Computer Methods and Programs in Biomedicine 246 (2024) 108048

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Unraveling motor imagery brain patterns using explainable artificial intelligence based on Shapley values

Sergio Pérez-Velasco [a,b,*], Diego Marcos-Martínez [a,b], Eduardo Santamaría-Vázquez [a,b], Víctor Martínez-Cagigal [a,b], Selene Moreno-Calderón [a], Roberto Hornero [a,b]

[a] Biomedical Engineering Group, E.T.S Ingenieros de Telecomunicación, University of Valladolid, Paseo de Belén 15, Valladolid, 47011, Spain
[b] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Spain

ARTICLE INFO

Keywords:
Brain-computer interface (BCI)
Motor imagery (MI)
Explainable artificial intelligence (XAI)
Shapley additive explanations (SHAP)
Deep learning (DL)
Sensorimotor rhythms (SMR)

ABSTRACT

Background and objective. Motor imagery (MI) based brain-computer interfaces (BCIs) are widely used in rehabilitation due to the close relationship that exists between MI and motor execution (ME). However, the underlying brain mechanisms of MI remain not well understood. Most MI-BCIs use the sensorimotor rhythms elicited in the primary motor cortex (M1) and somatosensory cortex (S1), which consist of an event-related desynchronization followed by an event-related synchronization. Consequently, this has resulted in systems that only record signals around M1 and S1. However, MI could involve a more complex network including sensory, association, and motor areas. In this study, we hypothesize that the superior accuracies achieved by new deep learning (DL) models applied to MI decoding rely on focusing on a broader MI activation of the brain. Parallel to the success of DL, the field of explainable artificial intelligence (XAI) has seen continuous development to provide explanations for DL networks success. The goal of this study is to use XAI in combination with DL to extract information about MI brain activation patterns from non-invasive electroencephalography (EEG) signals. *Methods.* We applied an adaptation of Shapley additive explanations (SHAP) to *EEGSym*, a state-of-the-art DL network with exceptional transfer learning capabilities for inter-subject MI classification. We obtained the SHAP values from two public databases comprising 171 users generating left and right hand MI instances with and without real-time feedback. *Results.* We found that *EEGSym* based most of its prediction on the signal of the frontal electrodes, i.e. F7 and F8, and on the first 1500 ms of the analyzed imagination period. We also found that MI involves a broad network not only based on M1 and S1, but also on the prefrontal cortex (PFC) and the posterior parietal cortex (PPC). We further applied this knowledge to select a 8-electrode configuration that reached inter-subject accuracies of 86.5% ± 10.6% on the Physionet dataset and 88.7% ± 7.0% on the Carnegie Mellon University's dataset. *Conclusion.* Our results demonstrate the potential of combining DL and SHAP-based XAI to unravel the brain network involved in producing MI. Furthermore, SHAP values can optimize the requirements for out-of-laboratory BCI applications involving real users.

1. Introduction

Brain-computer interfaces (BCIs) create an alternative pathway to connect users' brains with the environment [1]. A BCI system operates as a closed-loop system, composed of three stages: recording, processing, and feedback. For the recording stage, electroencephalography (EEG) is one of the most commonly used and affordable techniques in BCI systems. EEG has the advantages of being non-invasive, highly portable, and offering excellent temporal resolution [2]. An EEG-based BCI records the electrical brain activity through the placement of electrodes on the user's scalp. In the processing stage, a BCI processes these EEG recordings to discern the user's intentions [1]. Finally, feedback from the BCI's real-time processing is offered to the user. The feedback can take various forms, including visual feedback on a computer screen or prosthetic limb movement, among others [3]. Nevertheless, obtaining information from the EEG is not trivial due to its low spatial resolution and low signal-to-noise ratio (SNR). To overcome these challenges, BCIs rely on different paradigms that can create discernible

[*] Corresponding author at: Biomedical Engineering Group, E.T.S Ingenieros de Telecomunicación, University of Valladolid, Paseo de Belén 15, Valladolid, 47011, Spain.

E-mail address: sergio.perezv@uva.es (S. Pérez-Velasco).

https://doi.org/10.1016/j.cmpb.2024.108048
Received 11 April 2023; Received in revised form 22 January 2024; Accepted 23 January 2024
Available online 30 January 2024

patterns in the EEG. Among these paradigms, motor imagery (MI) has gained considerable attention due to its relevance for motor rehabilitation [4–6].

MI-based BCIs activate the primary motor cortex (M1) and related motor areas, similar to motor execution (ME) [6]. Due to this similarity, it has been shown that targeted treatments based on a closed-loop MI-based BCI with functional electrical stimulation feedback promotes brain plasticity and improves ME in stroke patients [6]. Previous studies have revealed that ME involves a complex network of sensory, association, and motor areas that coordinate in a hierarchical manner to produce normal movement [7]. The posterior parietal cortex (PPC), the prefrontal cortex (PFC), and the premotor cortex (PM) are involved in planning and preparing movement based on visual information related to both the movement goal and the limb state. The M1 and somatosensory cortex (S1) are involved in executing movement. The PFC was reported to activate 150 ms earlier than the PM, indicating its higher placement in the temporal hierarchy [7]. In visually-guided movements, the PPC associates visual information related to both the movement goal and to the state of the limb and its place in the temporal hierarchy seems to be variable. However, less is known about how these areas interact during MI-based BCI training. A symptom of this lack of knowledge is that MI-based BCIs usually restrict the feedback to only the EEG signal recorded in the M1 and S1 areas related to sensorimotor rhythms (SMR) [6,8,9]. Moreover, the pioneering studies that examined the behavior of the brain activity in ME have used electrocorticography (ECoG), or brain-penetrating microelectrodes [7], which are not feasible for most widespread neurorehabilitation applications due to their invasiveness. Therefore, there is a need for methods that can explore both spatial and temporal aspects of MI in non-invasive recordings of brain activity during MI-based BCI training.

Despite using the MI paradigm to enhance the SNR, decoding the intentions of the user from the EEG is still very challenging [1]. Furthermore, MI-based BCI's decoding precision is lower than that of other paradigms. For instance, event related potentials-based BCIs demonstrate greater than 90% accuracy [10] [11], while code-modulated evoked potentials-based BCIs exhibit more than 95% accuracy [12]. Traditionally, machine learning (ML) approaches have been used to decode users' intentions in MI-based BCIs. In an ML pipeline, the processing stage is composed of three different sub-stages: preprocessing, feature extraction with an optional feature selection, and classification [1,13–15]. Though this process has yielded suitable performances and has been in continuous improvement, its accuracy has been recently surpassed with the use of deep learning (DL) [10,16–19]. The processing pipeline is replaced by an end-to-end neural network that can classify the user's intentions from the pre-processed EEG signal. DL networks not only outperform their ML counterparts but also have the advantage of learning complex patterns from EEG data without the need for manual feature engineering. At the same time, DL networks can solve the shortcomings of ML in overcoming inter-subject and inter-session variability of EEG data [18].

While DL networks have improved performance, they are often considered a black box, making it difficult to retrieve information about the classification process [20]. Unlike multi-stage processing pipelines in ML, a single DL network responsible for all sub-stages can be less interpretable, and the patterns extracted by DL without human intervention might remain undiscovered. A trade-off between explainability and accuracy exists [20]. The field of explainable artificial intelligence (XAI) tries to overcome this disadvantage. XAI techniques can be categorized by the method used to extract information into (1) backpropagation and (2) perturbation [21]. Backpropagation-based XAI propagates the importance of the output backwards to the input, while perturbation-based XAI alters the input to observe the impact on the following layers of the model [21]. There have been previous works that have already adapted backpropagation and perturbation-based XAI for EEG data. For backpropagation-based XAI, an adaptation of layerwise relevance propagation (LRP) can be found [22]. These authors found that individual trials were classified by giving relevance to M1 and S1 regions 1000 to 3000 ms after MI onset and without the presence of real-time feedback [22]. Conversely, for perturbation-based XAI, there has been an adaptation of occlusion sensitivity analysis [23], and a procedure called *easyPEASI* based solely on perturbations to the input data [24]. Ieracitano et al. [23] found that the relevant brain sources used by their DL approach to distinguish between previous hand movement and resting state were located in the temporal lobe and in the central area of M1. Meanwhile, Nahmias et al. [24] found the most relevant frequency bands for pathology detection.

To approximate the different XAI approaches, the Shapley additive explanations (SHAP) were proposed [25]. SHAP values unified the most well-known XAI approaches by introducing a theoretical framework based on the properties: local accuracy, missingness, and consistency. They showed how previous approaches satisfied one or more of these concepts, while providing adaptations so that they could satisfy all of them. The modified explanation methods that comply with these three properties guarantee the uniqueness of the solution and a meaningful explanation of the model's predictions [25]. While an application that adapts SHAP values to EEG data has already been proposed [26], it was created for tree-based classifiers based on power spectral density features which cannot discern as complex patterns as a DL network can in EEG data. To the best of our knowledge, none of the previous studies have applied SHAP values to DL networks in an EEG context. An analysis based on SHAP values to analyze the decision process of DL models in MI classification could reveal high-level features and neural patterns that remain undiscovered in MI. Our study has important implications for both theory and practice of MI-BCI: it can enhance our understanding of how and when MI activates different brain regions, and it can improve the usability of MI-BCI by reducing the number of electrodes required while maintaining classification accuracy.

The main goal of this study is to analyze the relevance of the features for a DL model applied to MI tasks. Using SHAP values for the first time in this context allows us to gain insights into the brain's network involved in MI. Additionally, a second objective is to assist in channel selection, further improving the ease-of-use of EEG-based BCIs by reducing the number of needed electrodes, taking into account the information revealed by SHAP values. To achieve these objectives, we use two different public databases of non-invasive EEG recordings from healthy users performing MI tasks: one consisting of 109 users without feedback [27], and another one with 62 users with real-time visual feedback [9]. We apply a DL network called *EEGSym* [18] with state-of-the-art performance on inter-subject classification. *EEGSym* is fine-tuned on these datasets to generate SHAP values that indicate how each EEG channel contributes to the classification output.

2. Methods

2.1. Datasets and preprocessing

Two public datasets will be used to extract information with the SHAP-based XAI method. Both databases include trials with continuous imagination of sequentially opening and closing either the left or right hand. However, they present differences such as the presence of feedback or the sampling frequency. The protocol followed by both datasets is presented in Fig. 1. The feedback sessions had a first cue indicating the imagination to perform and provided real-time feedback 2 seconds after this onset. On the one hand, the Physionet dataset [27] is composed of trials pertaining to 109 healthy users without feedback, with only one session of 42-46 trials. The imagination period of each trial is constant and spans 3 seconds. EEG was recorded using 64 electrodes with a sampling frequency of 160 Hz for 105 users and 128 Hz for four of them. On the other hand, the Stieger2021 dataset [9] recorded the EEG of 62 healthy users during 7 to 11 sessions with feedback. In each session, 450 trials with 62 electrodes and a sampling frequency of 1000 Hz were recorded. The trial duration was between 4 and 10

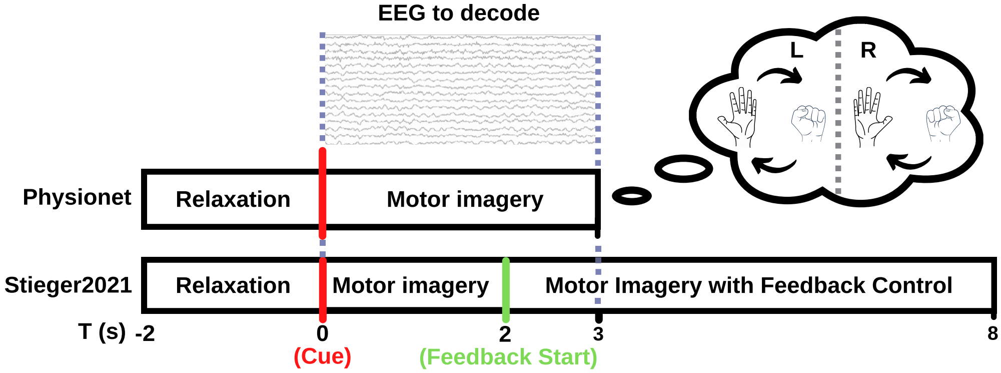

Fig. 1. Schematic of trials present in the public datasets, Physionet [27] and Stieger2021 [9]. The feature window selected for classification, and the motor imagery period.

seconds, eventually being reduced if the user reached the target during feedback.

To both datasets, we apply the following preprocessing [18]: (1) from both datasets, we extract the electrodes F7, F3, T7, C3, P7, P3, O1, Pz, Cz, F8, F4, T8, C4, P8, P4, and O2, (2) we apply a notch filter on the Physionet dataset [27] since it does not have the power line signal removed by hardware, (3) we perform common average reference (CAR) spatial filtering to these 16 electrodes, (4) we do a resampling to 128 Hz to homogenize both datasets, (5) we extract the trials with a time window length of 3 seconds after the onset, and (6) we apply channel-wise z-score standardization to each trial.

2.2. DL architecture

The open implementation of *EEGSym* [18] for 16 channels will be the model explained using the SHAP-based XAI method. *EEGSym* is a novel CNN for inter-subject MI classification that was presented in our previous study [18]. The implementation of *EEGSym* takes advantage of recent techniques developed for DL: residual connections, data augmentation, inter-subject transfer learning, and a siamese-network design that exploits the symmetry of the brain through the mid-sagittal plane. This CNN reached significantly higher accuracy in a binary MI inter-subject classification task than four previous CNNs designed for EEG classification, i.e., ShallowConvNet and DeepConvNet [16], EEG-Net [17], and EEG-Inception [10]. The combination of *EEGSym* and the DL techniques applied to it offered new state-of-the-art results for inter-subject MI classification.

This particular DL network is selected for two reasons. First, it achieved superior performance on inter-subject classification tasks, thanks to its training on the largest and most diverse set of users and MI strategies known to us. Second, the network was designed specifically for inter-subject classification, so it is expected to identify and prioritize patterns that are common across users. These properties should make the CNN particularly adept at extracting generalizable information about MI. Thus, our conclusions drawn from it are expected to be broadly applicable. Since the generalizability and stability of this approach to other networks are desirable, we also include results with the EEG-Inception and EEGNet networks in the supplementary material [10,17].

2.3. SHAP values for MI

There have been multiple approaches to explaining deep networks through the use of additive feature attribution methods [28,29,21]. To unify all these emerging techniques, SHAP values were proposed [25]. SHAP values are a game-theoretic method that studies how different players cooperate or compete with each other [25]. Applied to an EEG signal, the different parts of the EEG signal are the players, and the prediction is the outcome of their cooperation. The SHAP values of an EEG signal represent the contribution of each data instance to the prediction. Furthermore, SHAP values satisfy three properties that ensure they provide meaningful explanations of a model's predictions and can be used to understand the importance of different features in those predictions: local accuracy, missingness, and consistency [25]. Local accuracy ensures that the sum of SHAP values equals the difference between the prediction and the expected prediction. Satisfying missingness means that if a feature value is missing, then its SHAP value is zero, and thus SHAP values will not assign contributions to features that are not present. On the other hand, consistency ensures that if a feature increases its contribution to the prediction when other features are added or removed, then its SHAP value should not decrease [25].

While the exact computation of SHAP values is challenging in DL models, the SHAP Python package implements approximation methods based on previous additive feature attribution methods that calculate expected SHAP values [25]. We select the method called 'GradientExplainer' from the SHAP package, an adaptation that merges the concepts of integrated gradients [30], SHAP [25], and SmoothGrad [31]. This method extends the integrated gradients method, which is already an extension of Shapley values to infinite player games (the Aumann-Shappley values) [30]. 'GradientExplainer' method first defines a baseline input signal that is used as a reference point for comparison. This baseline signal is typically chosen as an all-zero signal [30]. This method computes the gradients (i.e., change in the predicted outcome) of the model's prediction with respect to the input signal at each point along the path from the baseline to the actual input signal. These gradients represent how much the model's prediction changes as each feature of the input signal changes along the path. Then, the SHAP values are computed by numerically approximating the area under the curve of the gradients along the path from the baseline to the actual input signal. This integral represents the total contribution of each feature to the model's prediction for that input signal [30]. The resulting SHAP values are a measure of the relative importance of each feature in the prediction. Positive SHAP values indicate data points that positively influence the classification, whereas negative values indicate points that are detrimental to the correct prediction. Conversely, values close to zero indicate features that do not contribute to any class prediction [25].

One of the key aspects to approximate integrated gradients [30] into SHAP values by ensuring local accuracy is the use of a background dataset instead of a single reference. This feature adds computational complexity but can be circumvented in our application to EEG data. Since the preprocessing of our signal will set the mean of each channel data to approximately 0, the average of infinite basal EEG signals with 0 mean will also have 0 mean according to the central limit theorem. This means that in our application the background dataset can be replaced by a matrix in which all points are 0 s in all channels and time instances. Furthermore, this reference will be a better approximation for calculating SHAP values than any background dataset selected with a finite number of examples. Meanwhile, integrated gradients inherently ensure the other two essential properties of missingness and consistency by using gradient computations: irrelevant features are assigned a SHAP value of zero, while features with higher gradients correspondingly receive increased SHAP values.

At this point, we have defined all the elements needed to obtain the feature attribution maps from each trial: the approximate method 'GradientExplainer', the MI data to evaluate, and the CNN model from which we will obtain the SHAP values. Regardless, what we present and analyze in this work is the averaged SHAP values obtained from each individual trial across subjects and datasets. First, we obtain the SHAP values from the prediction of the model applied to an individual EEG input signal referenced to our background. To obtain these values for each user, we trained EEGSym following a leave-one-subject-out (LOSO) procedure, as in the original paper [18]. For each subject's trials, we use the fine-tuned weights to the dataset that have not seen any trial of the current subject. We will take into consideration only correctly classified trials, as these are expected to embody the inter-subject common brain patterns associated with the MI task that we seek to unveil. We average the SHAP values maps obtained for left and right classes per user, and then we average across subjects. By averaging SHAP values first within subjects and then across different subjects, we create a more generalizable representation of the significant features implicated in the MI tasks. This aggregated analysis provides a robust interpretive framework that transcends individual variances, thereby offering insights into common neurophysiological mechanisms that underlie the MI task. This process ends with 4 feature attribution maps, one for each class in both dataset conditions: with and without feedback. In these feature attribution maps, the time series is represented on the horizontal axis while the channels are on the vertical axis. Accordingly, there is a direct correlation between SHAP values and the input signal's channels or time segments. By performing an axis-wise aggregation of the SHAP values, we can investigate the contribution that each region of the EEG has in predicting the class under analysis [25]. We can determine the relative importance of each channel or period of time for MI classification. However, it is important to note that a positive influence in one region does not necessarily mean the presence of a pattern, but rather the absence of any pattern that would detract from the classification. Therefore, to gain a comprehensive understanding of the classification process, we present feature attribution maps for both classes and examine both together.

An open-source adaptation of this SHAP-based XAI method to EEG signals can be found as part of the kernel functions within MEDUSA©, a Python-based software ecosystem to accelerate BCI and cognitive neuroscience research [32].

2.4. Channel selection based on SHAP values

To validate the usefulness of the information extracted from the adapted method for EEG data, we propose a channel selection based on SHAP values. Aggregating channel-wise the feature attribution maps generated with the 16-electrode configuration in the Physionet dataset [27] and the Stieger2021 dataset [9] will help us identify the channels that have the least contribution to the final classification in the *EEGSym* architecture [18]. To make a BCI system more cost-effective and quicker to install for out-of-laboratory use, we will identify the electrodes that have the strongest influence on correctly predicting MI. This selection will be based on the analysis of the aggregated SHAP values. Instead of assessing electrodes individually, we will examine them in pairs. Our comparison will involve the contributions of seven electrode pairs located on the left and right sides of the scalp (i.e., F7 and F8, F3 and F4, T7 and T8, C3 and C4, P7 and P8, P3 and P4, O1 and O2), as well as the central electrode pair (i.e., Cz and Pz). Given that 14 out of 16 electrodes are symmetrically placed, it would be incongruous to include only one electrode from each pair, even though it falls on the midsagittal plane of the scalp. Therefore, our initial strategy of evaluating electrodes in pairs logically extends to the inclusion of Cz and Pz as an additional comparative pair. The top four electrode pairs with the most significant contributions to classification will be selected. The accuracy of this new configuration will be compared to the 8 electrodes used in *EEGSym's* publication [18]. For comparative analysis, we will also derive an 8-electrode configuration based on SHAP values. Additionally, we will investigate how the significance of channels to the model's decision-making process influences the model's classification accuracy on both datasets. This will be done by selectively adding channel pairs according to their respective contributions.

3. Results

Figs. 2 and 3 show the feature attribution maps of class-averaged MI trials for each dataset as generated from the EEGSym algorithm. Supplementary materials provide analogous attribution maps obtained using the EEG-Inception and EEGNet models for comparison. SHAP values shown in these figures are obtained from the 3 seconds of signal after the onset, and the 16 channels used in the prediction. Positive SHAP values are colored red, and negative values are colored blue. In addition to the direct feature attribution map, there is a channel-wise average indicating the percentage-based contribution to the prediction. In the temporal axis, there is also an average of the SHAP value across all channels. Both temporal and spatial averages show the influence that different time segments and channels have on the classification of that class.

3.1. Physionet

Fig. 2 shows the SHAP values obtained for left and right hand MI events from the Physionet dataset [27], a dataset with no feedback. Only the SHAP values of correctly classified trials are averaged as mentioned in 2.3. *EEGSym* reaches an inter-subject accuracy of 88.6% ± 9.0% in this dataset, with 108 out of the 109 users attaining what is considered as BCI control (i.e., ≥70% accuracy) [18], indicating that almost every user provides useful information about their MI process.

For this binary classification, the first 1000 ms of EEG signal following the onset contributes the most to the prediction. More specifically, the region between 200 and 400 ms shows the strongest influence, which correlates with the two-choice reaction time distribution in humans [33]. Of note, the last 2000 ms of the signal are less relevant for the prediction in the temporal aggregated contribution, despite not being unimportant for some electrodes. Analyzing individually each channel, the attribution maps show that the first 1000 ms of the signal is also the most relevant for almost every channel. The electrodes in the PPC region, i.e., P3, P4, P7, P8, O1, and O2 showed stronger contributions between 100 and 300 ms after the cue. They were followed by the contribution of electrodes in the PFC region, i.e., F7 and F8, between 200 and 500 ms. Of note, the ipsilateral PFC seems to have a behavior that is noticeable earlier than the contralateral, starting 100 ms after the onset. The relevance of the M1 and S1 signal detected in C3 and C4 electrodes starts around 300 ms after onset and remains noticeable until 1500 ms. T7 and T8 electrodes show similar relevance to C3 and C4, and may exhibit earlier contributions in some cases.

Each channel's percentage value shows its relevance in predicting a specific class, thereby giving an idea of how important each channel is in terms of influencing the model's decisions. The percentage value is calculated by dividing the sum of the absolute SHAP values for that channel by the total sum of absolute SHAP values across all channels and time points. This information reveals that *EEGSym* finds the signal pertaining to the PFC on F7 and F8 channels as the most informative for classifying MI trials. Afterwards, we find the M1 and S1 signal registered on C3 and C4 electrodes. These four channels are responsible for almost 50% of the prediction. From the remaining 12 channels, we can highlight the importance given to the pair of T7 and T8, and to the pair of central electrodes Cz and Pz. For both MI classes, *EEGSym* has a

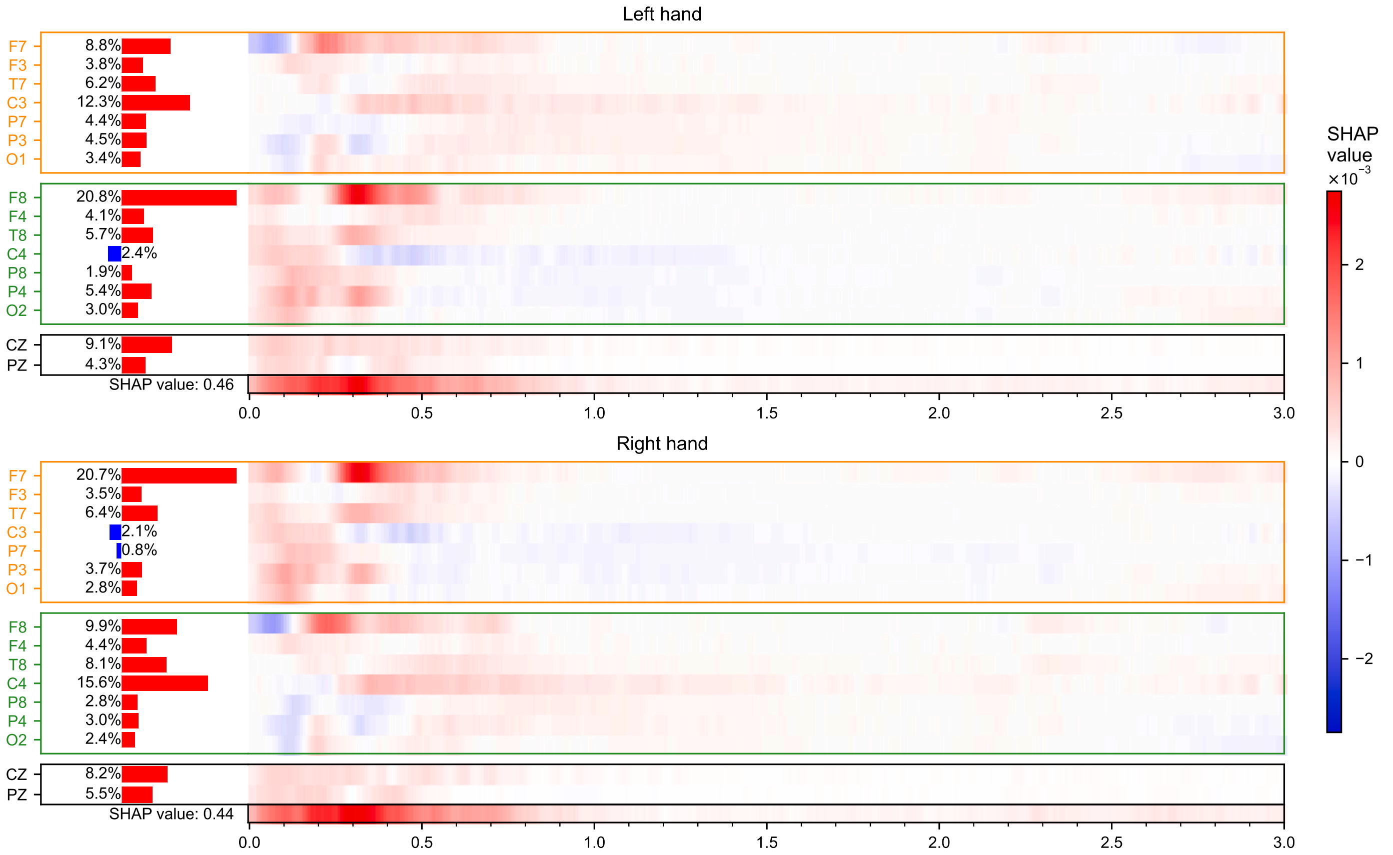

Fig. 2. Feature attribution map with SHAP values for MI events in the Physionet dataset [27] without feedback. In the vertical axis, 16 channels from the 10-10 system are presented. Channels corresponding to the left hemisphere of the scalp are marked in orange, those corresponding to the right hemisphere in green, and the central electrodes in black. The percentages indicate the relative importance of that electrode signal to the prediction providing insight into the significance of each channel in guiding the model's decisions. The horizontal axis shows the three seconds after the cue corresponding to the MI event. The positive SHAP values are marked in red. Meanwhile, blue color marks negative SHAP values which contribute to not predicting the target class. A red tempo-spatial region contributes to correctly predicting the target class.

symmetric attribution of the channels. While the signal from F8 and C3 electrodes is the most significant to correctly classify left hand MI, the signal found in F7 and C4 is used for right hand MI trials.

3.2. Stieger2021

The SHAP values obtained for MI events in the presence of feedback from the Stieger2021 dataset [9] are shown in Fig. 3. The inter-subject accuracy for this dataset is 90.2% ± 6.5%, with 61 out of the 62 users reaching BCI control. In this regard, only one user may not have provided actual MI information in the averaged feature attribution maps presented. The figure includes a green mark that indicates the start of the real-time visual feedback phase 2 seconds after the cue.

Similar to the non-feedback scenario of the Physionet dataset, the most important temporal region for making a prediction is found between 200 and 400 ms after the start of the imagination period. However, in this case, there is also another significant temporal region between 2200 and 2400 ms after the onset. Of note, this temporal window coincides with 200 and 400 ms after the start of the feedback phase, so both regions can be correlated with the two-choice reaction time distribution in humans [33]. The analysis that can be made of the first 2 seconds without feedback is similar to the one extracted from the Physionet dataset, the temporal hierarchy seems to be the same. Interestingly, the SHAP values show a clearer influence of the early ipsilateral region of the PFC, around 100 ms from the onset. Moving to the 1000 ms of signal with real-time visual feedback, SHAP values reveal a decreased influence of the PFC and an increased activity of the PPC recorded by the P3, P4, O1, and O2 electrodes. This PPC activity remains significant after the initial 200 ms, exhibiting a stronger contribution during the feedback phase. Furthermore, O1 and O2 also capture information from the occipital region, which is responsible for processing visual information [34]. These electrodes appear to be relevant in ongoing MI with real-time feedback beyond the first 500 ms. Additionally, the signal of the M1 detected by the C3 and C4 electrodes, demonstrates greater significance than that of the PFC 500 ms after the start of feedback.

The channel-wise aggregation indicates that the pair of electrodes positioned in the mid-sagittal plane, i.e. Cz and Pz, have the least contribution to the predicted class during the trial. In this case the signals of the PFC registered in the frontal electrodes, i.e. F7 and F8, are some of the most important for the prediction. They are followed by the signal of the M1, i.e. C3 and C4. The channels O1, O2, P3, P4, T7, and T8 have a similar contribution. We can also observe a symmetrical disposition between left and right hand MI feature attribution maps.

3.3. SHAP based 8-electrode configuration

As explained in 2.4, we will validate the contribution maps obtained for the Physionet dataset [27] and the Stieger2021 dataset [9] by selecting an electrode configuration based on SHAP values. The new 8-electrode configuration includes the 4 electrode pairs that have the higher aggregated absolute SHAP values, which means they have the strongest contribution to the prediction. The electrode pairs with the higher contribution in the combination of the two datasets are F7 and F8, C3 and C4, T7 and T8, and P3 and P4, in that order. These 8 electrodes concentrate 72% of the contribution to the classification of both

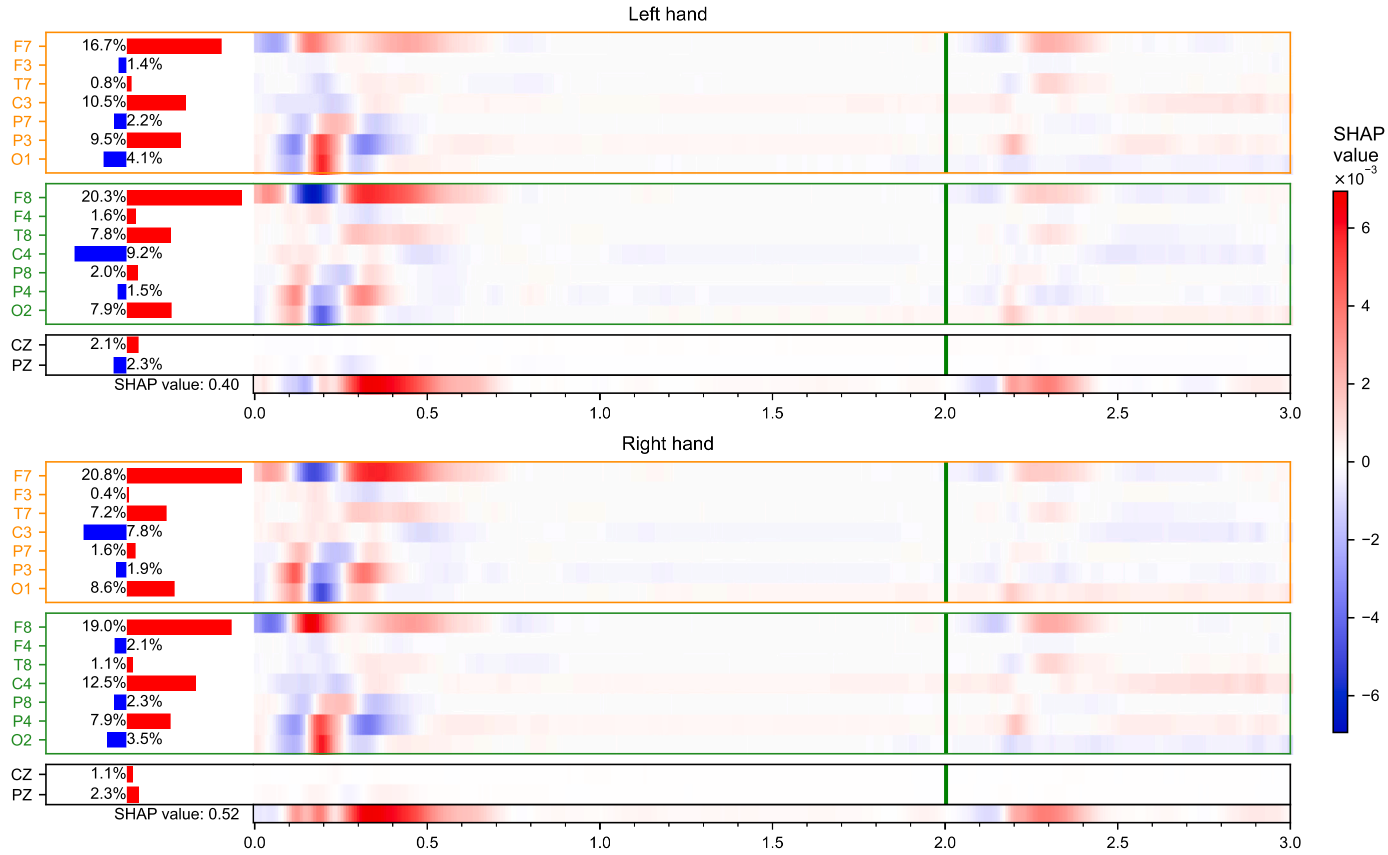

Fig. 3. Feature attribution map with SHAP values for MI events in the Stieger2021 dataset [9] with feedback. In the vertical axis, 16 channels from the 10-10 system are presented. Channels corresponding to the left hemisphere of the scalp are marked in orange, those corresponding to the right hemisphere in green, and the central electrodes in black. The percentages indicate the relative importance of that electrode signal to the prediction providing insight into the significance of each channel in guiding the model's decisions. The horizontal axis shows the three seconds after the cue corresponding to the MI event. Start of the real-time visual feedback phase 2 seconds after the onset is marked in green. The positive SHAP values are marked in red. Meanwhile, blue color marks negative SHAP values which contribute to not predicting the target class. A red tempo-spatial region contributes to correctly predicting the target class.

classes. This is the new 8-electrode configuration selected through the analysis of the SHAP values for our datasets. This SHAP based configuration will be compared with the 8-electrode configuration used in the original publication of *EEGSym* [18], which was comprised of F3, F4, C3, C4, P3, P4, Cz, and Pz. *EEGSym's* original configuration comprised only 45% of the contribution to the prediction. The original 8-electrode configuration was selected based on its balanced positioning within the 10-20 standard, coupled with its broad spatial coverage of the scalp, although it was not specifically optimized for performance. Table 1 summarizes the results of this comparison between both configurations. The new configuration has significantly higher mean accuracies (p-value < 0.01) than *EEGSym*'s original 8-electrode configuration on the Physionet dataset [27]. This comparison was assessed with the Wilcoxon signed rank test [35], correcting the false discovery rate (FDR) with the Benjamini-Hochberg approach [36]. Furthermore, the correlation between each channel pair's relevance and the model's classification accuracy is illustrated in Fig. 4. The inclusion order of channel pairs, based on their contribution across both datasets, is as follows: F7-F8 (34.3%), C3-C4 (18.1%), T7-T8 (10.8%), P3-P4 (9.4%), O1-O2 (8.9%), Cz-Pz (8.7%), F3-F4 (5.3%), and P7-P8 (4.5%).

4. Discussion

4.1. Insights on MI

Our analysis of the feature attribution maps presented in Figs. 2 and 3 can be conducted in both temporal and spatial dimensions of the input signal. The overall influence of each channel showed that the neural activity of the PFC region, registered on the frontal electrodes F7 and F8, is the most relevant for making MI predictions in the first 1000 ms. The involvement of frontal brain electrical activity in MI-based BCIs has been previously studied in the literature [37]. The PFC plays a role in planning and preparing movement based on information related to both the movement goal and limb state in ME. In this MI task, the PFC's contribution appeared to be more closely related to the imagined goal, as its involvement is more pronounced in the non-feedback scenario. SHAP values have also highlighted the contribution of the PPC region, which is more substantial in the real-time feedback scenario. The participation of PPC in visually-guided ME is known to associate visual information [7]. Correspondingly, Fig. 3 showed a higher contribution

Table 1
Comparison of binary classification performance of SHAP-based selection of 8-electrode configurations.

Study	Physionet [27] Accuracy(%)	Stieger2021 [9] Accuracy(%)
Pérez-Velasco et al. [18]	84.5 ± 9.7	88.4 ± 6.5
Present study	**86.5 ± 10.6***	**88.7 ± 7.0**

Accuracy (%): mean accuracy and standard deviation in percentage obtained between all subjects in a subject-independent scheme. The best results for each dataset are marked in **bold**. Statistical differences between the mean accuracies were assessed with Wilcoxon signed rank test, correcting the false discovery rate (FDR) with Benjamini-Hochberg approach. Obtaining significant differences is marked with *(p-value < 0.01).

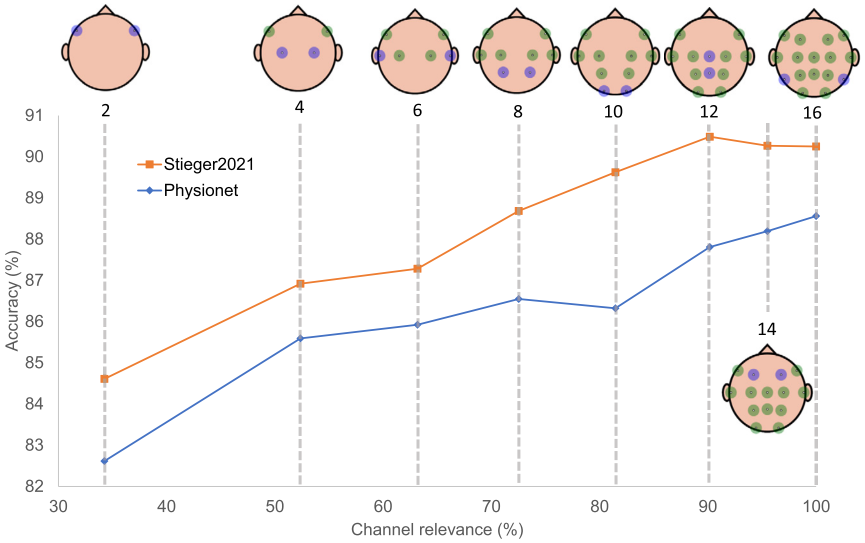

Fig. 4. Correlation between classification accuracy on the Physionet [27] and the Stieger2021 [9] datasets and the relevance of each channel, where relevance is quantified by the proportion of absolute SHAP values for a given channel relative to the total absolute SHAP values across all channels and time intervals, presented as a percentage. Each point on the graph represents a specific electrode configuration used in the prediction, with the latest pair of electrodes added to the configuration highlighted in blue, and the electrode pairs from preceding configurations shown in green.

of PPC-related electrodes in the second after visual feedback started, i.e., P3, P4, P7, P8, O1 and O2. This behavior suggests that MI involves the participation of the PPC in the presence of real-time visual feedback in the same manner as ME. Furthermore, the temporal hierarchy of the brain regions in MI, as described in sections 3.1 and 3.2, appears to be consistent with that in ME [7].

From these findings, we hypothesize that *EEGSym* is able to detect the association occurring in the PPC, the planning carried out in the PFC, and the imagination in the M1 and S1. These non-invasive recordings have allowed us to extract insights from MI using a combination of DL and the adaptation of SHAP-based XAI method. MI likely involves a complex network to generate imagined movement. This behavior should be taken into account when designing a BCI application intended for controlling an external device.

4.2. XAI based channel selection performance

We presented a SHAP-based channel selection that resulted in a reduced 8-electrode configuration from the 16 electrodes evaluated, using the methodology described in subsection 3.3. The 8-electrode configuration consisted of F7, F8, T7, T8, C3, C4, P3, and P4. The approximate physical locations are represented in Fig. 5. This general configuration reduces costs and setup duration by enabling the use of more affordable EEG caps with fewer electrodes, as opposed to the more expensive and time-consuming approach of individually selecting channels online with caps that have a higher electrode count. The XAI based channel selection suggests that this 8-electrode configuration more effectively captures neural activity during this MI task. It appears that the PFC planning activity is better recorded by the F7 and F8 electrode positions than by F3 and F4. Meanwhile, the neural activity in M1 and S1 seems to be distributed between the paired electrode positions of C3/C4, and T7/T8.

The comparison presented in Table 1 showed that this new configuration obtained better mean accuracies in both datasets than the 8-electrode configuration of the original article of *EEGSym* [18]. Although the original configuration was not specifically fine-tuned, it still achieved state-of-the-art performances in MI classification. Nevertheless, we increased the accuracy of an 8-electrode configuration from 84.5% to 86.5% in the Physionet dataset [27], and from 88.4% to 88.7% in the Stieger2021 dataset [9]. Furthermore, the visual representation presented in Fig. 4 helps to understand more in-depth how incremental additions of electrode pairs, based on their SHAP value-derived relevance, influence the overall model accuracy.

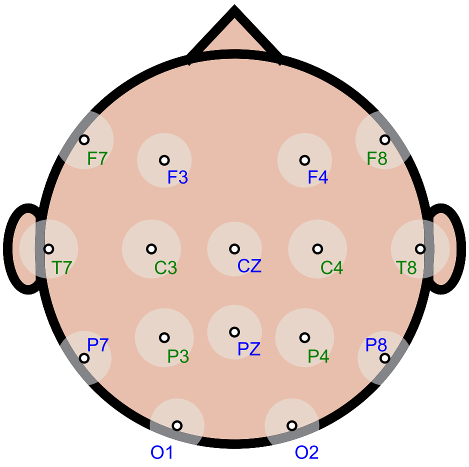

Fig. 5. 16-electrode configuration analyzed in this work. Selected electrodes for the 8-electrode configuration are labeled in green.

4.3. Contributions

In this study, we adapted the SHAP method ‘GradientExplainer’ for DL networks trained to decode MI from EEG. This adaptation has been open-sourced as part of the MEDUSA© Kernel Python package [32]. We applied this XAI-based method to explore the feature attribution maps generated by the state-of-the-art CNN on inter-subject MI classification, *EEGSym* [18]. This analysis was performed on two datasets comprising 171 subjects. Despite the common practice of using electrodes centered around the PM, M1, and S1 areas [6,8,9], XAI showed that MI involves a complex network that also includes the PPC and PFC. With the combination of DL and the adapted SHAP method, we discovered a temporal hierarchy among these regions in MI tasks. Furthermore, we identified a new 8-electrode configuration that can be employed in actual BCI applications, reducing the preparation time and allowing for the use of cheaper EEG caps.

4.4. Limitations and future work

Despite the promising insights on MI obtained from the application of SHAP-based XAI method and their correlation with previous knowledge on ME, we also acknowledge the following limitations that ought to be addressed in the future. On the one hand, we chose this paradigm to make use of *EEGSym* [18], which offers the best inter-subject capabilities and should show common patterns to a wide variety of subjects. However, by analyzing a binary classification problem we are limiting the information that can be extracted with XAI. It would be desirable to apply this analysis to a DL network trained to distinguish at least between left/right hand MI and the resting state. Furthermore, to strengthen the generalizability of our findings, it is recommended to apply this approach to a larger and more diverse sample of subjects and databases. Special attention should be given to including data from target populations for MI-based BCIs, such as stroke patients and individuals with movement disorders. Importantly, the 8-electrode configuration identified in this study holds promise for practical, online applications, making it particularly relevant for these target groups. Taking into account that MI seems to involve a hierarchical contribution of a wide variety of brain regions, we suggest that future MI-BCIs use electrode configurations that are not restricted to the M1 and S1 areas, which will increase decoding accuracy and in turn result in faster user adaptation. In this study, we corroborated the explanations obtained through a selection of an 8-channel configuration that could be compared with the one obtained in the original work [18]. Nevertheless, it would enrich the usefulness of this explanation method to study more in depth the relation between the percentage contribution to the model's decision and classification accuracy of the model presented in Fig. 4. Another equally important future line of work is to apply this validated method with different DL networks to other paradigms that are less studied in the literature and further advance the knowledge about the functioning of the brain.

5. Conclusion

Our XAI-based method for DL networks applied to EEG provides feature attribution maps through SHAP values. These maps shed light on the spatio-temporal distribution of the input signal, revealing significant contributions from the PFC and PPC, in addition to the well-known contributions from the M1 and S1. Our analysis also demonstrates the temporal hierarchy among these regions. These findings suggest that sensory, association, and motor areas play a crucial role in MI tasks, and MI-based BCIs should consider focusing on this broader network. The results indicate that the frontal channels F7 and F8, followed by central electrodes C3 and C4, are the most relevant for classification. By implementing a channel selection process based on SHAP values, we achieved a considerable improvement in accuracy on the Physionet dataset, reaching 86.5% ± 10.6%, and on the Carnegie Mellon University's dataset, achieving 88.7% ± 7.0% with a reduced 8-electrode configuration. Our XAI method, based on SHAP values, enables the discernment of important regions in the input signal for DL networks. This advances the knowledge of BCI paradigms employing these techniques, while potentially optimizing the EEG recording devices used, without compromising performance.

CRediT authorship contribution statement

Sergio Pérez-Velasco: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Diego Marcos-Martínez:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Eduardo Santamaría-Vázquez:** Writing – review & editing, Supervision, Software, Investigation, Formal analysis, Conceptualization. **Víctor Martínez-Cagigal:** Writing – review & editing, Visualization, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Selene Moreno-Calderón:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Roberto Hornero:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: R. Hornero reports financial support was provided by MCIN/AEI/10.13039/501100011033. S. Pérez-Velasco, D. Marcos-Martínez, E. Santamaría-Vázquez, V. Martínez-Cagigal, and R. Hornero report administrative support was provided by 'Centro de Investigación Biomédica en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)'. S. Pérez-Velasco and D. Marcos-Martínez report financial support was provided by 'Consejería de Educación de la Junta de Castilla y León'.

Acknowledgements

This research has been developed under the grants PDC2021-120775-I00, TED2021-129915B-I00, RTC2019-007350-1 and PID2020-115468RB-I00 funded by MCIN/AEI/10.13039/501100011033/ and European Regional Development Fund (ERDF) 'A way of making Europe'; under the R+D+i project 'EUROAGE+ Red Internacional de Investigación, Innovación y Transferencia de Tecnologías para la Promoción del Envejecimiento Activo' ('Cooperation Programme Interreg V-A Spain-Portugal POCTEP 2021–2027') funded by 'European Commission' and ERDF; and by 'Centro de Investigación Biomédica en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)' through 'Instituto de Salud Carlos III' co-funded with ERDF funds. S. Pérez-Velasco and D. Marcos-Martínez were in receipt of a PIF grant by the 'Consejería de Educación de la Junta de Castilla y León'.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpb.2024.108048.

References

[1] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, T.M. Vaughan, Brain–computer interfaces for communication and control, Clin. Neurophysiol. 113 (6) (2002) 767–791.

[2] J.R. Wolpaw, E.W. Wolpaw, Brain-Computer Interfaces: Principles and Practice, Oxford University Press, 2012.

[3] A. Ramos-Murguialday, M. Schürholz, V. Caggiano, M. Wildgruber, A. Caria, E.M. Hammer, S. Halder, N. Birbaumer, Proprioceptive feedback and brain computer interface (BCI) based neuroprostheses, PLoS ONE 7 (10) (2012) e47048.

[4] D.T. Bundy, L. Souders, K. Baranyai, L. Leonard, G. Schalk, R. Coker, D.W. Moran, T. Huskey, E.C. Leuthardt, Contralesional brain-computer interface control of a powered exoskeleton for motor recovery in chronic stroke survivors, Stroke 48 (7) (2017) 1908–1915.

[5] A. Moldoveanu, O.M. Ferche, F. Moldoveanu, R.G. Lupu, D. Cinteza, D. Constantin Irimia, C. Toader, The TRAVEE system for a multimodal neuromotor rehabilitation, IEEE Access 7 (2019) 8151–8171.

[6] M. Sebastián-Romagosa, W. Cho, R. Ortner, N. Murovec, T. Von Oertzen, K. Kamada, B.Z. Allison, C. Guger, Brain computer interface treatment for motor rehabilitation of upper extremity of stroke patients—a feasibility study, Front. Neurosci. 14 (October) (2020) 1–12.

[7] L.E. Miller, N. Hatsopoulos, Neuronal activity in motor cortex and related areas, in: J. Wolpaw, E.W. Wolpaw (Eds.), Brain–Computer InterfacesPrinciples and Practice, Oxford University Press, 2012, pp. 15–44.

[8] M.H. Lee, O.Y. Kwon, Y.J. Kim, H.K. Kim, Y.E. Lee, J. Williamson, S. Fazli, S.W. Lee, EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy, GigaScience 8 (5) (2019) 1–16.

[9] J.R. Stieger, S. Engel, H. Jiang, C.C. Cline, M.J. Kreitzer, B. He, Mindfulness improves brain–computer interface performance by increasing control over neural activity in the alpha band, Cereb. Cortex 31 (1) (2021) 426–438.

[10] E. Santamaria-Vazquez, V. Martinez-Cagigal, F. Vaquerizo-Villar, R. Hornero, EEG-inception: a novel deep convolutional neural network for assistive ERP-based brain-computer interfaces, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (12) (2020) 2773–2782.

[11] Y. Yu, Y. Liu, E. Yin, J. Jiang, Z. Zhou, D. Hu, An asynchronous hybrid spelling approach based on EEG–EOG signals for Chinese character input, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (6) (2019) 1292–1302.

[12] V. Martínez-Cagigal, J. Thielen, E. Santamaría-Vázquez, S. Pérez-Velasco, P. Desain, R. Hornero, Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): a literature review, J. Neural Eng. 18 (6) (2021).

[13] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, A. Cichocki, Correlation-based channel selection and regularized feature optimization for MI-based BCI, Neural Netw. 118 (2019) 262–270.

[14] J. Jin, R. Xiao, I. Daly, Y. Miao, X. Wang, A. Cichocki, Internal feature selection method of CSP based on L1-norm and Dempster–Shafer theory, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (2021) 4814–4825.

[15] L. Pan, K. Wang, L. Xu, X. Sun, W. Yi, M. Xu, D. Ming, Riemannian geometric and ensemble learning for decoding cross-session motor imagery electroencephalography signals, J. Neural Eng. 20 (6) (2023) 066011.

[16] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420.

[17] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces, J. Neural Eng. 15 (5) (2018) 1–30.

[18] S. Perez-Velasco, E. Santamaria-Vazquez, V. Martinez-Cagigal, D. Marcos-Martinez, R. Hornero, EEGSym: overcoming inter-subject variability in motor imagery based BCIs with deep learning, IEEE Trans. Neural Syst. Rehabil. Eng. 30 (2022) 1766–1775.

[19] Y. Xie, K. Wang, J. Meng, J. Yue, L. Meng, W. Yi, T.-P. Jung, M. Xu, D. Ming, Cross-dataset transfer learning for motor imagery signal classification via multi-task learning and pre-training, J. Neural Eng. 20 (5) (2023) 056037.

[20] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[21] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: 34th International Conference on Machine Learning, ICML 2017, vol. 7, 2017, pp. 4844–4866.

[22] I. Sturm, S. Lapuschkin, W. Samek, K.R. Müller, Interpretable deep neural networks for single-trial EEG classification, J. Neurosci. Methods 274 (2016) 141–145.

[23] C. Ieracitano, N. Mammone, A. Hussain, F.C. Morabito, A novel explainable machine learning approach for EEG-based brain-computer interface systems, Neural Comput. Appl. (Mar 2021) 0123456789 (Dl).

[24] D.O. Nahmias, K.L. Kontson, Easy perturbation EEG algorithm for spectral importance (easyPEASI), in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2020, pp. 2398–2406.

[25] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: NIPS 2017, vol. 32, 2017, pp. 1208–1217.

[26] H. Alsuradi, W. Park, M. Eid, Explainable classification of EEG data for an active touch task using Shapley values, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), in: LNCS, vol. 12424, 2020, pp. 406–416.

[27] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (Jun 2000).

[28] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 (7) (2015) e0130140.

[29] M. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": explaining the predictions of any classifier, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, vol. 46, Association for Computational Linguistics, 2016, pp. 97–101.

[30] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: 34th International Conference on Machine Learning, ICML 2017, vol. 7, 2017, pp. 5109–5118.

[31] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, SmoothGrad: removing noise by adding noise, jun 2017.

[32] E. Santamaría-Vázquez, V. Martínez-Cagigal, D. Marcos-Martínez, V. Rodríguez-González, S. Pérez-Velasco, S. Moreno-Calderón, R. Hornero, MEDUSA©: a novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research, Comput. Methods Programs Biomed. 230 (2023) 107357.

[33] R. Psotta, The visual reaction time distribution in the tasks with different demands on information processing, Acta Gymn. 44 (1) (2014) 5–13.

[34] E. Beam, C. Potts, R.A. Poldrack, A. Etkin, A data-driven framework for mapping domains of human neurobiology, Nat. Neurosci. 24 (12) (2021) 1733–1744.

[35] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80.

[36] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc., Ser. B, Methodol. 57 (1) (1995) 289–300.

[37] D. Marcos-Martínez, V. Martínez-Cagigal, E. Santamaría-Vázquez, S. Pérez-Velasco, R. Hornero, Neurofeedback training based on motor imagery strategies increases EEG complexity in elderly population, Entropy 23 (12) (2021) 1–19.