A Functional and Machine Learning-based Brain-Computer Interface for Mind Reading

Jessica Leoni, Silvia Carla Strada, Mara Tanelli
Politecnico di Milano, Dipartimento di Elettronica Informazione e Bioingegneria (DEIB), Milan, Italy
jessica.leoni@polimi.it, silvia.strada@polimi.it, mara.tanelli@polimi.it
Alice Mado Proverbio
Milan Center for Neuroscience, Università di Milano-Bicocca, Milan, Italy
mado.proverbio@unimib.it

***Abstract*—Brain-computer interfaces (BCIs) have revolutionized the way humans interact with machines, particularly for patients with severe motor impairments. However, current BCI systems have limited functionality due to the restricted pool of stimuli that can be distinguished from the electroencephalogram signal. Event-related potentials offer a promising solution, as each semantic category of visual and auditory stimuli (either imagined or perceived) evokes a response with a distinct shape that can be recognized by a BCI system. However, nowadays, the proposed BCIs relying on event-related potentials employ paradigms that requires patient's perception of the eliciting stimulus.**

In this work, we propose a novel BCI system that combines functional data analysis and machine-learning techniques to decode patients' thoughts from their EEG traces. Our system leverages functional data representation to preserve the time information of the collected signal and relies on an ad-hoc designed hierarchical ensemble classifier to gradually recognize event-related potentials associated with 14 different semantic categories of imagined stimuli. We validated our method on an extensive dataset collected from 20 volunteers. Our approach represents a significant advance in BCI technology as it recognizes semantic categories of stimuli based on event-related potentials evoked by imagination alone, without the need for external stimulation. Furthermore, we propose an approach to quantify the importance of each EEG channel in the decision-making process of the classifier, which can help reduce the number of embedded electrodes required for data acquisition, promoting patients' comfort. In conclusion, our work offers a promising step towards the development of a passive mind-reading interface that can recognize the consciousness of specific stimulus categories, even when they are not externally presented.

List of Acronyms

Acronym	Extended
BCI	Brain Computer Interface
EEG	Electroencephalogram
ERP	Event-Related Potentials
FDA	Functional Data Analysis
fPCA	Functional Principal Component Analysis
k-NN	k Nearest Neighbours
PCA	Principal Component Analysis
SSVEP	Steady-State Visual Evoked Potentials

I. Introduction

Brain Computer Interface (BCI) is intended to create a direct connection between the human brain and computerized devices, enabling individuals to operate such devices without peripheral muscle involvement [1]. While non-medical BCI applications exist, these systems are critical for patients with conditions such as locked-in syndrome, offering a means to reconnect with the external world. Several BCIs have been proposed over the years, leveraging information provided by brain electrical activity. Depending on the specific application, brain activity can be acquired either invasively through electrocorticogram (ECoG) or non-invasively through electroencephalogram (EEG) [2], [3]. EEG is commonly preferred as non-invasive, but scalp-measured brain electrical activity has lower signal-to-noise ratio, which remains a significant challenge when developing an effective BCI [4].

In more detail, non-invasive BCI relies on specific depolarizations in the EEG signal, which are identified and associated with corresponding user actions. To this extent, a dictionary is shared between the user and the machine that associates each possible stimulus with the intended action. While most BCIs rely on motor imagery, Steady-State Visual Evoked Potentials (SSVEP), or P300, they typically perform binary classification only, recognizing the presence or absence of expected patterns [5]. Also, active patient participation is required for SSVEP and stimuli perception paradigms. SSVEP is elicited by flashing a light at a specific frequency to the user, who must concentrate on the stimulus. To expand the pool of possible actions, multiple frequencies can be used, with each frequency associated with a specific one [6]. Stimuli perception is also required in eliciting P300, as it is a positive deflection of electric brain activity that occurs 300ms after stimulus recognition [7], indicating the presence of conscious mental attention processes [8].

Despite their effectiveness, BCIs that rely on SSVEP and P300 have limitations in terms of available functionalities [9]. This is because P300 is the same regardless of the stimulus that elicits it, and SSVEP is limited by the resolution of the system to detect depolarization changes due to the different flashing frequencies [10]. To address these limitations, researchers have explored the use of Event-Related Potentialss (ERPs), which offer a wider range of evoked responses that can be associated with different actions [11]. ERPs are time-locked responses to stimuli, and their shape depends on the category of the stimulus presented.

To exploit ERPs potential, neuroscientists have been extensively studied their morphology to identify reliable and characteristic markers. In more detail, the intent is to leverage the domain knowledge to define a set of rules that an expert BCI system can use to recognize different ERPs and perform the corresponding action. Additionally, these studies can provide insight into the cognitive processes underlying perception and provide a better understanding of this phenomenon. For example, in a recent work [12], researchers analyzed ERPs generated by 10 stimuli categories from both visual and auditory domains to extract reliable markers for the perception process. By interpreting the spatio-temporal coordinates of the brain activity, they identified relevant voltage peaks and statistically investigated their behavior with respect to the specific stimulus category.

Despite the interpretability benefits provided by an expert BCI system, the definition of hard rules can be challenging due to the complexity of EEG data. Indeed, EEG typically contains multiple channels, which refer to different brain regions. Moreover, the morphology and amplitude of brain waves varies across individuals, requiring a time-consuming fine-tuning procedure for each new user. To address these challenges, research has been conducted to leverage machine- and deep-learning techniques to automatically identify ERPs, assessing promising results [13]. The advantage of these techniques is that they not need any prior knowledge concerning the investigated domain. A considerable machine-learning approach is presented in [14], where ERPs are recognized belonging to 14 different semantic categories of stimuli. The proposed approach largely overcomes the accuracy threshold of 70% for each category of stimulus, which is considered as the minimum requirement to guarantee a meaningful BCI communication [15].

In addition to perceived stimuli, the use of imagined stimuli in BCI has been also explored for patients who are unable to produce observable responses, such as those in a coma or severe locked-in state [16]. Indeed, mind-reading applications would considerably expand the BCI potential in clinical field; for instance, they could investigate the presence of brain activity in vegetative patients. Again, studies have been conducted that demonstrate the potential of motor imagery-based BCI systems in enabling communication and control of external devices. For instance, in [17] a BCI system based on motor imagery was developed to allow patients to control robots. Similarly, in [18], a motor imagery-based BCI system was designed to control a wheelchair. Additionally, in [19], a similar paradigm was proposed to control a cursor on a computer screen. These studies highlight the potential of motor imagery-based BCI systems in improving the quality of life for individuals with severe motor disabilities.

Despite effective at moving objects, motor imagery BCIs still provide a reduced pool of functionalities. Therefore, researchers investigated ERP-based BCIs using imagined stimuli as an alternative to provide a wider set of actions to the patients. These works have been inspired by evidence reported by recent studies that have found an overlap in neural processing related to perceived and imagined stimuli [20]. In more detail, studies demonstrated that similar brain regions activate in response to the same stimulus, whether it is perceived or imagined [21]. Furthermore, multi-voxel approach revealed that similar sensory visual features can be identified in both perceived and imagined stimuli, with additional activity in the anterior fronto-temporal region during imagery tasks due to attention and memory processes [22]. For auditory stimuli, the secondary auditory cortex activates similarly in both perceived and imagined stimuli, while the primary auditory cortex tends to be more active in response to perceived stimuli [23]. However, recent findings have provided evidence of an activation of Heschl gyri (A1) during music imagery [24]. To further explore these findings, a recent study has identified reliable markers in imagery ERPs to develop a robust set of hard rules for implementing a mind-reading BCI system [25].

Despite the promising results assessed by neuroscientists at identifying a reliable set of marker to distinguish imagery ERPs, few approaches have been proposed that resort to machine- and deep-learning techniques. Prior works, such as those presented in [26] and [27], leverage deep-learning techniques to distinguish between ERPs that are elicited by imagery of different stimuli. Specifically, the former focused on distinguishing between ERPs that are elicited by imagery of homes and human faces, achieving an accuracy of 68%. On the other hand, the latter study combined the potential of convolution neural networks and genetic algorithms to distinguish between ERPs that are associated with imagery of dogs, airplanes, and houses, achieving 60% accuracy. However, these performances are not enough to create a reliable BCI system.

In this study, we introduce a novel machine learning-based BCI system that can learn characteristic patterns to accurately classify ERPs associated with different semantic categories of stimuli. To assess the performance of our proposed approach, we conducted an extensive experimental campaign involving 20 volunteers, who were presented with 40 stimuli belonging to 10 semantic categories and asked to both perceive and imagine them. The proposed system prove high accuracy, largely exceeding the 70% threshold for both perception and imagery ERPs. To reduce the number of electrodes needed, we also investigate how the number of considered channels affects the classifier's performance. Our findings indicate that many electrodes can be neglected, allowing to improve the patient comfort. These promising results lay the groundwork for the development of effective mind-reading BCI systems relying on stimuli imagery only.

A. Novel Contributions

In this work, we present a novel BCI system that relies on imagined stimuli, providing two main contributions that have not been previously explored in the field. First, we propose an automatic method that uses Functional Data Analysis (FDA) and machine-learning to robustly attribute an ERP to the semantic category of the eliciting stimulus. To the best of the authors' knowledge, this is the first time that FDA has been applied to the ERP recognition problem. It has been used as effective in retaining the time information and reducing data dimensionality by means of Functional Principal Component Analysis (fPCA). Additionally, we have designed a hierarchical classifier architecture that simplifies the multi-category stimuli classification problem as a series of binary classifications, which enhances the stimuli recognition performance.

The second contribution of this work is an ad-hoc designed approach to quantify the importance of each channel in the decision-making process of the hierarchical classifier. This approach calculates the Kendall τ correlation coefficient between each EEG channel and the first principal component identified by fPCA, providing an effective tool to reduce the number of electrodes embedded in the acquisition cap and improve patients' comfort. Furthermore, this approach allows for automatic identification of relevant brain areas involved in stimuli classification, which turns out to be coherent with neuroscientific knowledge.

To validate the proposed BCI system, grand averaged ERPs triggered by 10 semantic categories of stimuli, both visual and auditory, were collected during an extensive experimental campaign involving 20 volunteers. Performance was evaluated through k-fold cross-validation to ensure robustness when considering new ERPs belonging to the same subjects analyzed in training, and leave-one-out validation to investigate the impact of BCI illiteracy on hierarchical classifier's accuracy. The results showed that the proposed BCI system assess outstanding performance on both perception and imagery ERPs, vastly exceeding the 70% threshold for effective communication.

The most significant achievement of this work is demonstrating the ability to recognize the imagined stimulus that triggers an ERPs, represents a significant step forward in the field of mind-reading applications that are effective even in patients who are in a coma or affected by locked-in syndrome.

The rest of the paper is organized as follows: Section II provides insights concerning the experimental procedure used to collect the ERPs data considered in this work. Then, Section III details the method designed to attribute each ERPs to the semantic category of the stimulus that has elicited it. Section IV discusses the system's performance. Section V explains the approach designed to investigate EEG channels' importance and presents the obtained results. Finally, Section VI remarks on the main contributions provided and depicts future work.

II. Experimental Setup

This Section outlines the experimental methodology designed for ERPs data collection. Specifically, it includes information regarding the participants and the stimuli used, as well as details concerning the procedure of the stimulation runs each volunteer underwent. Additionally, it reports the specifications of the acquisition devices utilized and the preprocessing techniques implemented on the collected signals. For additional information regarding the data collection procedure, please refer to [12], [25], where further neuroscientific insights into the collected ERPs' components are also provided.

A. Experimental Protocol

The data collection process involved a group of 20 volunteers (13 female and 7 male) with an average age of 23.9 ± 3.34 years, whose characteristics are summarized in Table I. All participants were right-handed according to the Edinburgh Inventory Questionnaire [28]. Furthermore, they had normal or corrected-to-normal vision and hearing, and did not experience deficits in language comprehension, reading, or spelling. None of the volunteers had been previously diagnosed with a psychological or psychiatric disorder or drug abuse. The experimental protocol adhered to the Helsinki Declaration of 1964 and was approved by the Ethics Committee of Bicocca University (protocol number RM-432). Each participant provided written informed consent before participating in the data collection process.

Each volunteer was instructed of the experimental protocol which has the same structure for all participants. First, participants were asked to wear Sennheiser electronic gmbH headphones and sit in an anaechoic and faradized cabinet, 114cm away from a HR VGA color monitor located outside. They were also instructed to stay still and avoid any ocular or body movements while fixing their gaze on a red point displayed at the center of the screen. The experiment comprised 8 runs of visual stimuli and 4 runs of auditory stimuli categories. Each visual run had a duration of approximately 3 minutes, while each auditory run lasted for 2 minutes and 30 seconds. 40 different stimuli belonging to the same category were presented in each run. Considering all the runs, each subject was presented with 40 different stimuli for each of the 10 categories, 7 of which were visual and 3 were auditory. Therefore, there were a total of 280 visual stimuli and 120 auditory stimuli. Each visual stimulus consists of a 18.5×13.5 cm image that was presented at the center of the monitor for 1500ms, surrounded by a white background. Auditory stimuli, instead, consists of 1500ms recordings reproduced by the headphones from an iPhone 7 and a Huawei P10. The visual stimuli were matched for sensory properties such as luminance, color, and size. Additionally, human-related visual and auditory stimuli were matched based on perceptual properties such as sex and age, while written and spoken words were matched based on linguistic properties. Specifically, 40

TABLE I: Experimental Campaign: Subjects. This Table reports some insights concerning the 20 volunteers involved in the experiment.

Volunteer ID	Age	Gender	Dominance
1	29	F	0.48
2	20	F	0.81
3	29	F	0.90
4	23	F	0.48
5	20	F	0.62
6	23	F	1.00
7	30	M	0.90
8	22	F	0.90
9	25	F	1.00
10	23	F	0.86
11	26	M	0.81
12	22	F	0.86
13	19	M	0.20
14	26	M	0.76
15	28	M	1.00
16	26	F	0.71
17	22	F	1.00
18	23	M	1.00
19	20	F	0.62
20	21	M	0.71

TABLE II: Experimental Campaign: Semantic Categories of Stimuli. This Table reports some insights concerning the categories of the stimuli employed in the experiment.

Sensory Domain	Stimulus Category	Composition
Visual	Infant Face	20 faces of baby males and 20 females
	Adult Face	20 faces of adult males and 20 females
	Animal	40 heads of different mammals
	Bodies	20 dressed bodies of females and 20 males
	Written Words	40 frequent Italian words
	Objects	40 non manipulable and familiar objects
	Checkerboards	40 different checkerboards
Auditory	Music	40 different piano recordings
	Emotional Vocalization	40 voices of crying, fear, laughter, and surprise
	Spoken Words	40 Italian words by 20 male and 20 female

common Italian words were selected. Auditory stimuli has an intensity levels ranging from 20 to 30 dB, which was normalized and leveled by intensity and volume. Further details regarding the stimuli can be found in Table II.

The stimuli were presented to the volunteers using Eevoke Software for audiovisual presentation (ANT, Enschede, The Netherlands). The proposition of each stimulus was followed by a mental imagery task. In more detail, when the proposed stimulus was removed, an intra-stimulus interval of $500 \pm 100ms$ followed, in which a gray screen was displayed. Subsequently, a yellow frame appeared on the screen, indicating that the volunteer has to imagine the picture or sound that was just presented. After 2000ms, the yellow frame disappeared, and the volunteer was given an inter-trial interval of $900 \pm 100ms$ before the next stimulus was presented. 2000ms are given to the volunteers for imaging each stimulus as evidence in the literature is reported that at least 500ms are required to figure an alphabet letter, and this interval considerably increases considering auditory stimuli [29]. Last, to guarantee that the volunteers were concentrated during the experiment, they were informed beforehand that at the end of the experiment they would be asked to complete a questionnaire related to the stimuli presented.

B. Acquisition Procedure and Collected Dataset

During the experiment, EEG data was continuously recorded by a cap embedding 126 electrodes that sample at 512Hz and placed according to the 10/5% system by Oostenveld and Praamstra [30]. The electrodes' impedance was kept below $5k\Omega$. The electrodes record both EEG and electrooculogram (EOG), and use the linked mastoids (M1, M2) as reference leads. ANT software was used to acquire and clean the data. In detail, it applies a band-pass filter between 0.016 and 30Hz to all the EEG channels, and between 0.016 and 70Hz for the EOG channels. Artifacts caused by eye movements, blinks, or excessive muscle potential were also rejected using peak-to-peak amplitude exceeding 50µV as criteria, leading to a rejection rate of 5%.

Furthermore, ANT acquisition software performs ERPs grand averaging to improve the signal-to-noise ratio of the potential elicited by each stimulus [31]. Specifically, for both the perception and imagery recordings, the 40 responses of each subject associated with stimuli belonging to the same semantic category were first synchronized based on the trigger onset and then averaged. The resulting collections of averaged ERPs for perception and imagery stimuli constitute the datasets produced at the end of the experimental procedure. Both perception and imagery datasets consisted of 200 ERPs, 10 per subject, one for each category of stimulus. In the perception dataset, each ERP lasted 1500ms, with 100ms ore-stimulus baseline; In the imagery dataset, each ERP lasts 2000ms, with a 100 pre-stimulus baseline.
An example of the ERPs collected in the perception and imagery datasets is reported in Figure 1, where Subject 1's

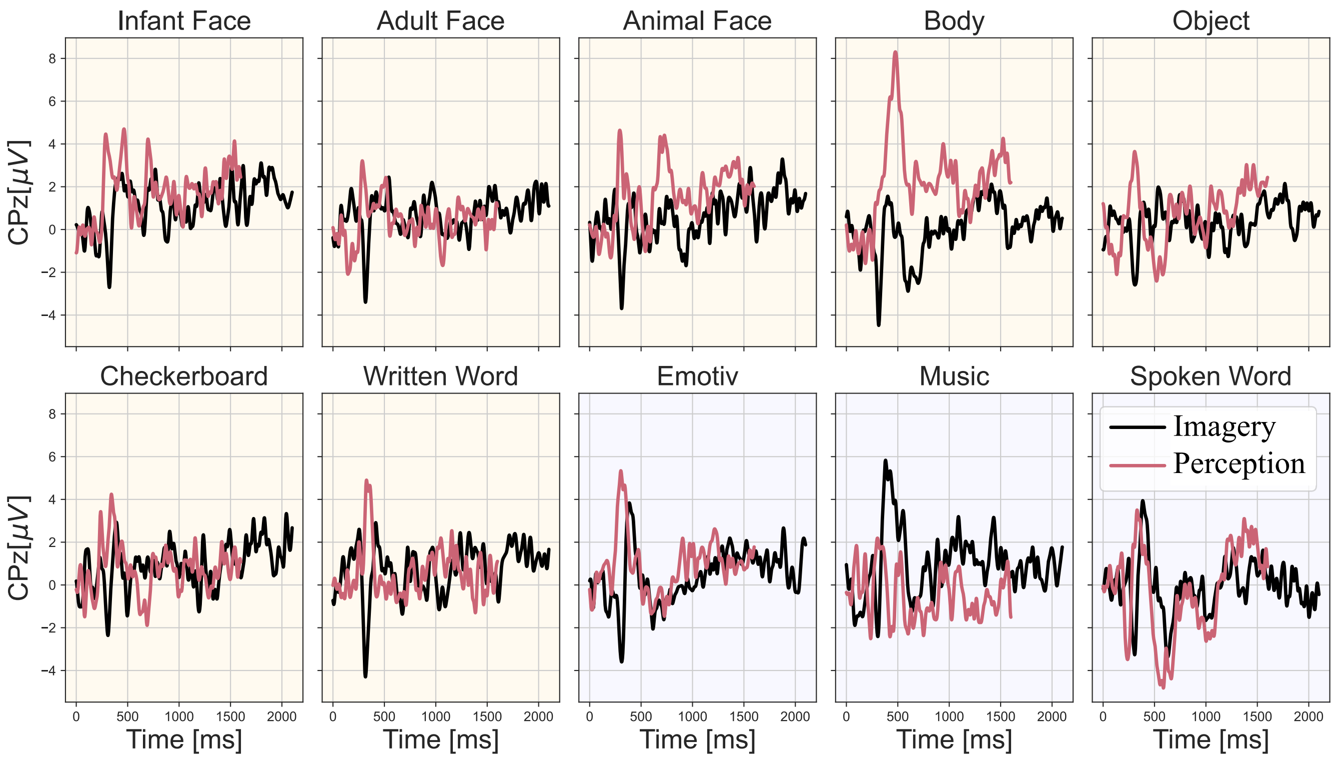


Fig. 1: Perception vs Imagery: CP1 Trend.
This Figure refers to Subject 1 and shows the trends measured by CPz channel for all the semantic categories of stimuli, comparing perception and imagery ERPs. Please notice that, to ease the comparison, the plotted trends have been subjected to baseline correction.

trends of the CPz channel are presented for all the stimulus categories. It is interesting to notice that, as reported in Section I,

P300 component was smaller and later in the imagery than in perception datasets). This result further suggests how imagery is weaker and noisier imaginative experience, as opposed to the more vivid and detailed perceptual experience

[25].

III. Proposed Method

This Section details the pipeline developed to recognize the category of the stimulus related to the measured ERPs, and presents the performance metrics used for evaluation. The pipeline is composed of several stages, starting with pre-processing. During this stage, each response undergoes baseline correction and extraction of the region of interest. The windowing stage follows, and the extracted windows are transformed into a set of functions using FDA.

Afterward, fPCA is performed, and a set of representative features is computed. The resulting instances are then provided to the ad-hoc designed hierarchical machine-learning classifier, which predicts the categories. The presented method was evaluated according to k-fold cross-validation and leave-one-out validation. A comprehensive set of metrics has been considered, to provide a reliable understanding of the proposed approach's performance and enable comparison with other methods in the literature.

A. Pre-Processing

The first step of the presented ERPs classification pipeline is pre-processing. As reported in Figure 2, it consists of two stages: baseline correction and frame selection. As grand averaged ERPs are affected by noise, baseline correction is essential. Indeed, this step aims to remove the resting-state activity from the signal, making sure that any measured voltage changes are due to the stimulus, and not to the ongoing brain processes [32]. For each grand averaged ERP, the mean voltage recorded before the onset of the stimulus is subtracted channel by channel for each grand averaged ERP. After this procedure, the unnecessary channels, *i.e.*, the ones related to the ocular (vEOG, hEOG) and mastoids (M1, M2) are removed.

Then, frame selection is performed, to improve the signal-to-noise ratio [32]. Considering imagery stimuli categories, we consider as relevant the 400 to 1000ms range, while for perception stimuli categories, we maintain the grand averaged ERP's portion included in 100 to 1000ms range. The choice of

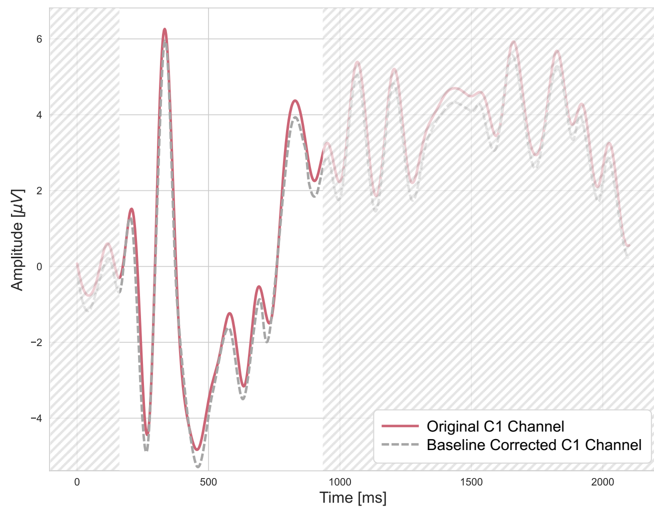


Fig. 2: Pre-processing.
This Figure shows the pre-processing pipeline applied to each channel of each recording, which consists of baseline correction and removal of the frames not related to the grand averaged ERPs. The reported case refers to CP1 channel of a recorded while proposing an image of an infant face to Subject 1. As it belongs to the perception dataset, the frame between 100ms and 1000ms is kept.

considering a smaller interval in imagery ERPs is due to the lack of P300 component. Indeed, as reported in Section I, it is related to attention processes, so it is not typically observed in potentials that are not evoked by an external stimulation [33].

B. Windowing

Windowing is a technique widely used in machine learning to enhance model training and provide robustness to time shifts in data collection. It involves dividing the pre-processed signals into smaller segments of fixed length leveraging a window that slides of a specific factor, defined as slope. Windowing is also essential when designing an online BCI to enable real-time ERPs classification. Indeed, the window and the slope determine the time required to get the first prediction and the update rate, respectively. It follows that window size is a crucial parameter that must be properly fine-tuned. Wider windows correspond to a reduced number of instances and longer wait to get the first prediction, while smaller windows may not accurately represent the grand averaged ERP behavior, leading to inconsistent results. Two main categories of approaches to fine-tune the window size are a priori and a posteriori. A priori approaches optimize the window size using a cost function that measures the class separability provided by the features extracted from each evaluated window, while a posteriori approach relies on the performance of the classifier in recognizing the classes from the extracted features while varying the window size. In the design of the proposed BCI, we opted for a posteriori approach and evaluated different window sizes ranging from 100ms to 600ms. For each one, we extracted the corresponding set of features and measured the performance of the classifier in recognizing the semantic category of the stimuli by considering the f1-score as the cost function. The fine-tuning was conducted separately for the perception and imagery datasets. In both cases, the optimal window size was found to be 500ms, while the slope was set to 1s.

C. Features Extraction

After the windowing stage, the perception and imagery datasets consist of two collections of grand averaged ERPs' windows. In more detail, they are composed of a set of 6600 and 9300 windows 500ms long. As reported in Figure 3a, each window collects the samples measured in the window time frame by the 122 channels. Formally, we can define the perception and the imagery datasets as χ^{pro} and χ^{fig}, where each χ is a collection the windows $ERP_{i,j,w}$ extracted from the grand averaged ERP recorded for the i^{th} subject when triggered by the j^{th} stimulus:

$$\chi = \{ERP_{1,1,1}, \dots, ERP_{i,j,w}, \dots, ERP_{St,Su,W}\}$$

provided that St is the number of stimuli categories in the dataset, *i.e.*, 10, Su the number of volunteers, *i.e.*, 20, and W the number of overall windows extracted from all the grand averaged ERP. Moreover, each $ERP_{i,j,w}$ collects all the discrete samples collected in the w^{th} window frame:

$$ERP_{i,j,w} = [ERP_{i,j,w}(1,a), \dots, ERP_{i,j,w}(k,t), \\ \dots, ERP_{i,j,w}(K,b)]$$

where K is the number of channels, *i.e.*, 122, while a and b are the extremes of the grand averaged ERP window's domain, *i.e.*, $ERP_{i,j,w} : [a,b]$ and $b - a = 500ms$.

The high number of channels can cause a decrease in classifier performance due to the curse of dimensionality problem [34]. This occurs because as more features are added, the dimension of the space in which the instances are represented increases, causing each instance to cover a smaller volume, making it more challenging for the classifier to learn a robust decision function without overfitting. One common technique used to address this problem is Principal Component Analysis (PCA), which reduces the instances dimensionality while preserving most of the information [35]. However, PCA does not account for the temporal dependency of grand averaged ERPs, for which the shape of the curves is a critical factor that cannot be neglected in their classification. To overcome this issue, we decided to represent the data using FDA, which allows us to use fPCA. fPCA is an extension of PCA that accounts for the time dimension and is specifically designed to handle data where observations can be represented as functions, such as EEG time-series [36].

Therefore, we transform our data resorting to FDA techniques. According to this representation, each window is composed of a set of 122 functions fitted to the discrete samples provided by each EEG channel. To fit the functions, B-splines are used, which are piecewise polynomial functions defined over a sequence of knots that partition the function domain in intervals. Within each interval, the B-spline is a polynomial of a fixed degree. The degree of smoothness is determined by the number and location of the knots, as well as the order of the B-splines. B-splines are preferred over Fourier or other options because they handle irregularly spaced data and allow for flexible modeling of complex functions. Additionally, the use of B-splines can help to maintain important features of the original data, such as peaks and troughs, while removing noise [37].

In the designed approach, the EEG channels' functions are approximated using 5 knots (m) and a polynomial of order (r) equal to 2. In more detail, let $ERP_{i,j,w}(k,t)$, $t = [a,b]$, be the set of N_w discrete samples associated to the k^{th} channel of the grand averaged ERP window referred to the j^{th} subject elicited by the i^{th} stimulus. Let also τ_n, $n = 1, 2, \ldots, m$ be a set of m knots equally spaced in the function domain. The knots divide the window interval $[a,b]$ into $m+1$ subintervals, where $a = \tau_1 \leq \tau_2 \leq \cdots \leq \tau_m \leq \tau_{m+1} = b$. It follows that the basis smoothing of the grand averaged ERP window's data can be defined by a linear combination of the fitted splines:

$$\widehat{ERP}_{i,j,w}(k,t) = \sum_{n=1}^{m+2r-1} B_{n,r}(k,\tau) \tag{1}$$

where $\widehat{ERP}_{i,j,w}(k,t)$ is the smoothed function, $B_{n,r}(k,\tau)$ is the B-spline of order r associated to the n^{th} knot, and ω_n are the weights attributed to each B-spline in the linear combination, estimated by least squares. As reported in Figure 3b, at the end of this step, the window consists of a set of 122 functions.

As previously mentioned, FDA representation enables us to leverage fPCA to reduce instances dimensionality, while preserving temporal information [38]. In fPCA, the objective is to decompose the data into a set of orthogonal functions known as principal components, which capture the main sources of variability in the data. Each functional principal component is associated with a score that represents the amount of explained data variance. Usually, the first few principal components capture the largest sources of variability, so that the others can be neglected. To perform fPCA, we first center and scale the windows to have zero mean and unit variance. Then, we compute the eigenfunctions and eigenvalues of the covariance operator, which describes the variation in the data over time and is defined as:

$$\begin{aligned} &Cov[ERP_{i,j,w}(k,t), ERP_{i,j,w}(l,t)] = \\ &\frac{1}{N_w - 1} \sum_{t=1}^{N_w} \big[\big(ERP_{i,j,w}(k,t) - \overline{ERP}_{i,j,w}(k,t)\big) \\ &\qquad \big(ERP_{i,j,w}(l,t) - \overline{ERP}_{i,j,w}(l,t)\big) \big] \end{aligned}$$

where $ERP_{i,j,w}(k,t)$ represents the k^{th} channel function of the w window at time t, $\overline{ERP}_{i,j,w}(k,t)$ is its mean across all time points, and N_w is the number of time points for each channel in the window. As we are considering functions, the functional covariance must also be computed from the sample covariance by solving the following integral equation:

$$\begin{aligned} &\int_a^b Cov[ERP_{i,j,w}(k,t), I(ERP_{i,j,w}(l,t))]dt \\ &= \lambda \int_a^b I(ERP_{i,j,w}(k,t)) I(ERP_{i,j,w}(l,t))dt \end{aligned}$$

where λ is the eigenvalue associated with each eigenfunction and $I(\cdot)$ is the integral operator. The solution to this problem can be expressed as a series expansion:

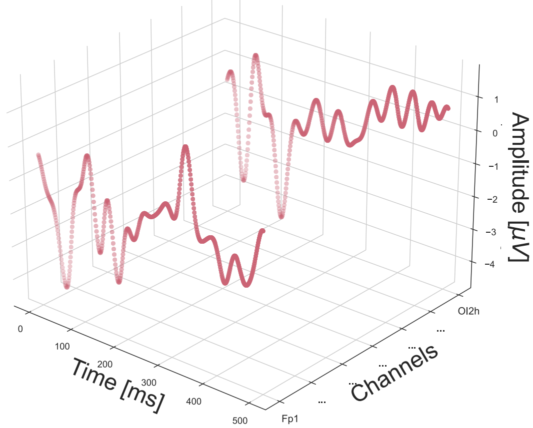
$$I(t) = \sum_{p=1}^{\infty} \sqrt{\lambda_p} \phi_p(t) \xi_p \tag{2}$$

where $\phi_p(t)$ are the eigenfunctions of the covariance operator, λ_p are the corresponding eigenvalues, and ξ_p are the coefficients of the expansion. Then we can project each window onto the fPCA space to obtain a lower-dimensional representation by retaining only the information provided by a reduced set of principal components.

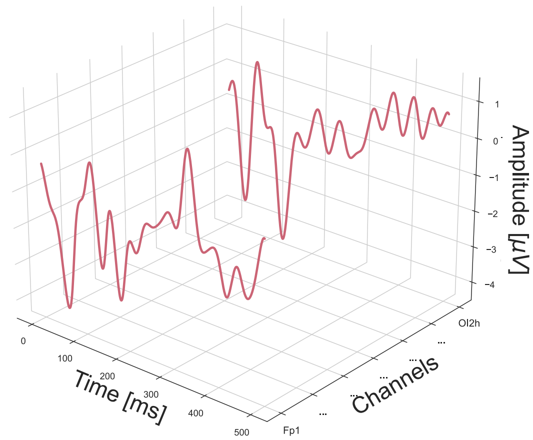
The first functional principal component captures most of the variance (over 95%) in both perception and imagery datasets; therefore this is the only one retained. It follows that fPCA reduces each window dimensionality from 122 channels to 1 functional principal component. As reported in Figure 3c, to highlight its variations within the window, we also calculate the first derivative of the component. Finally, we extract statistical features, *i.e.*, mean, standard deviation, minimum, and maximum from both the component and its first derivative. By extracting these 8 features for each window, we can provide a concise summary of the included information to the machine learning classifier, guaranteeing robust training performance.

D. Classifier Learning and Evaluation

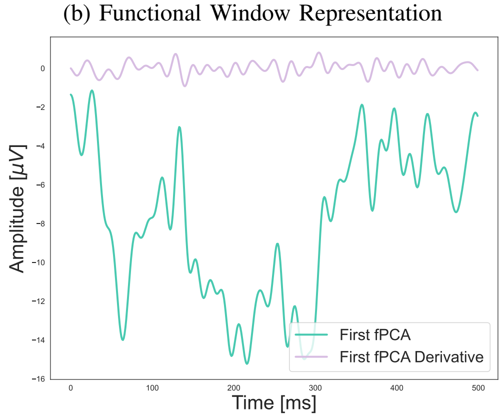
After the features extraction process, each window of the grand averaged ERPs is represented by eight features, including the mean, standard deviation, minimum, and maximum of the first functional principal component and its derivative. These instances are presented to the ad-hoc designed hierarchical classifier for training and evaluation purposes. The evaluation procedure was first conducted considering k-fold cross-validation with a stratified split, where 25% of the data is used for testing purposes. This enables us to assess the model's predictive capabilities when presented with grand averaged ERPs from the same subjects considered in training. Additionally, we investigate the model's performance on new users by resorting to leave-one-out validation, where the classifier is trained on all the instances in the dataset except for those belonging to one user, which constitute the test set. We repeat this procedure for all subjects, and the average evaluation metrics are considered.



(a) Original Window



(b) Functional Window Representation



(c) Extracted First fPCA Component and Derivative

Fig. 3: fPCA Computation. Each 500ms window is composed of the samples collected for the the 122 channels in that time range. To compute fPCA, the first step is to estimate from the samples of each channel the respective function. Then, fPCA estimates the principal components that best explains the variance of the channel functions. As in each window the first component explains more than 95% of the variance, it is the only one considered. From this component and its derivative the features are extracted.

To design the hierarchical structure, stimuli categories are grouped into fictional macro-categories according to a semantic perspective to create a tree structure, whose architecture is reported in Figure 4. Grouping similar categories together improves model interpretability and, depending on the end-user's requirements, allows for high-level predictions while maintaining a correct semantic. Each binary classification is performed by a k Nearest Neighbours (k-NN) machine-learning model. k-NNk-NN identifies the k nearest neighbors to a new data point in the training set and classifies it based on the most commonly represented class among the neighbors [39]. One of the advantages of k-NN is that it is a non-parametric algorithm that makes no assumptions about the underlying data distribution. The key parameter in this model is the number of neighbors considered, k, which we fine-tuned and set to 5. Each k-NN in the tree is trained and evaluated individually, and overall architecture performance is also estimated.

The f1-score is considered the main evaluation metric due to its ability to provide a balanced view of performance, unlike accuracy, which can be misleading in datasets with unbalanced class distribution. It is defined as the harmonic mean of precision and recall :

$$f1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \cdot 100. \tag{3}$$

Precision and recall can be computed from the confusion matrix, a statistical tool used to compare actual and predicted classes. In more detail, the precision is the ratio of true positives to the total number of instances attributed by the classifier to the positive class:

$$Precision = \frac{TP}{TP + FP} \cdot 100 \tag{4}$$

while recall is the ratio of true positives to the total number of instances belonging to the positive class:

$$Recall = \frac{TP}{TP + FN} \cdot 100 \tag{5}$$

In these equations, TP represents true positives, *i.e.*, instances correctly classified as positive; FP represents false positives, *i.e.*, instances incorrectly classified as positive, and FN represents false negatives *i.e.*, instances incorrectly classified as negative.

IV. Experimental Results Evaluation and Discussion

This Section presents and analyzes the performance of the proposed BCI system at recognizing the stimuli associated to each ERP window in both perception and imagery datasets. To this end, k-fold cross-validation and leave-one-out validation have been leveraged.

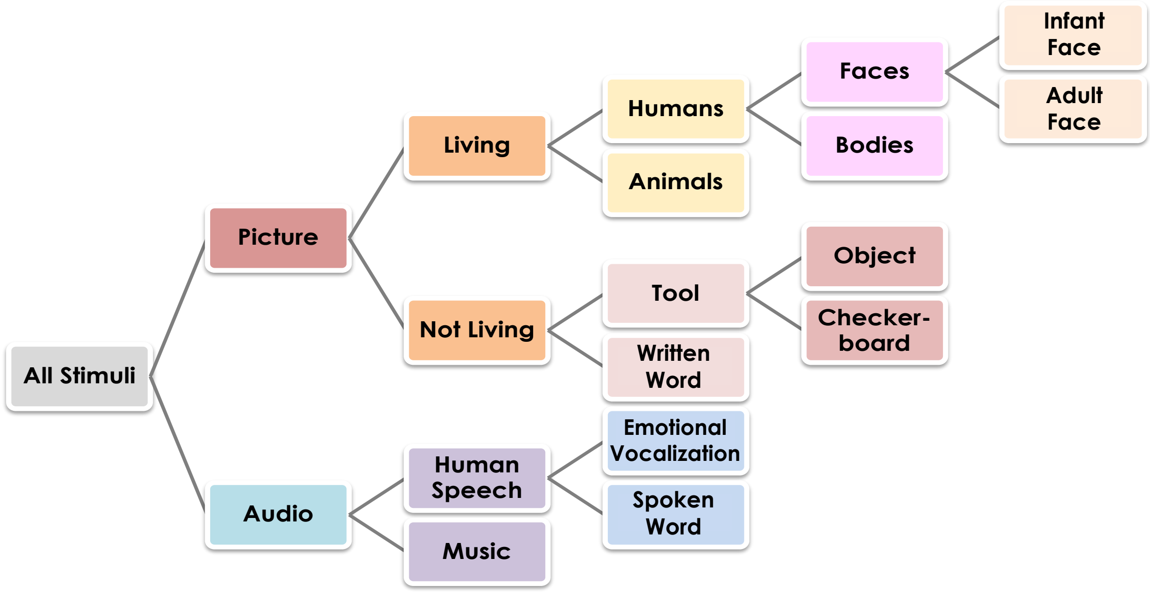


Fig. 4: Hierarchical Classifier Architecture.
This Figure shows the hierarchical classification architecture designed to identify the grand averaged ERPs. Each binary split is performed by a machine-learning model, ad-hoc trained to distinguish the two classes.

A. k-Fold Cross-Validation

To evaluate our method's ability to recognize new grand averaged ERPs belonging to the same subjects as those in the training set, we utilized 10-fold cross-validation. This approach is widely adopted in machine-learning as it ensures reliable and consistent estimation of the classifier's performance [40]. We first evaluated the performance on the perception datset to prove the effectiveness of the designed approach and to provide a benchmark for comparing other works in the literature. At each iteration, we stratified partitioned the dataset instances between a training set (75%) and a test set (25%). Stratified sampling was used to ensure similar class distribution in the training and test sets, which is crucial in case of imbalanced datasets. The performance metrics were computed as the average assessed on the test set for each iteration. We evaluated the performance of each binary classifier individually and within the hierarchical architecture.

Then we also applied the same procedure to the imagery dataset. Tables III and IV present the precision, recall, and resulting f1-score of each binary k-nn classifier on the perception and imagery datasets, respectively. The classifiers achieved an average f1-score of 94.26% and 97.34%, with both datasets performing best for written and spoken words. However, the most frequently misclassified stimulus was the infant face. Upon further investigation, it was discovered that 7.63% and 1.89% of the grand averaged ERPs windows associated with infant face were incorrectly attributed to adult face during the split considering perception and imagery datasets, respectively.

Then, we evaluated the performance of the binary k-NN classifiers within the hierarchical structure. To this extent, we provided all instances in the dataset to the first binary classifier, which aimed to distinguish between visual and auditory stimuli categories. The instances were then passed on to the following binary classifiers according to the predicted category, until a leave of the hierarchical structure is reached. Table V and VI show the resulting precision, recall, and f1-scores for the perception and imagery datasets, respectively. The f1-score assessed to 91.10% and 94.13% for the perception and imagery datasets, respectively. The infant face was the least recognized class, and a higher percentage of wrongly predicted windows should be attributed to the animal stimulus. This could be due to baby schema, *i.e.*, some animals sharing

TABLE III: Classification Report for Binary Classifiers: *Perception*. This Table reports the precision, recall, and F1-Score for the grand averaged ERPs in perception dataset when predicted by the each binary classifier independently. The support is equal to 660 samples for each class.

Stimulus	Evaluation Metrics		
	Precision	Recall	F1-Score
Infant Face	76.22	92.51	83.18
Adult Face	92.71	94.53	93.13
Animal	99.99	85.12	92.05
Body	99.99	93.71	96.02
Checkerboard	91.81	94.34	92.97
Object	94.52	90.21	92.72
Written Word	99.99	99.99	99.99
Emotional Vocalization	95.51	97.76	96.24
Spoken Word	99.99	99.99	99.99
Music	97.31	95.52	96.33

TABLE IV: Classification Report for Binary Classifiers: *Imagery*. This Table reports the precision, recall, and F1-Score for the grand averaged ERPs in imagery dataset when predicted by the each binary classifier independently.The support is equal to 930 samples for each class.

Stimulus	Evaluation Metrics		
	Precision	Recall	F1-Score
Infant Face	87.15	98.58	92.57
Adult Face	98.23	96.36	97.65
Animal	99.99	95.63	97.97
Body	99.99	95.12	97.83
Checkerboard	98.71	97.31	98.24
Object	97.15	98.29	98.43
Written Word	99.99	99.99	99.99
Emotional Vocalization	94.62	96.16	95.41
Spoken Word	99.99	99.99	99.99
Music	96.31	94.22	95.34

TABLE V: Classification Report for Hierarchical Classifier: *Perception*. This Table reports the precision, recall, and F1-Score for the grand averaged ERPs in perception dataset when predicted by the hierarchical classifier. The support is equal to 660 samples for each class.

Stimulus	Evaluation Metrics		
	Precision	Recall	F1-Score
Infant Face	97.	80.	88.
Adult Face	90.	93.	91.
Animal	93.	83.	88.
Body	91.	90.	91.
Checkerboard	91.	93.	92.
Object	80.	95.	87.
Written Word	96.	96.	94.
Emotional Vocalization	94.	94.	94.
Spoken Word	89.	92.	91.
Music	93.	96.	94.

TABLE VI: Classification Report for Hierarchical Classifier: *Imagery*. This Table reports the precision, recall, and F1-Score for the grand averaged ERPs in imagery dataset when predicted by the hierarchical classifier. The support is equal to 660 samples for each class.

Stimulus	Evaluation Metrics		
	Precision	Recall	F1-Score
Infant Face	99.99	93.24	90.79
Adult Face	93.17	96.81	94.95
Animal	96.59	94.74	95.48
Body	95.29	93.43	94.24
Checkerboard	94.78	97.36	95.21
Object	93.31	96.67	94.14
Written Word	97.11	96.96	96.15
Emotional Vocalization	91.86	98.15	94.93
Spoken Word	93.52	95.33	94.65
Music	99.99	89.14	94.51

physical characteristics with babies, as suggested by ethologist Konrad Lorenz's research in 1943, eliciting similar responses in humans. [41], [42].

B. Leave-One-Out Validation

The purpose of leave-one-out validation is to assess the performance of a BCI system at classifying ERP windows belonging to a new subject. Despite high leave-one-out validation performance would be desired, as it would guarantee that the system is effective on new patients without requiring classifier re-training, it is well known in the literature that differences in human physiology and brain patterns cause BCI illiteracy [43]. To evaluate the performance of our system on new users, we trained the hierarchical architecture on all but one user and repeated the procedure for all subjects, considering the average performance. The results for the perception and imagery datasets at each split level are reported in Table VII and VIII, which also provide a comparison with k-fold cross-validation outcomes. Despite our method assesses high f1-score even on new users in both perception and imagery datasets, the classifier trained and evaluated on the same subjects performed better than those trained on some subjects and tested on a new one.

Additionally, Figure 5 shows the f1-score obtained by the hierarchical classifier according to leave-one-out validation for each subject, considering the perception and imagery datasets. The results demonstrate inter-subject variability in performance, with some subjects achieving higher recognition accuracy than others. This finding is consistent with previous studies indicating that similarity in cortical organization or patterns of brain activity can influence BCI performance [44]. Similarly, evidence is reported in literature that individuals who had more similar EEG patterns to the training set achieved higher classification accuracy in BCI tasks [45].

Last, the trend in predictions over time for Subject 1 was investigated and reported in Figure 6. The top subplots show the predictions made when the hierarchical classifier was trained on 75% of the instances equally distributed by stimuli categories and subjects. The bottom subplots show the predictions made by an instance of the hierarchical classifier trained on the instances of all the subjects, except for Subject 1, who constituted the test set. Little classification error occurred for both perception and imagery stimuli categories according to k-fold cross-validation. Additionally, by performing a majority voting and attributing to the grand averaged ERP the label most attributed to its windows, the performance can be further enhanced. The same applies to the hierarchical classifier trained according to leave-one-out validation. However, in this case, more errors occur, mostly in discriminating adult and infant faces or music and emotional vocalization stimuli categories.

TABLE VII: F1-Score in Perception Dataset. This Table reports the F1-Score assessed by each classifier embedded in the grand averaged ERPs recognition pipeline designed to recognize the perception stimuli categories when evaluated according to k-fold cross-validation and leave-one-out validation.

Split Node	**k-Fold**		**Leave-One-Out**	
	F1-Score	**Train Support**	**F1-Score**	**Train Support**
Audio vs Picture	95.50	4955	81.36	6380
Living vs Not Living	91.0	3465	74.03	4488
Human Speech vs Music	97.35	1485	83.33	1914
Humans vs Animals	94.24	1980	82.95	2552
Tool vs Written Word	92.32	1485	89.39	1914
Emotional Vocalization vs Spoken Word	96.67	990	90.91	1276
Faces vs Body	90.91	1485	89.39	1914
Object vs Checkerboard	94.85	990	90.91	1276
Infant Face vs Adult Face	90.61	990	90.91	1276

TABLE VIII: F1-Score in *Imagery Dataset*. This Table reports the F1-Score assessed by each classifier embedded in the grand averaged ERPs recognition pipeline designed to recognize the imagery stimuli categories when evaluated according to k-fold cross-validation and leave-one-out validation.

Split Node	**k-Fold**		**Leave-One-Out**	
	F1-Score	**Train Support**	**F1-Score**	**Train Support**
Audio vs Picture	95.99	6974	79.35	8990
Living vs Not Living	96.31	4882	68.20	6324
Human Speech vs Music	96.52	2092	77.42	2697
Humans vs Animals	96.67	2790	87.90	3596
Tool vs Written Word	97.56	2092	90.32	2697
Emotional Vocalization vs Spoken Word	95.27	1395	83.87	1798
Faces vs Body	94.99	2092	90.32	2697
Object vs Checkerboard	97.63	1395	83.87	1798
Infant Face vs Adult Face	96.54	1395	86.79	1798

V. EEG Channels Reduction

This Section presents the approach ad-hoc designed to quantify the importance of each channel in recognizing the semantic category of stimulus associated with each ERP, with the aim of improving the users' comfort by reducing the acquisition setup.

It is known in the literature that PCA can make it difficult to interpret the contributions of each input attribute, as it produces a linear combination of the inputs. Similarly, fPCA can have similar interpretability issues. To address this problem, we design an approach to estimate the contribution of each channel in determining the first functional principal component. Specifically, for each grand averaged ERP related to the same stimulus we compute the correlation between each channel and the first functional principal component. Despite several formulation are provided to estimate correlation, we use Kendall τ coefficient, as it is non-parametric and more robust to noise [46]. The channels that are most correlated with the first principal component, considering all subjects, are considered the most relevant. Indeed, those channels are the most correlated to the first principal component, which is the one leveraged to extract the features proposed to the classifier to identify relevant patterns on which rely its predictive process. This procedure is repeated for each grand averaged ERPs in perception and imagery datasets separately.

Figure 7 shows the outcomes of the process, reporting the channels as colored dots, larger the most they are important. It turns out that, regardless the specific stimulus, central, dorso-lateral and centro-parietal brain area are the most considered by the hierarchical classifier to perform the recognition task. Last, it is also possible to see that the prefrontal cortex is most considered in perception than in imagery stimuli categories, where the centro-parietal is most involved. It is worth to mention that the correlation is computed over the whole frame

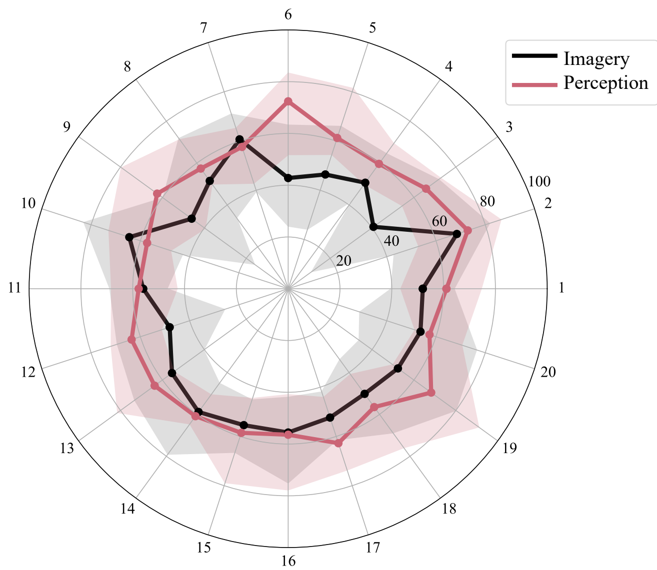


Fig. 5: Leave-One-Out Validation Results.
This Figure shows the average F1-Score assessed by each subject according to leave-one-out validation procedure.

of the grand averaged ERP considered, *i.e.*, between 100ms and 1000ms for the perception dataset and 400ms to 1000ms for the imagery. Therefore, the preponderance of frontal and central activity is also due to the cognitive processes related to the perception of the stimulus.

According to importance results it is possible to reduce the acquisition setup. Therefore, we first averaged the Kendall τ coefficient computed for each ERP. Then, according to the average correlation coefficient, we sort the channels, from the most to the least important. Table IX reports the 10 most important channels in perception and imagery datasets. In both datasets the most important channels' set includes electrodes of central, dorsolateral prefrontal/frontocentral and centro/parietal brain areas, *e.g.*, C1, C2, and Cz, CCP1h and CCP2h, FC2, and FFC1h, FFC2h, and FFC4h. The results of our study align with the physiological findings reported in [12], [25]. This result prove that our approach is able to identify important channels located in brain areas that are known to be involved in stimulus processing, without relying on prior knowledge. In more detail, many of the channels that are known to contain relevant markers according to previous studies were reported as important according to the presented approach. This is the case of C1, C2, Cz, CPz, FFC1h, FFC2h, AF2, AF3, AF4, AFz, Fz, FPz, P3, and P4. Exceptions are channels in the midline occipital and left occipitotemporal regions, including Oz, Iz, P7, P8, PPO9h, and PPO10h, which are known to be relevant for distinguishing between word, checkerboard, object, human, and animal face stimuli, but were among the least important for the hierarchical classifier. Nevertheless, our system was still able to accurately distinguish between these stimuli categories, even with the limited contribution of these channels.

TABLE IX: Channels' Importance. This Table reports the 10 channels that, one average, are most important to recognize the stimuli categories associated to the grand averaged ERPs in perception and imagery datasets.

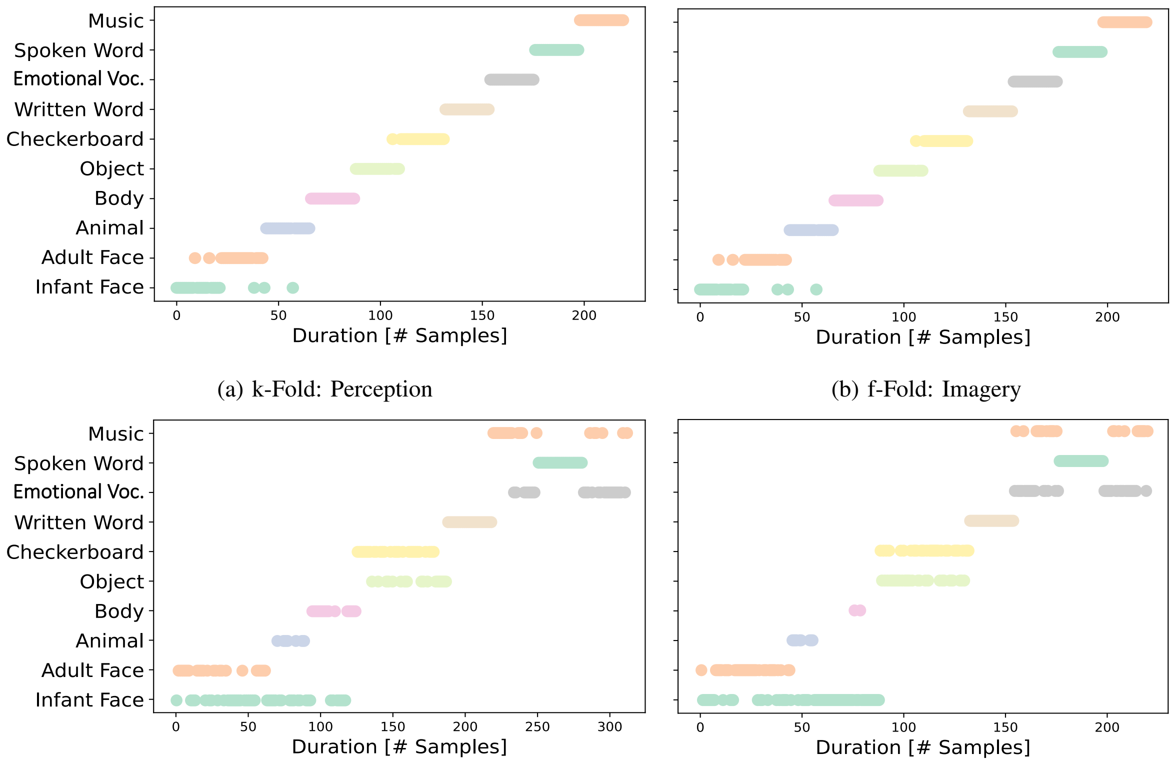
Perception		**Imagery**	
Channel's Name	**Kendall τ**	**Channel's Name**	**Kendall τ**
FC2	76.72	CCP2h	77.27
FFC3h	77.68	C2	77.02
CCP1h	78.30	C1	76.87
FFC4h	78.44	Cz	76.58
FFC1h	78.63	CCP1h	76.03
FFC2h	78.75	FFC2h	75.25
CCP2h	79.17	FFC4h	75.07
Cz	79.62	FFC1h	74.33
C1	79.79	FC2	74.01
C2	79.90	CCP3h	73.86

At this point, the hierarchical classifier has been iteratively re-trained and evaluated according to k-fold cross-validation by removing one channel at time, from the least to the most important one. Figure 8 reports the overall f1-score assessed by the hierarchical classifier on the perception and imagery datasets, respectively. In perception dataset, the first important channel alone, provides 74.32% f1-score; in imagery 80.09%. Also, in perception dataset to provide 90% f1-score 93 electrodes are required; in imagery 37 electrodes are enough to assess the same performance. According to the performance requirements of the specific application, this analysis provides valuable insights to design a cap which embeds only the minimum number of important electrodes.

VI. Conclusion

This study represents a significant advancement in the development of BCI and their potential applications in clinical settings. By leveraging FDA and machine learning techniques, we have demonstrated the feasibility of developing a passive BCI that recognizes specific semantic categories of stimuli based on grand averaged EEG data. Notably, our work focuses on recognizing imagery stimuli categories, which has not been addressed by previous studies. Additionally, our proposed approach maintains the temporal dependency of the EEG data, thanks to functional data representation and simplifies the stimuli categories recognition pipeline by means of an ad-hoc designed hierarchical ensemble classifier, resulting in high accuracy performance.

To validate the proposed approach, we extensively evaluated it on grand averaged ERPs evoked by stimuli belonging to 10 different semantic categories that have been perceived and imagined by 20 volunteers. Our approach achieved outstanding performance in classifying both perceived and imagery stimuli categories, with high accuracy rates of 96.37% and 83.11%,



(a) k-Fold: Perception

(b) f-Fold: Imagery

(c) Leave-One-Out: Perception

(d) Leave-One-Out: Imagery

Fig. 6: Predictions for Subject. This Figure compares the stimuli categories predicted for Subject 1's data by the hierarchical classifier when trained and evaluated according to k-fold cross-validation (top), and leave-one-out validation (bottom). Left subplots refer to perception dataset; right subplots to the imagery one.

respectively, according to k-fold cross-validation and hold-out validation.

The potential clinical applications of our work are promising, as it provides a new avenue for mind-reading and communication with patients who have neurological disorders, such as locked-in syndrome or coma. Future work will focus on optimizing the hierarchical classification architecture, expanding the range of stimuli categories, and evaluating its real-time performance to advance its practical applications.

References

[1] D. J. McFarland and J. R. Wolpaw, "Brain-computer interfaces for communication and control," *Communications of the ACM*, vol. 54, no. 5, pp. 60–66, 2011, doi: 10.1145/1941487.1941506.

[2] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of eeg-based brain–computer interface paradigms," *Journal of neural engineering*, vol. 16, no. 1, pp. 1–21, 2019, doi: 10.1088/1741-2552/aaf12e.

[3] K. J. Miller, D. Hermes, and N. P. Staff, "The current state of electrocorticography-based brain–computer interfaces," *Neurosurgical focus*, vol. 49, no. 1, pp. 1–8, 2020, doi: 10.3171/2020.4.FOCUS20185.

[4] S. K. Mudgal, S. K. Sharma, J. Chaturvedi, and A. Sharma, "Brain computer interface advancement in neurosciences: Applications and issues," *Interdisciplinary Neurosurgery*, vol. 20, pp. 1–8, 2020, doi: 10.1016/j.inat.2020.100694.

[5] A. T. Chan, J. C. Quiroz, S. Dascalu, and F. C. Harris, "An overview of brain computer interfaces," in *Proc. 30th Int. Conf. on Computers and Their Applications*, 2015, pp. 1–7.

[6] D. Zhu, J. Bieger, G. Garcia Molina, and R. M. Aarts, "A survey of stimulation methods used in ssvep-based bcis," *Computational intelligence and neuroscience*, vol. 2010, pp. 1–12, 2010, doi: 10.1155/2010/702357 .

[7] J. Polich, "Neuropsychology of p300," *The Oxford handbook of event-related potential components*, 2012, doi: 10.1093/oxfordhb/9780195374148.013.0089.

[8] A. M. Proverbio and A. Zani, "Electromagnetic manifestations of mind and brain," in *The cognitive electrophysiology of mind and brain*. Elsevier, 2003, pp. 13–40, doi: 10.1016/B978-012775421-5/50004-2.

[9] B. Z. Allison, E. W. Wolpaw, and J. R. Wolpaw, "Brain–computer interface systems: progress and prospects," *Expert review of medical devices*, vol. 4, no. 4, pp. 463–474, 2007, doi: 10.1586/17434440.4.4.463.

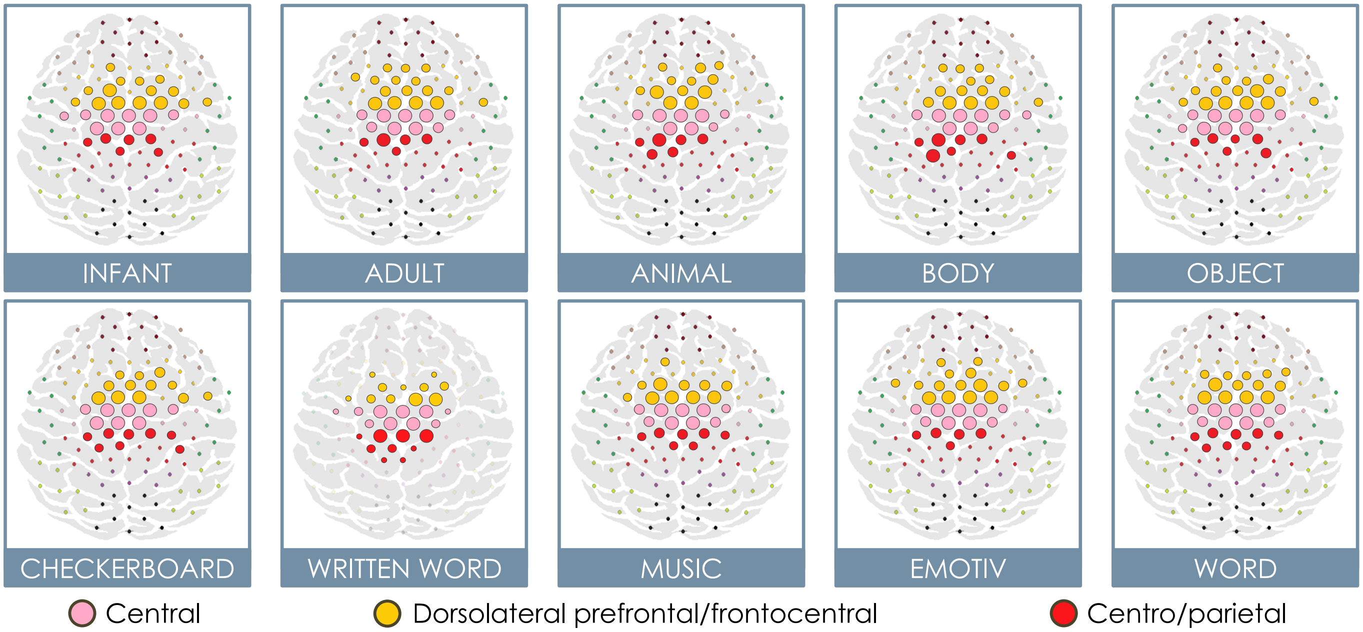
[10] Y. Li, C. S. Nam, B. B. Shadden, and S. L. Johnson, "A p300-based brain–computer interface: Effects of interface type and screen size," *Intl. Journal of Human–Computer Interaction*, vol. 27, no. 1, pp. 52–68, 2010, doi: 10.1080/10447318.2011.535753.

[11] B. Z. Allison, S. Dunne, R. Leeb, J. D. R. Millán, and A. Nijholt, *Towards practical brain-computer interfaces: bridging the gap from research to real-world applications*. Springer Science & Business Media, 2012.

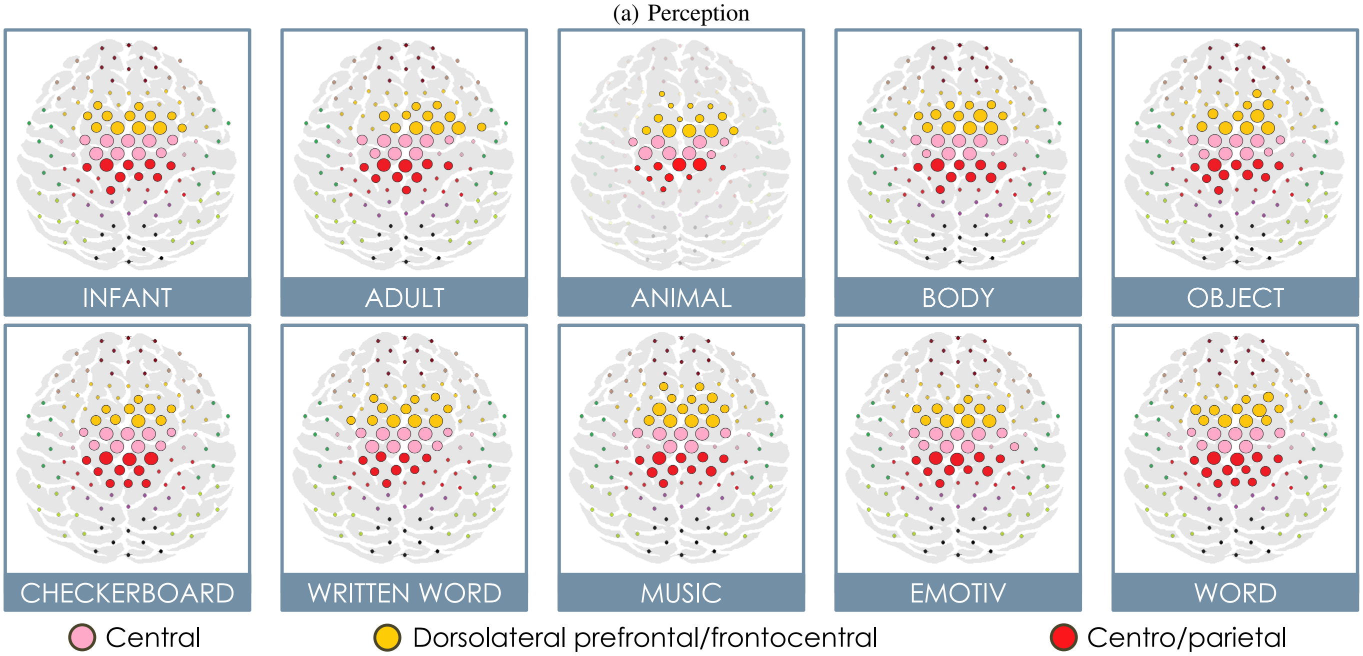
[12] A. M. Proverbio, M. Tacchini, and K. Jiang, "Event-related brain potential markers of visual and auditory perception: A useful tool for brain computer interface systems," *Frontiers in behavioral neuroscience*, vol. 16, pp. 1–22, 2022, doi: 10.3389/fnbeh.2022.1025870.

[13] S. Aggarwal and N. Chugh, "Review of machine learning techniques for eeg based brain computer interface," *Archives of Computational Methods in Engineering*, pp. 1–20, 2022, doi: 10.1007/s11831-021-09684-6.

[14] J. Leoni, S. C. Strada, M. Tanelli, K. Jiang, A. Brusa, and A. M. Proverbio, "Automatic stimuli classification from erp data for augmented communication via brain–computer interfaces," *Expert Systems with Applications*, vol. 184, pp. 1–14, 2021, doi: 10.1016/j.eswa.2021.115572.

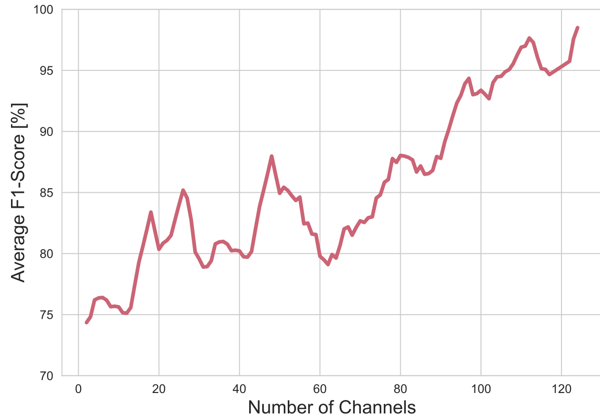


(a) Perception

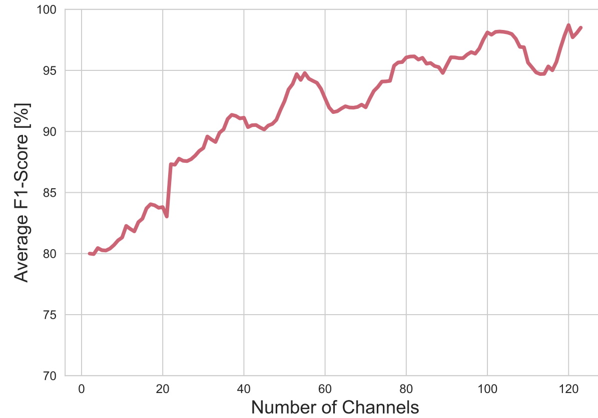


(b) Imagery

Fig. 7: Channels Importance.
This Figure shows, for each considered stimulus, the channels that, on average, are most correlated to the first fPCA component, *i.e.,* that are most considered in the classification process.



(a) Perception



(b) Figured

Fig. 8: Setup Semplification. This Figure shows the average F1-Score assessed on grand averaged ERPs classification as a function of the number of considered channels. Channels removal is performed according to their importance.

[15] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of erp components—a tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011, doi: 10.1016/j.neuroimage.2010.06.048.

[16] J. N. Mak and J. R. Wolpaw, "Clinical applications of brain-computer interfaces: current state and future prospects," *IEEE reviews in biomedical engineering*, vol. 2, pp. 187–199, 2009, doi: 10.1109/RBME.2009.2035356.

[17] J. Zhang and M. Wang, "A survey on robots controlled by motor imagery brain-computer interfaces," *Cognitive Robotics*, vol. 1, pp. 12–24, 2021, doi: 10.1016/j.cogr.2021.02.001.

[18] D. Huang, K. Qian, D.-Y. Fei, W. Jia, X. Chen, and O. Bai, "Electroencephalography (eeg)-based brain-computer interface (bci): A 2-d virtual wheelchair control based on event-related desynchronization/synchronization and state control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 3, pp. 379–388, 2012, doi: 10.1109/TNSRE.2012.2190299.

[19] D. Huang, P. Lin, D.-Y. Fei, X. Chen, and O. Bai, "Decoding human motor activity from eeg single trials for a discrete two-dimensional cursor control," *Journal of neural engineering*, vol. 6, no. 4, pp. 1–12, 2009, doi: 10.1088/1741-2560/6/4/046005.

[20] J. Pearson, "The human imagination: the cognitive neuroscience of visual mental imagery," *Nature reviews neuroscience*, vol. 20, no. 10, pp. 624–634, 2019, doi: 10.1038/s41583-019-0202-9.

[21] N. Dijkstra, S. E. Bosch, and M. A. van Gerven, "Shared neural mechanisms of visual perception and imagery," *Trends in cognitive sciences*, vol. 23, no. 5, pp. 423–434, 2019, doi: 10.1016/j.tics.2019.02.004.

[22] T. Naselaris, C. A. Olman, D. E. Stansbury, K. Ugurbil, and J. L. Gallant, "A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes," *Neuroimage*, vol. 105, pp. 215–228, 2015, doi: 10.1016/j.neuroimage.2014.10.018.

[23] A. R. Halpern, R. J. Zatorre, M. Bouffard, and J. A. Johnson, "Behavioral and neural correlates of perceived and imagined musical timbre," *Neuropsychologia*, vol. 42, no. 9, pp. 1281–1292, 2004, doi: 10.1016/j.neuropsychologia.2003.12.017.

[24] A. M. Proverbio and F. Pischedda, "Measuring brain potentials of imagination linked to physiological needs and motivational states," *Frontiers in Human Neurosciences*, vol. 17, 2023, doi: 10.3389/fnhum.2023.1146789.

[25] A. M. Proverbio, M. Tacchini, and K. Jiang, "What do you have in mind? erp markers of visual and auditory imagery," *Brain and Cognition*, vol. 166, pp. 1–11, 2023, doi: 10.1016/j.bandc.2023.105954.

[26] F. R. Llorella, G. Patow, and J. M. Azorín, "Convolutional neural networks and genetic algorithm for visual imagery classification," *Physical and Engineering Sciences in Medicine*, vol. 43, pp. 973–983, 2020, doi: 10.1007/s13246-020-00894-z.

[27] G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani, "End-to-end deep image reconstruction from human brain activity," *Frontiers in computational neuroscience*, vol. 13, pp. 1–11, 2019, doi: 10.3389/fncom.2019.00021.

[28] D. Salmaso and A. M. Longoni, "Problems in the assessment of hand preference," *Cortex*, vol. 21, no. 4, pp. 533–549, 1985, doi: 10.1016/S0010-9452(58)80003-9.

[29] H. F. Crovitz, D. Rosof, and H. Schiffman, "Timing oscillation in human visual imagery," *Psychonomic Science*, vol. 24, pp. 87–88, 1971, doi: 10.3758/BF03337905.

[30] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution eeg and erp measurements," *Clinical neurophysiology*, vol. 112, no. 4, pp. 713–719, 2001, doi: 10.1016/S1388-2457(00)00527-7.

[31] D. Regan, "Human brain electrophysiology," *Evoked potentials and evoked magnetic fields in science and medicine*, 1989.

[32] S. J. Luck, *An introduction to the event-related potential technique*. MIT press, 2014.

[33] T. W. Picton, S. Bentin, P. Berg, E. Donchin, S. Hillyard, R. Johnson Jr, G. Miller, W. Ritter, D. Ruchkin, M. Rugg *et al.*, "Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria," pp. 1–26, 2000, doi: 10.1111/1469-8986.3720127.

[34] M. Köppen, "The curse of dimensionality," in *5th online world conference on soft computing in industrial applications (WSC5)*, 2000, pp. 4–8.

[35] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, 2014, doi: 10.1039/C3AY41907J.

[36] J. Ramsay and B. Silverman, "Principal components analysis for functional data," *Functional data analysis*, pp. 147–172, 2005.

[37] P. H. Eilers and B. D. Marx, "Flexible smoothing with b-splines and penalties," *Statistical science*, vol. 11, no. 2, pp. 89–121, 1996, doi: 10.1214/ss/1038425655.

[38] P. Hall and M. Hosseini-Nasab, "On properties of functional principal components analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 109–126, 2006, doi: 10.1111/j.1467-9868.2005.00535.x.

[39] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[40] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2, doi: 10.1007/978-0-387-21606-5.

[41] K. Lorenz, *The natural science of the human species: An introduction to comparative behavioral research*. Mit Press, 1997.

[42] M. L. Glocker, D. D. Langleben, K. Ruparel, J. W. Loughead, R. C. Gur, and N. Sachser, "Baby schema in infant faces induces cuteness perception and motivation for caretaking in adults," *Ethology*, vol. 115, no. 3, pp. 257–263, 2009, doi: 10.1111/j.1439-0310.2008.01603.x.

[43] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, and S.-W. Lee, "Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy," *GigaScience*, vol. 8, no. 5, pp. 1–16, 2019, doi: 10.1093/gigascience/giz002.

[44] N. Birbaumer and L. G. Cohen, "Brain–computer interfaces: communication and restoration of movement in paralysis," *The Journal of physiology*, vol. 579, no. 3, pp. 621–636, 2007, doi: 10.1113/jphysiol.2006.125633.

[45] F. Nijboer, N. Birbaumer, and A. Kübler, "The influence of psychological state and motivation on brain–computer interface performance in patients with amyotrophic lateral sclerosis–a longitudinal study," *Frontiers in Neuropharmacology*, vol. 4, pp. 1–13, 2010, doi: 10.3389/fnins.2010.00055.

[46] N. S. Chok, "Pearson's versus spearman's and kendall's correlation coefficients for continuous data," Ph.D. dissertation, University of Pittsburgh, 2010.