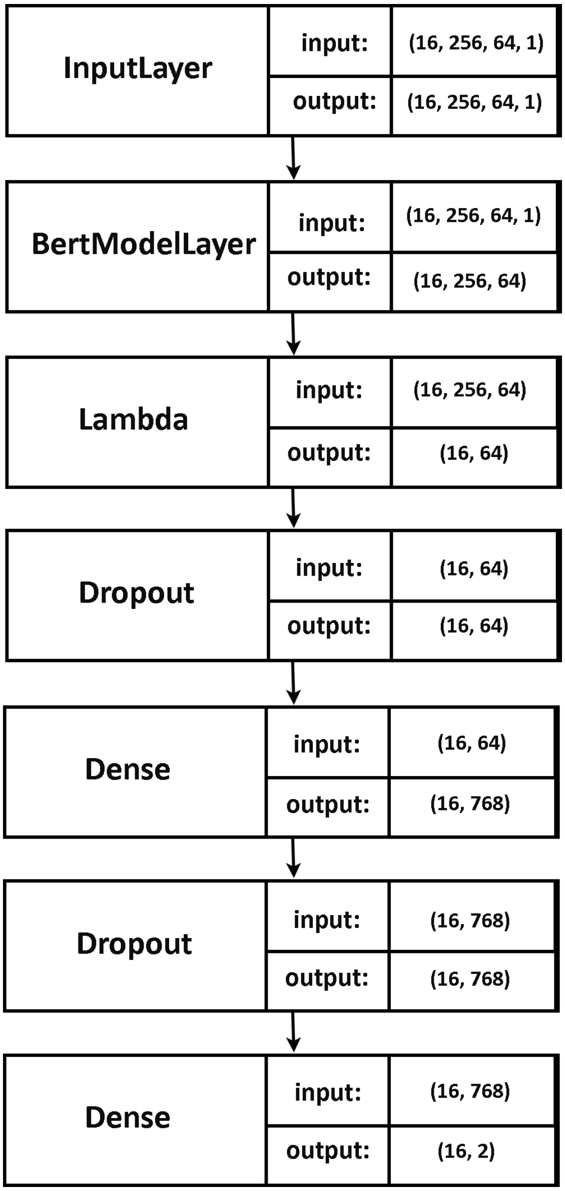

FIGURE 3. Architecture of the 64 channels model.

dow's beginning. Different colors represent windows; in this case, there are four windows.

B. TREATING EEGS AS SEQUENCES OF WORDS

This research has adapted BERT models based on an analogy between EEGs and texts. In a BERT-based model, the first layer is word-embedding, which takes a word as input and returns its vector representation. In the case of EEGs, each timestamp of all the channels has been considered the minimum input data unit. Then, the input vector has a set of values of a time window using the different channels of an EEG. This input uses a vector with the length of the number of channels in the EEG. Each vector's value is a brain activity measure for a channel at a particular time. In this case, the embedding layer is removed since our data has a numerical representation. Notice that the lack of the embedding layer reduces the size of the classification models and thus saves training time. Moreover, a large amount of data is needed to obtain good embeddings, around billions of words for good word embeddings [32].

C. BERT-BASED MODELS TO CLASSIFY EEGS

To compare the results, we develop two experiments based on BERT models with the same architecture. First, a model is trained and focused on processing the 28 most interior channels, assuming that the peripherical channels add noise and predict if it comes from a person with PD or not. Then, a model is implemented processing a 64-channels-EEG which means using all the information collected in the EEGs.

1) THE TRAINING STAGES

Both experiments are trained, including all the EEGs (both tappings and resting state) for each individual (training strategy 1) and then removing the corresponding to a resting state (training strategy 2). The motor task that has been chosen is finger-tapping, consisting of a self-cued repetitive opposition of the thumb and index of each hand is one of the most informative tasks included in clinical evaluations such as the UPDRS. The reason is that the hand has a pre-dominant somatotopic representation in basal ganglia and is one of the earliest locations of motor alterations identified in the disease [30]. On the other hand, the resting state has been extensively used in functional magnetic resonance imaging (fMRI) to study functional connectivity among specific brain regions organized into networks [26]. These networks' dynamics and disruption may be associated with various diseases. The resting-state has been extensively used to study EEG microstates [29] that are altered in PD depending on the dopamine administration [48].

Then, we have to split the dataset into training, validation, and test subsets. Train and validation comprise the training stage, and then the test stage is used to do new classifications of EEGs. In this case, 80% for training and validation applying 5-fold cross-validation, and 20% of the cases were used for the test (examples never seen by the model during training). The different subsets were chosen randomly in terms of individuals but always maintained the percentage of patients and controls. Although BERT-based models can work with unbalanced classes [37], this double validation allows us to eliminate the bias produced by the choice of data and to identify failures during the training process through the use of the CV method, and to verify the generalization capacity of the model by means of a test blind set. The split into train/validation/test sets was carried out guaranteeing patient independence, and then the division into windows was performed. The classification models give a result that belongs to a particular instant time of an EEG for a specific class. The final classification probability is the average of the probabilities for each EEG fragment.

Processing EEGs is a complex task due to a large number of values. This fact is reflected in the times needed to train the proposals with a CPU AMD Ryzen Threadripper 2950x 16 core, four NVidia GeForce RTX 2080 11 Gb RAM GPUs running on Ubuntu 20.04.3. The training uses a Python script that uses 49 libraries. The 28 channels solution is trained during five epochs for three days, 12 hours, and 29 minutes with all the EEGs and one day and 12 hours without the resting state. In the case of 64 channels, it is trained during five epochs and needs three days, 16 hours, and 48 minutes in the first case and one day and 12 hours in the second. Regarding parameters to be trained: the solution for 28 channels has 583,842, and the one for 64 channels has 1,374,978.

2) 64-CHANNELS-EEG MODEL

The 64-channels-EEG architecture is in Figure 3. It implements a BERT model followed by a classification module without the embedding part. The input is already a dense vector representation of the information from the EEG values. BERT module has 6 Transformer encoders having 64 neurons and four attention-heads. Between each Transformer, we set a feed-forward layer with 1,536 neurons, Gaussian Error Linear Units (GELU) function, and a dropout of 0.3. The classification module is a multilayer perceptron with an input layer of 64 neurons, a hidden layer of 268 neurons, followed by a SoftMax output over two classes representing the PD or Non-PD possible labels for the EEGs. Figure 3 illustrates the model and Table 1 summarizes its parameters.

TABLE 1. Parameters of the 64 channels model.

Module		Parameter	Value
BERT		"attention_probs_dropout_prob"	0.3
		"hidden_act"	"gelu"
		"hidden_dropout_prob"	0.3
		"hidden_size"	64
		"initializer_range"	0.02
		"intermediate_size"	1536
		"max_position_embeddings"	5120
		"num_attention_heads"	4
		"num_hidden_layers"	6
		"type_vocab_size"	2
		"vocab_size"	30522
Classification	Dropout layer	dropout	0.5
	Dense layer	units	768
		activation	"tanh"
	Dropout layer	dropout	0.5
	Dense layer	units	2
		activation	"softmax"

3) 28-CHANNELS-EEG MODEL

Goncharova *et al.* [22] claim that electrodes situated on peripheric areas of the brain are more suitable for collecting noise. Considering that, the 64-channel-EEG baseline model is replicated but uses only the most interior 28 channels. The elimination of peripheric electrodes does not affect the central electrodes, which recollect the information from the primary motor and sensitive areas. These areas expect to reflect most of the changes produced by the dopaminergic stimulation changes in the disease. The reason for selecting these particular channels is two-fold. First, to confirm the previous hypothesis, and second because they allow us to maintain the four attention-heads in the model's architecture.

D. CLASSIFYING PARKINSON'S PATIENT

Trained approaches compile accuracy, specificity, sensitivity, and precision as metrics. Since we are dealing with a medical use case, the metrics should consider false positives and false negatives [33]. In this work, a false positive is a healthy person misdiagnosed with PD. A false negative is a person with PD diagnosed as healthy.

After training the 28 channel models with both training sets (with and without resting-state EEGs) during five epochs, we obtained results from Table 2. It contains the four metrics for both pieces of training, separating training validation and splitting with its standard deviation.

TABLE 2. Evaluation of the 28 channels models with both trainings.

Model	Training without resting states			Training with resting states		
	Train	Valid.	Test	Train	Valid.	Test
Accuracy	75.98% ± 6.59	65.67% ± 5.16	70.16% ± 10.01	72.30% ± 4.71	68.75% ± 8.66	71.30% ± 5.58
Specificity	52.05% ± 30.01	61.89% ± 27.47	30.53% ± 25.06	21.65% ± 24.92	19.92% ± 20.64	18.08% ± 20.64
Sensitivity	87.47% ± 10.05	82.73% ± 11.52	90.00% ± 10.00	96.58% ± 44.13	92.20% ± 44.16	98.00% ± 44.72
Precision	79.22% ± 20.57	71.20% ± 12.65	72.15% ± 19.44	72.05% ± 17.27	70.65% ± 19.19	70.47% ± 17.61

As seen in Table 2, we can interpret the results by considering the bias-variance trade-off [9]. First, bias seems accurate in some metrics, as diagnostic accuracy is slightly over 80% [44]. In terms of variance, the model trained without resting tests has good results in all metrics except specificity due to its differences between stages. Similar results happened when resting states.

If we analyze the results in-depth, we can see that the variability of results for the true and false negative (sensitivity and specificity) without resting states is lower than using these tests but still significantly high. In this experimentation (28 channels), we have less data than in the other case by dispensing with one of the EEG tests. However, percentage-wise, the difference between the classes is maintained. This result affects the specificity metric, as we can see in the results. In addition to having a significantly low value, its standard deviation exhibits high values, around 30%. When the resting test remained unused, we did not observe significant differences between the precision and accuracy metrics results.

Analyzing the results of all the tests, we found the following. On the one hand, sensitivity, a metric responsible for providing the rate of true positives, has values above 90% in all stages of experimentation (that is, train, validation, and test). Still, it exhibits very high deviation values, around 44% in all cases. On the other hand, specificity, the metric responsible for providing the rate of true negatives, has very low values, around 20%. When performing a 5-fold strategy, we find high variability in the results of each fold for the true negatives (Non-PD predictions that are non-PD) and false negatives (Non-PD predictions that are PD). Given that the model is trained with two classes, one being the majority, these results lead us to think that the model may be over-training the majority class (PD). Therefore, fluctuations can be found in each fold when the prediction of non-PD is produced.

The results regarding the precision and accuracy of the model do not differ much in both pieces of training. Where we do find differences is in the sensitivity and specificity metrics. From the above evaluation, we can conclude that class imbalance has negatively impacted the training. Although this could be a problem, BERT has demonstrated promising results by working with imbalanced datasets [37]. Now, we train a 64-channel model to verify the collected noise hypothesis commented above. So, the model has been trained with the same data and conditions during five epochs. Table 3 compiles the information of the four metrics for training, validation, and test.

TABLE 3. Evaluation of the 64 channels models with both trainings.

Model	Training without resting states			Training with resting states		
	Train	Valid.	Test	Train	Valid,	Test
Accuracy	93.45% ± 2.82	84.67% ± 3.79	86.41% ± 7.08	76.04% ± 5.85	68.41% ± 10.08	67.93% ± 5.59
Specificity	93.08% ± 1.66	84.94% ± 11.52	96.00% ± 8.94	54.11% ± 9.47	44.26% ± 22.28	36.00% ± 8.93
Sensitivity	93.64% ± 4.21	84.56% ± 9.00	81.63% ± 11.58	86.54% ± 1.63	79.98% ± 9.33	83.96% ± 10.99
Precision	96.59% ± 26.93	92.15% ± 23.90	97.61% ± 23.72	79.77% ± 19.69	75.00% ± 19.37	72.35% ± 5.59

As can be seen in Table 3, the 64 channels model has better results than the 28 channels one. Results for training without resting tests seem very good in terms of bias and variance, except for precision due to its high deviation. However, it should be noted that, in the test case, for the false positives, there is a fold that contains very different values from the rest of the folds. Therefore, these results alter the measurements of the metrics that include this value. Since it only appears in one of the folds, we de-duce that it is a specific event derived from a data division and not from an error in the training.

Only the false negative values have shown a specific variability in the folds in the case using resting states, much less than in the previous cases. This fact is reflected in the values obtained from the metrics and their standard deviation, where low values of the specificity metric still prevail with high variability between the training processes. However, the results of this experiment do not indicate an affectation by the imbalance of classes since there are no significant variations in the case of the True Negatives, while in False Negatives, said fluctuation has dropped considerably. This issue may occur because, considering more data, the model can better relate the information, minimizing the effect caused by class imbalance. We can corroborate this result with the increased precision and accuracy metrics concerning the model of 28 channels using all data.

E. COMPARISON WITH BASELINES

As a final way to check the performance of our model, we are comparing it with two classical deep learning models widely used with EEGS: CNNs and RNNs with Gated Recurrent Units (GRUs). Both models are inspired by Shi *et al.* [50] but have been adapted to our data. We trained both models in the same conditions as our BERT model with underfitting results. So, we decided to augment the number of epochs to obtain well-trained models. The results of this comparison are in Table 4.

TABLE 4. Evaluation of the 64 channels models with both trainings.

Model	Train accuracy	Test accuracy
64 channels BERT model	93.45% ± 2.82	86.41% ± 7.08
CNN	94.97% ± 0.04	93.49% ± 0.06
RNN	76.12% ± 0.04	70.92% ± 0.09

As seen above, our model improves the results of the RNNs but is slightly worse than the CNNs. However, it should be considered that we needed more epochs to obtain a non-underfitting model. We also want to remember that this work aims to demonstrate that powerful NLP techniques like BERT can be used in biosignal processing. In fact, there is a tendency to use these models in other fields. For example, He *et al.* [24] uses BERT for image classification. In this way, the next step would be to test the performance of BERT and EEGs in a more complex problem that could be difficult to solve with CNNs.

V. CONCLUSION AND FUTURE WORKS

The main aim of this work has been to develop a neural model that could differentiate between Parkinson's patients and healthy subjects using EEGs as time series and taking advantage of NLP techniques. For this purpose, first, we have collected a set of EEGs from PD subjects and controls. Parkinson's EEGs have been recorded in several conditions, considering that there may be significant changes according to the degree of the disease or even with motor activation. Then, we retrained different versions of the BERT model to prove our hypothesis. Also, additional training strategies have been developed to achieve the results.

We obtain two main conclusions. First, EEGs without resting states help the models discriminate better between Parkinson's patients and healthy controls than only finger tapping EEGs. Secondly, the model corresponding to a 64 channels model best differentiates between PD and healthy subjects. To summarize, our main conclusion is that 64 channels model without resting EEG was the best option in this case. Results in different metrics are around 86% of performance classifying EEGs between a patient with Parkinson's and a healthy subject.

This value may occur because a BERT model requires more data to perform training, and removing part of the electrodes does not contribute to improving the results of the classification problem. We could also think that, in the case of PD, the affected area extends to peripheral regions; therefore, these electrodes also contain information about the disease. New training with an intermediate number of channels will be required to test this hypothesis.

However, it draws our attention that when comparing 28 without resting tests and 28 with all tests, we did not find much difference between the precision and accuracy results, only in the sensitivity and specificity metrics that seem to be influenced by class imbalance. This fact makes us think that motor tests are significant when diagnosing PD, while the resting test plays a secondary role.

The results of the 64 channels experiment with all tests differ from those of 64 without resting states. Since when comparing 28 channels with all tests and 28 channels without resting test, we do not find significant differences between their precision values. We may deduce that in the case of 64, everything the training has been insufficient. Remember that the hyperparameters in each experiment are the same to facilitate the comparison and evaluation of the results depending on the channels and EEG tests performed.

This study is not without limitations. Firstly, we cannot determine why resting tests are crucial in the model but are not enough to differentiate them when studied separately. In future studies, a more considerable amount of EEG recordings will help us to reinforce our conclusions. Secondly, further studies should be done with more EEGs in the resting state. Another study that could help us understand the differences between EEGs with electrodes alongside the entire scalp (64 channels) and only central electrodes (28 channels) could be an analysis by zones.

In future works, apart from experimenting with an intermediate number of channels, there is an interest in studying the brain connectivity in PD. For example, we divide the brain into several zones, using a BERT model for each of them, and then making a final diagnosis based on the previous models. Another exciting study uses Graph Convolutional Neural Networks alongside graph theory metrics by modeling Parkinson's EEGs as graphs. Finally, we want to make another diagnosis of PD patients that could evaluate how advanced the disease is.

REFERENCES

[1] M. P. Adams, A. Rahmim, and J. Tang, "Improved motor outcome prediction in Parkinson's disease applying deep learning to DaTscan SPECT images," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104312.

[2] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizures using EEG signals," *Comput. Biol. Med.*, vol. 100, pp. 270–278, Sep. 2018.

[3] E. V. Altay and B. Alatas, "Association analysis of Parkinson disease with vocal change characteristics using multi-objective metaheuristic optimization," *Med. Hypotheses*, vol. 141, Aug. 2020, Art. no. 109722.

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[5] C. Babiloni, M. F. De Pandis, F. Vecchio, P. Buffo, F. Sorpresi, G. B. Frisoni, and P. M. Rossini, "Cortical sources of resting-state electroencephalographic rhythms in PD related dementia and Alzheimer's disease," *Clin. Neurophysiol.*, vol. 122, no. 12, pp. 2355–2364, 2011, doi: 10.1016/j.clinph.2011.03.029.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[7] P. Baldi and P. J. Sadowski, "Understanding dropout," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2814–2822.

[8] B. Barz and J. Denzler, "Deep learning on small datasets without pre-training using cosine loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1371–1380.

[9] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 32, pp. 15849–15854, Aug. 2019.

[10] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Nov. 1985.

[11] N. Betrouni, A. Delval, L. Chaton, L. Defebvre, A. Duits, A. Moonen, A. F. G. Leentjens, and K. Dujardin, "Electroencephalography-based machine learning for cognitive profiling in Parkinson's disease: Preliminary results," *Movement Disorders*, vol. 34, no. 2, pp. 210–217, Feb. 2019.

[12] Y. Bengio, A. C. Courville, I. J. Goodfellow, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[13] B. R. Bloem and F. Stocchi, "Move for change part I: A European survey evaluating the impact of the EPDA charter for people with Parkinson's disease," *Eur. J. Neurol.*, vol. 19, no. 3, pp. 402–410, Mar. 2012.

[14] H.-C. Cheng, C. M. Ulane, and R. E. Burke, "Clinical progression in PD and the neurobiology of axons," *Ann. Neurol.*, vol. 67, no. 6, pp. 715–725, 2010, doi: 10.1002/ana.21995.

[15] A. Coenen and O. Zayachkivska, "Adolf beck: A pioneer in electroencephalography in between Richard caton and Hans Berger," *Adv. Cognit. Psychol.*, vol. 9, no. 4, pp. 216–221, Dec. 2013.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[17] B. M. Eskofier, S. I. Lee, J.-F. Daneault, F. N. Golabchi, G. Ferreira-Carvalho, G. Vergara-Diaz, S. Sapienza, G. Costante, J. Klucken, T. Kautz, and P. Bonato, "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson's disease assessment," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 655–658.

[18] A. Ezen-Can, "A comparison of LSTM and BERT for small corpus," 2020, *arXiv:2009.05451*.

[19] V. J. Geraedts, M. Koch, M. F. Contarino, H. A. M. Middelkoop, H. Wang, J. J. van Hilten, T. H. W. Bäck, and M. R. Tannemaat, "Machine learning for automated EEG-based biomarkers of cognitive impairment during deep brain stimulation screening in patients with Parkinson's disease," *Clin. Neurophysiol.*, vol. 132, no. 5, pp. 1041–1048, May 2021.

[20] L. A. W. Gemein, R. T. Schirrmeister, P. Chrabaszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, and T. Ball, "Machine-learning-based diagnostics of EEG pathology," *NeuroImage*, vol. 220, Oct. 2020, Art. no. 117021.

[21] P. Gloor, "Hans Berger on electroencephalography," *Amer. J. EEG Technol.*, vol. 9, no. 1, pp. 1–8, 1969.

[22] I. I. Goncharova, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "EMG contamination of EEG: Spectral and topographical characteristics," *Clin. Neurophysiol.*, vol. 114, no. 9, pp. 1580–1593, 2003.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.

[25] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[26] M. P. van den Heuvel and H. E. H. Pol, "Exploring the brain network: A review on resting-state FMRI functional connectivity," *Eur. Neuropsychopharmacol.*, vol. 20, no. 8, pp. 519–534, 2010, doi: 10.1016/j.euroneuro.2010.03.008.

[27] B. Karlık and B. Hayta, "Comparison machine learning algorithms for recognition of epileptic seizures in EEG," in *Proc. IWBBIO*, 2014, pp. 1–12.

[28] S. Kaur, H. Aggarwal, and R. Rani, "Diagnosis of Parkinson's disease using deep CNN with transfer learning and data augmentation," *Multimedia Tools Appl.*, vol. 80, no. 7, pp. 10113–10139, Mar. 2021.

[29] A. Khanna, A. Pascual-Leone, C. M. Michel, and F. Farzan, "Microstates in resting-state EEG: Current status and future directions," *Neurosci. Biobehavioral Rev.*, vol. 49, pp. 105–113, Feb. 2015, doi: 10.1016/j.neubiorev.2014.12.010.

[30] J. L. Lanciego, N. Luquin, and J. A. Obeso, "Functional neuroanatomy of the basal ganglia," *Cold Spring Harbor Perspect. Med.*, vol. 2, no. 12, Dec. 2012, Art. no. a009621, doi: 10.1101/cshperspect.a009621.

[31] Y. Li, Y. Chang, and H. Lin, "Statistical machine learning in brain state classification using EEG data," *Open J. Big Data*, vol. 1, no. 2, pp. 19–33, 2015.

[32] Q. Li, S. Shah, X. Liu, and A. Nourbakhsh, "Data sets: Word embeddings learned from tweets and general data," in *Proc. 11th Int. AAAI Conf. Web Social Media*, May 2017, pp. 1–9.

[33] Y. Lee, J.-M. Kwon, Y. Lee, H. Park, H. Cho, and J. Park, "Deep learning in the medical domain: Predicting cardiac arrest using deep learning," *Acute Crit. Care*, vol. 33, no. 3, pp. 117–120, Aug. 2018.

[34] A. M. Maitin, J. P. Romero Muñoz, and Á. J. García-Tejedor, "Survey of machine learning techniques in the analysis of EEG signals for Parkinson's disease: A systematic review," *Appl. Sci.*, vol. 12, no. 14, p. 6967, Jul. 2022.

[35] C. Marras, J. C. Beck, J. H. Bower, E. Roberts, B. Ritz, G. W. Ross, and C. M. Tanner, "Prevalence of PD across north America," *NPJ Parkinson's Disease*, vol. 4, no. 1, pp. 1–7, 2018.

[36] D. S. Marín, H. Carmona, M. Ibarra, M. Gámez, "Enfermedad de Parkinson: Fisiopatología, diagnóstico y tratamiento," *Revista de la Universidad Industrial de Santander*, vol. 50, no. 1, pp. 79–92, 2018, doi: 10.18273/revsal.v50n1-2018008.

[37] F. Muslim, A. Purwarianti, and F. Z. Ruskanda, "Cost-sensitive learning and ensemble BERT for identifying and categorizing offensive language in social media," in *Proc. 8th Int. Conf. Adv. Inform., Concepts, Theory Appl. (ICAICTA)*, Sep. 2021, pp. 1–6.

[38] P. Nagabushanam, S. T. George, and S. Radha, "EEG signal classification using LSTM and improved neural network algorithms," *Soft Comput.*, vol. 24, pp. 1–23, Nov. 2019.

[39] S. L. Oh, Y. Hagiwara, U. Raghavendra, R. Yuvaraj, N. Arunkumar, M. Murugappan, and U. R. Acharya, "A deep learning approach for PD diagnosis from EEG signals," *Neural Comput. Appl.*, vol. 32, pp. 1–7, Aug. 2018.

[40] D. Pawuś and S. Paszkiel, "The application of integration of EEG signals for authorial classification algorithms in implementation for a mobile robot control using movement imagery—Pilot study," *Appl. Sci.*, vol. 12, no. 4, p. 2161, Feb. 2022.

[41] C. R. Pereira, S. A. T. Weber, C. Hook, G. H. Rosa, and J. P. Papa, "Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics," in *Proc. 29th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2016, pp. 340–346.

[42] V. Podgorelec, "Analyzing EEG signals with machine learning for diagnosing Alzheimer's disease," *Electron. Electr. Eng.*, vol. 18, no. 8, pp. 61–64, Oct. 2012.

[43] R. B. Postuma, D. Berg, M. Stern, W. Poewe, C. W. Olanow, W. Oertel, J. Obeso, K. Marek, I. Litvan, A. E. Lang, G. Halliday, C. G. Goetz, T. Gasser, B. Dubois, P. Chan, B. R. Bloem, C. H. Adler, and G. Deuschl, "MDS clinical diagnostic criteria for Parkinson's disease: MDS-PD clinical diagnostic criteria," *Movement Disorders*, vol. 30, no. 12, pp. 1591–1601, Oct. 2015, doi: 10.1002/mds.26424.

[44] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino, "Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis," *Neurology*, vol. 86, no. 6, pp. 566–576, 2016.

[45] G. Ruffini, D. Ibañez, M. Castellano, L. Dubreuil-Vall, A. Soria-Frisch, R. Postuma, J.-F. Gagnon, and J. Montplaisir, "Deep learning with EEG spectrograms in rapid eye movement behavior disorder," *Frontiers Neurol.*, vol. 10, p. 806, Jul. 2019.

[46] J. S. Saavedra Moreno, P. A. Millán, and O. F. Buriticá Henao, "Introducción, epidemiología y diagnóstico de la enfermedad de Parkinson," *Acta Neurológica Colombiana*, vol. 35, no. 3, pp. 2–10, Aug. 2019, doi: 10.22379/24224022244.

[47] 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.

[48] J. I. Serrano, M. D. del Castillo, V. Cortés, N. Mendes, A. Arroyo, J. Andreo, E. Rocon, M. del Valle, J. Herreros, and J. P. Romero, "EEG microstates change in response to increase in dopaminergic stimulation in typical Parkinson's disease patients," *Frontiers Neurosci.*, vol. 12, p. 714, Oct. 2018.

[49] A. H. Shahid and M. P. Singh, "A deep learning approach for prediction of PD progression," *Biomed. Eng. Lett.*, vol. 10, no. 2, pp. 227–239, 2020.

[50] X. Shi, T. Wang, L. Wang, H. Liu, and N. Yan, "Hybrid convolutional recurrent neural networks outperform CNN and RNN in task-state EEG detection for Parkinson's disease," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 939–944.

[51] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6809–6817.

[52] E. Tolosa, M. J. Marti, F. Valldeoriola, and J. L. Molinuevo, "History of Levodopa and dopamine agonists in Parkinson's disease treatment," *Neurology*, vol. 50, no. 6, pp. S2–S10, Jun. 1998, doi: 10.1212/wnl.50.6_suppl_6.s2.

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

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[55] W. F. Velicer and P. C. Molenaar, "Time series analysis for psychological research," in *Handbook of Psychology*, I. Weiner, J. A. Schinka, and W. F. Velicer, Eds., 2nd ed. 2012, doi: 10.1002/9781118133880.hop202022.

[56] C. S. von Bartheld, J. Bahney, and S. Herculano-Houzel, "The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting," *J. Comparative Neurol.*, vol. 524, no. 18, pp. 3865–3895, Dec. 2016.

[57] W. Wang, J. Lee, F. Harrou, and Y. Sun, "Early detection of PD using deep learning and machine learning," *IEEE Access*, vol. 8, pp. 147635–147646, 2020.

[58] S. Waninger, C. Berka, S. K. Marija, S. Korszen, P. D. Mozley, C. Henchcliffe, and A. Verma, "Neurophysiological biomarkers of Parkinson's disease," *J. Parkinson's Disease*, vol. 10, no. 2, pp. 471–480, 2020.

[59] T. Wilaiprasitporn, A. Ditthapron, K. Matchaparn, T. Tongbuasirilai, N. Banluesombatkul, and E. Chuangsuwanich, "Affective EEG-based person identification using the deep learning approach," *IEEE Trans. Cogn. Developmental Syst.*, vol. 12, no. 3, pp. 486–496, Sep. 2019.

[60] H. Bin Yoo, E. O. D. L. Concha, D. De Ridder, B. A. Pickut, and S. Vanneste, "The functional alterations in top-down attention streams of Parkinson's disease measured by EEG," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, Dec. 2018.

[61] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, and A. Etemad, "Classification of hand movements from EEG using a deep attention-based LSTM network," *IEEE Sensors J.*, vol. 20, no. 6, pp. 3113–3122, Mar. 2020.

[62] Y. Zhang, J. Chen, J. H. Tan, Y. Chen, Y. Chen, D. Li, L. Yang, J. Su, X. Huang, and W. Che, "An investigation of deep learning models for EEG-based emotion recognition," *Frontiers Neurosci.*, vol. 14, Dec. 2020, Art. no. 622759.

ALBERTO NOGALES received the master's degree in software engineer and artificial intelligence, and the Ph.D. degree focused in the field of semantic web and social network analysis from the University of Alcalá, Spain. He has experience in several European research projects and publications in conferences and journals from JCR. His education has been completed working in Technische Universität Wien (Austria), Tallinna Tehnikaülikool (Estonia) and University of Málaga (Spain). Since November 2017, he has been works as a Postdoctoral Researcher and a Lecturer at Universidad Francisco de Victoria and CEIEC Research Center. His research interests include deep learning and its use in signal processing and predictive models.

ÁLVARO J. GARCÍA-TEJEDOR received the Ph.D. degree in biochemistry. He has been a Lecturer at UCM, UC3M, and Universidad Francisco de Vitoria, where he is currently an Associate Professor in artificial intelligence. He has been involved in more than 20 research and development projects mainly in AI. As a researcher, he started doing mathematical modeling of biological systems. Since 2007, he heads CEIEC, the Research Institute of the UFV for technological innovation with a social accent. He has produced several serious games for transmission of educational and cultural content and integration of people with intellectual disabilities. His current research interests include neural models, deep learning techniques, and their applications in several fields.

ANA M. MAITÍN received the B.Sc. degree in physics, with a specialization in the theoretical branch, and the M.S.S. degree in biomedical physics from the Complutense University of Madrid, Spain. She has worked in different projects within the physical and medical fields, such us the modeling of the cellular mechanics at the Complutense University of Madrid, and the statistical and non-linear analysis of electroencephalograms at the Francisco de Vitoria University (Spain). She is currently completing her Ph.D. degree in biotechnology, medicine and bio-sanitary sciences at the Francisco de Vitoria University as the recipient of the Ph.D. Fellowship. Moreover, she is conducting her research work at the Center for Studies and Innovation in Knowledge Management (CEIEC), where she has focused on the development of machine learning techniques, especially in deep learning techniques, and their application to different problems within medicine.

ANTONIO PÉREZ-MORALES was born in Cáceres, in 1998. He received the degree in software engineering from the Francisco de Vitoria University, Madrid, where he was awarded Optimus for his performance during his degree. During those years, he worked as an Audit Intern at Mapfre and as an AI Intern Researcher at CEIEC, where he developed his degree project, on which he leveraged the power of BERT for medical applications. He is currently works as the Head of Integrations at Valispace, a Lisbon based startup that develops a software as a service solution to agilise hardware development for engineering companies.

MARÍA DOLORES DEL CASTILLO received the Ph.D. degree in physics. She started a Research career at the Centre for Automation and Robotics (CAR) CSIC–UPM. Her research interests include artificial intelligence and machine learning disciplines, cognitive science, and brain–computer interfaces. Application fields of her research have evolved from automating industrial process and generating robots' behavior, passing by knowledge discovery in exhaustive data volumes to more recently computational linguistics, and cognitive modeling and technological platforms for human–machine interfacing.

JUAN PABLO ROMERO received the master's degree in neurobiochemistry, biotechnology, and neuropsychology, and the Ph.D. degree in neurodegenerative diseases mortality from the Complutense University of Madrid. He is a Neurologist specializing in movement disorders and brain damage rehabilitation. He is currently a Professor of neurology and neuroanatomy at the Francisco de Vitoria University, Madrid. He is the main Researcher and the Chief of the Neurorehabilitation of Movement Disorders and Brain Damage Research Group with several research lines funded by national and international grants. His research interests include the non-invasive neuromodulation applied to the rehabilitation of cognitive and motor functions on Parkinson's Disease and brain damage and biosignal processing for identifying disease progression markers.