Automatic stimuli classification from ERP data for augmented communication via Brain-Computer Interfaces

Jessica Leoni, Silvia Carla Strada
Politecnico di Milano, Dipartimento di Elettronica Informazione e Bioingegneria (DEIB), Milan, Italy
jessica.leoni@polimi.it, silvia.strada@polimi.it
Mara Tanelli
Politecnico di Milano, Dipartimento di Elettronica Informazione e Bioingegneria (DEIB), Milan, Italy
Istituto di Elettronica e Ingegneria dell'Informazione e delle Telecomunicazioni, Torino, Italy
mara.tanelli@polimi.it
Kaijun Jiang, Alessandra Brusa, Alice Mado Proverbio
Milan Center for Neuroscience, Università di Milano-Bicocca, Milan, Italy
k.jiang2@campus.unimib.it, a.brusa@campus.unimib.it, mado.proverbio@unimib.it

***Abstract*—Brain-Computer Interfaces (BCI) are technological systems that allow providing communication capabilities to individuals who may lack them for different reasons: communication-related neural inabilities, such as autism, or lock-in syndromes possibly due to illnesses or accidents. In all these situations, if neural activities exist, some form of communication can be enabled by an automatic interpretation of neural signals, which may lead to an interaction between the patient and the outside world. In this work, we prove such a path can be viable showing that ERP data can be suitably processed with machine-learning techniques obtaining a very good accuracy in discriminating among a detailed family of stimuli. Furthermore, we analyze if such an accuracy degrades as a function of the number of electrodes used in the process, with a look at what can be achieved with those available on commercial EEG wearables. The promising results obtained in this work open the way to the design of portable devices for augmented communication systems.**

I. Introduction

According to the definition given by Wolpaw et al. in [1], the term Brain-Computer Interface (BCI) refers to a communication system that allows the user to interact with a machine without using nerves or peripheral muscles. The system consists of the user, who transmits his commands via electrophysiological impulses, and the machine, which decodes and executes them. This exchange of information is enabled by a translation algorithm, the core of the system, whose task is to convert the electrophysiological impulses in a set of machine instructions, allowing the two actors to communicate efficiently and effectively.

Although this technology was born to allow people with severe motor disabilities to recover their lost autonomy, the recent progress achieved both from a technological and algorithmic point of view gives birth to countless general-purpose applications. Therefore, what was initially conceived as a means of compensation for subjects whose functionalities had been compromised, has now the potential to augment human brain capabilities connecting it directly to any electrical device. Thanks to the spread of researches in this field, the BCI market value is expected to grow at a Compound Annual Growth Rate (CGAR) of 11.1%, reaching 2.67 billion dollars by 2026, according to a report produced by Reports and Data [2].

The electrophysiological signal mainly used by this technology is the electroencephalogram (EEG) that represents the temporal evolution of brain electrical activity. This signal is usually detected with non-invasive techniques, based on the use of electrodes, properly applied on the scalp of the subject. The optimal number of electrodes depends on the final application. As reported by Farquhar et al. in [3], the best practice would be using as many electrodes as possible, to increase the spatial resolution. However, employing a large number of electrodes is feasible in contexts such as hospitals or research laboratories, while it is particularly unsuitable for general-purpose applications.

Each electrode measures voltage fluctuations resulting from ionic current within the neurons in the nearest brain region, so the set of acquired signals allows extracting information on both temporal and spatial evolution of the brain activity. However, each measured signal contains the product of the overall cerebral activity. Considering the number of tasks that the brain has to coordinate at the same time, such as the motion of the eyes, the regulation of breath and heartbeat, or the muscle activation, being able to recognize the stimulus that the subject intends to transmit directly from the EEG is a challenging task. Therefore, the BCI often uses event-related potentials (ERPs), which are the time-locked electrophysiological response to a certain stimulus. The ERP signal then reflects the average neural activity over several repetitions of a stimulus, synchronized over time respect to the triggering process. The subject is stimulated using a specific sensory, cognitive, or motor trigger, repeated multiple times. The ERP is obtained by synchronously averaging the multiple EEG

signals concerning the same stimulus. In this way, the zero-mean gaussian noise associated with the background brain activity, the other biological signals and to electromagnetic interference, is reduced, while the response associated with the trigger event, considered similar at each repetition of the stimulation, is emphasized. In addition to improving the signal-to-noise ratio, synchronous averaging techniques are used to make the ERP robust to the inter-subjective variability of long latency depolarization. To achieve this goal the averaging process is performed also considering the average ERPs related to the same stimulus but for different subject. This process is defined as *grandaverage*.

To date, at best of our knowledge, there is not in the literature an algorithm able to recognize more stimuli, of different nature, from EEG or ERPs signals detected by a pool of electrodes placed on the scalp. Just few models in the literature are able to operate multi-class classification, but they aim to differentiate several emotional states induced in a subject, rather than to identificate a specific stimulus. An example is proposed by Fan et al. in [4], who managed to distinguish five emotional states with an average accuracy greater than 75% leveraging the EEG signal recorded by 14 electrodes. In fact, the algorithms used in the BCI, operate binary classification. Their task is to recognize the presence or absence of the potential evoked by a specific stimulus, a priori known. In particular, the systems that exploit those algorithms are based on what is called the *oddball paradigm*, used for the first time in 1975 during researches on event-related potentials conducted by Squires et al. and reported in [5].

Following what reported by Krusienski et al in [6] and by Pfurtscheller et al in [7], the main algorithms used in the literature to recognize the presence of the potential evoked by a specific stimulus within an EEG or ERP signals are:

- Linear classifiers. This class includes algorithms such as linear discriminant analysis (LDA) and support vector machines (SVMs);
- Deep learning techniques. This class includes, in particular, all those algorithms that use artificial neural networks (ANNs).

However, both classes of algorithms mentioned above are affected by limitations. Linear classifiers are not able to capture complex correlations within the dataset and require a priori assumptions about the probability distribution of the classes to be separated. In particular, as explained by Blankertz et al. in [8], this last condition makes these algorithms applicable just if some restrictive conditions are respected. The artificial neural networks, instead, are characterized by complex architectures, and this makes particularly complicated to reconstruct the rules used in attributing an instance to a determined class. This problem cannot be considered of secondary importance. As Altmann et al. report in [9], the interpretability of a machine-learning model is as important as the accuracy of the prediction itself.

Therefore, aiming to overcome these limitations, we propose:

- Two classification models, able to perform a multi-class distinction, recognizing 14 stimuli of different nature processing and analyzing the EEG signal detected by 126 electrodes;
- A ranking method for the criteria adopted in the decision-making process, to allow its easy interpretation. The ranking is produced according to the importance that each feature has in determining the predictions;
- A decision method, that allows choosing the most suitable representation of the data to optimize the performances in terms of accuracy of the predictions and interpretability of the decision-making process.

In addition, particular attention has been paid to adapt our algorithm also to the acquisition devices for general-purpose applications already available on the market. So, we analyze the prediction error's trend reducing the number of employed electrodes, aiming at identifying the value that optimizes the trade-off between accuracy in performing the classification task and cumbersomeness of the device.
The rest of the document is structured as follows: Section II provides an overview of the results reported so far in the literature concerning the algorithms used in BCI systems, Section III defines the statement of the problem addressed. Section IV explains the experimental setup used for data acquisition, while Section V illustrates the pre-processing, features extraction and classification phases performed. Section VI: shows the results obtained, which are then discussed in Section VII. Finally, Section VIII reports the conclusions and the future developments for the proposed work.

II. Related Works

This section aims to present the main results reported in the literature related to classification algorithms used in BCI.

The classification methods used so far are concerned with solving binary problems. Their purpose is to identify in the signal proposed the presence of a characteristic depolarization, caused by a specific stimulus, a priori known. Considering this classification task, as reported by Kubler in [10], many authors, to guarantee meaningful communication between man and machine, consider 70% as the minimum acceptable accuracy for those algorithms.

Classification techniques can be further distinguished on the basis of the evoked potential they intend to recognize. Among these, the large positive deflection usually recorded after 300ms post-stimulus, the P300 component, which reflects cognitive encoding, working memory and attentional processes, such as target discrimination, certainly has the prevalence. This potential can be triggered by systems such as the P300 speller, proposed by Farwell and Donchin in 1998 and illustrated in [11].

The particular attention to the identification of the P300 is also due in part to the second edition of the BCI competition, held in 2003. In fact, one of the proposed challenges asked the participants to correctly classify the instances contained in a dataset properly collected by the Institute of Medical

Psychology and Behavioral Neurobiology of the University of Tübingen in Germany. The dataset contains the EEG signal measured by a pool of 64 electrodes, with a sampling frequency of 240Hz, during several runs of P300 spellers [12]. Among the various researchers that have analyzed this dataset, the state of the art is given by the work proposed by Rakotomamonjy et al. in 2015, and retrievable at [13]. In particular, in [14], signals from just 10 electrodes are considered.The classification was performed leveraging SVMs. Cecotti et al. in [15], propose to perform the same classification task leveraging convolutional neural networks. Finally, in [16], Zhan et al. propose to consider the signals collected from only 16 of the electrodes provided and, after a filtering phase, they perform the classification task leveraging spatial-temporal LDA. Other works that still operate using the P300 speller are based on different datasets collected independently by the researchers.

However, the systems based on experimental setup as the P300 speller are somehow in contradiction with the very definition of BCI, as they assume that the subject is able to control the musculature responsible for eyes' movement. For this reason, BCI systems have been developed to recognize different type of stimuli, which do not require any muscular control. For example, in the work of Lopez-Gordo et al. [17], the stimulus was auditory, based on the dichotic listening task, described by Meyer et al. in [18]. This type of recognition exploits the innate activity of the human brain of being able to focus its attention on a precise sound source, filtering all the others. Since a classical example of this phenomenon is the ability of the person to concentrate on the voice of a specific interlocutor in a crowded room, it is widely known in literature as the *cocktail party effect*. Research on it was conducted by Squires et al., Bronkhorst et al. and Shinn-cunningham et al., respectively [5], [19], and [20].

Finally, the last example was provided in 1995 by Gupta et al. [21]. In this study, the evoked potential is triggered by error-related negativity feedback (ERN), i.e. by proposing to the subject a mistaken or unusual event, and it was measured from six electrodes. In particular, they matched images representing different objects to a vocally reproduced name. In some cases, this corresponded to the object depicted in the image, while in others it did not. Depending on whether the condition was match or non-match, the measured evoked potential assumed different shapes that can be recognized leveraging neural networks.

III. Problem Statement

In this paper, we propose a solution that goes beyond the ones widely discussed in the literature. We provide an innovative contribution in the field of research on BCI technologies, proposing two classification algorithms, one based on boosted trees and the other on the use of the artificial neural networks. Both of them are:

1) Multi-class, able to recognize 14 different stimuli, both visual and auditory, used to trigger a subject's neural reaction, starting from the a priori non known evoked potential that they generate;
2) Free from a priori assumptions. Both models are based on deep learning methods, which do not require any definition of the probability distribution for the classes of interest;
3) Easy to be interpreted. A ranking of the features used by the classifier is returned, based on the importance they had in determining the prediction. Moreover, both models have been suitably combined in a hierarchical structure. In this way, the final class predicted is the result of a series of binary classifications, in which the stimulus is progressively differentiated.

In addition to the process of classification itself, particular attention has been dedicated to the preliminary phases of signals' pre-processing and features extraction. In particular, the innovative contributions given are:

1) A method of debiasing that, at best of our knowledge, has not yet been applied in the literature. This method aims to subtract from the ERP time-series the mean value of the first 100ms of recording. Since this time window is immediately prior to the stimulus, it contains in fact only noise;
2) A rationale to extract only the portion of the ERP signal which is informative for classification purposes;
3) The comparison of different data representations, related to temporal, spatial and morphological characteristics of the measured ERPs;
4) The explicit evaluation of the prediction error based on the number of used electrodes, to identify the minimum number of electrodes for high classification performance. This is key in view of a future implementation of our methods to acquisition devices on the market.

IV. Experimental Setup

Twenty-one university students (10 men and 11 women) ranging in age from 19 to 30 years (mean age=23 years) participated in the study as unpaid volunteers. All of them had a normal or corrected-to-normal vision and normal hearing threshold. They were strictly right-handed as assessed by the Oldfield Inventory (mean score = 0.88; SD = 0.13) and reported no history of drug abuse or neurological or mental disorders. Experiments were conducted with the understanding and written consent of each participant according to the Declaration of Helsinki (BMJ 1991; 302: 1194), with approval from the Ethics Committee of the University of Milano-Bicocca (prot. N°. RM-2019-193).

A. Stimuli

To each of the 21 subjects, 14 different stimuli belonging to three different macrotypes were presented, without a precise order. Each stimulus has been presented to each subject 40 times. The static pictures stimuli set was constituted by nine different stimulus categories (40 pictures for each category), and target landscapes (36 pictures). The auditory set was constituted by a total of 40 words, 40 emotional vocalizations,

40 piano musical fragments, and 12 natural sounds that acted as rare targets. The dynamic visual stimuli set included 40 biological movements and 40 mechanical movements, plus eight videos of natural scenarios that acted as targets. All images had a white background, and the stimuli were presented in the center of the screen. For the living static stimuli, they were balanced for gender and are composed by infant, adult, and animal faces and dressed bodies. The non-living static category includes tools, objects and checkerboards, with different colors. Linguistic stimuli included 20 nonsense letter strings, and 20 italian words. These images were presented in random order, randomly mixed with the 36 landscape pictures. Each slide was presented for 1500 ms with an inter-stimulus interval (ISI) varying between 900-1000ms. The outer background was black. Auditory stimuli also lasted 1500 milliseconds. They included 40 utterances of Italian words (20 spoken by a woman, and 20 by a man), 40 emotional vocalization, and 40 piano musical fragments were also presented. The targets for auditory stimuli were natural sounds. All the auditory stimuli were normalized and leveled in intensity. All the videos were cut to the same duration, 1500 ms. There were 40 mechanical movements (e.g., moving trains, elevators, helicopters, cars) and 40 biological movements, which were also balanced by gender. The depicted actors performed actions such as gymnastics, hand moving, and so on.

B. Procedure

Participants comfortably sat in a faradized and acoustically shielded cubicle in front of a PC monitor located at 114cm from the subject's eyes. They were asked to fixate the center of the screen where a red dot served as a fixation point. Static stimuli were presented in random order at the center of the screen in eight different randomly mixed short runs lasting approximately 2 min and 5s. Auditory stimuli were presented in random order in three different randomly mixed short runs lasting approximately 1 min and 50s. Video stimuli were presented in random order at the center of the screen in two different randomly mixed short runs lasting approximately 1 min and 50s. To keep the subject focused on the visual stimuli, the task consisted of responding, as accurately and quickly as possible, to target photos or videos or sounds by pressing a response key with the index finger of their left or right hand.

C. EEG Recording and Analysis

The EEG was continuously recorded from 126 scalp sites (ANT Software, Enschede, The Netherlands) at a sampling rate of 512 Hz using tin electrodes mounted in an elastic cap (Electro-Cap) and arranged according to the international 10–5 system defined by Oostenveld and Praamstra in [22]. Horizontal (hEOG) and vertical (vEOG) eye movements were also recorded. Linked mastoids served as the reference lead. The EEG and EOGs were filtered with a half-amplitude bandpass of 0.016–70Hz. Electrodes' impedance was maintained below 5KOhm. EEG epochs were synchronized with the onset of stimulus presentation. Computerized artifact rejection was performed prior to averaging. The artifact rejection criterion was a peak-to-peak amplitude exceeding 50μV. This procedure resulted in a rejection rate of about 5%.

V. Proposed Methodology

In this section, we illustrate the phases of pre-processing and features extraction that led us to the definition of five data representations. Subsequently, we explain the design of the classification model produced to recognize the stimulus from the ERPs it generates.

A. Synchronous Averaging

Synchronous averaging is a process that is applied to enhance the signal components, reducing the presence of te superimposed noise. The noise is composed by the background brain activity, the electromagnetic interference that affects the transmission and other biological signals, as electromiogram and electrooculogram. Syncronous averaging process is based on two assumptions:

1) The evoked potential has invariable latency and shape between trials;
2) The noise is a zero-mean gaussian random process of variance σ^2, uncorrelated between trials.

Therefore, considering the case of a single electrode placed on the scalp of a subject for K trials, the signal measured in each k-th trial will be constituted by the contribution of both the ERP, s(t) and the noise, n(t,k).

$$x(t,k) = s(t) + n(t,k) \tag{1}$$

Notice that the ERP depends only on time, as it is considered equal to itself at each iteration, while noise depends both on time and on the iteration considered.

The mean value of the signal measured in N trials will be

$$\bar{x}(t) = \frac{1}{N}\sum_{k=1}^{N} x(t,k) = s(t) + \frac{1}{N}\sum_{k=1}^{N} n(t,k) \tag{2}$$

and its variance is $\frac{\sigma^2}{N}$.

As the noise is considered a zero-mean gaussian process, for N tending to infinity, the signal obtained as output of the synchronous averaging process would contain only the ERP. Averaging is a techniques vastly leveraged in the literature to improve signal to noise ratio (SNR) and its effects is even more beneficial as the number of averaged trials increases. .

Let us define $x_{i,j,k,el}$ as the evoked potential time-series measured by the el-th electrode placed on the scalp of the j-th subject in response to the k-th repetition of the i-th stimulus. Therefore $i = 1, ..., 14$ is number of stimuli, $j = 1, ..., 21$ is number of subjects considered in the study, $k = 1, ..., 40$ is number of repetitions of the stimulus on each subject and $el = 1, ..., 126$ is the number of electrodes placed on the scalp. Each time-series contains 820 samples, as the electrodes measure the evoked potentials from -100ms to 1500ms, for a total duration of 1600ms, at a frequency of 512Hz.

The synchronous average was first performed on the evoked potentials measured during the different repetitions of the same

stimulus to the same subject. The achieved signal, $x_{i,j,el}$, represents the ERP triggered by the i-th stimulus on the j-th subject, measured by the el-th electrode.

$$x_{i,j,el} = \frac{\sum_{k=1}^{40} x_{i,j,k,el}}{40} \tag{3}$$

Subsequently, a second synchronous averaging process was performed to obtain the grandaverages. In this way we obtain the average potentials caused by the i-th stimulus and measured by the el-th electrode, defined $x_{i,el}$.

$$x_{i,el} = \frac{\sum_{j=1}^{21} x_{i,j,el}}{21} \tag{4}$$

Each X_i is the set composed by 126 time-series, and describes the average potential evoked by the i-th stimulus and detected on the scalp by each of the electrodes used. These trends are robust to intra-individual and intra-stimulation variability, as they are obtained as averages of single trials.

B. Data Representations

The dataset consists of the 126 ERPs related to 14 different stimuli. Since the sampling frequency of the electrodes is 512Hz and each time-series has a duration of 1600ms, each ERP is composed of 820 samples. The measurements of four out of the electrodes, M1, M2, EOGH, and EOGV, were not considered relevant to perform the classification task and therefore were discharged. This is because M1 and M2 are the reference electrodes, while EOGH and EOGV measure the ocular movement, which is not informative for this application. Therefore, the actual number of considered electrodes is 122.

The collected time-series have been organized in more structured datasets. In particular, the three approaches proposed consist of a spatial representation, a temporal representation, and a representation based on the morphological characteristics of the ERPs. Moreover, of the first two representations, both a naive version, in which all the 820 registered samples are considered, and a cut version, in which is considered only the informative portion were built. The informative portion starts at 0ms, synchronously with the stimulus' onset, and ends at 900ms, when it is already exhausted. Notice that the stimulus recording starts at -100ms, but the stimulus is triggered at 0ms. All representations have been individually proposed to both the classification algorithms. The performances obtained were then compared in terms of prediction accuracy and interpretability of the results, to determine the optimal representation.

1) Spatial Representation: In spatial representation, each feature corresponds to the measurements recorded by the same electrode. Therefore, the number of features is 122. The number of instances is equal to the total number of samples measured by a single electrode. Each instance represents the spatial distribution of the cerebral electrical activity in a fixed time instant. Each feature contains the information related to the activity measured in a specific area of the scalp, i.e. the area adjacent to the corresponding electrode.

There are two versions of this representation; in the naive one, all the measured samples have been maintained, while in the cut one just the significant portion of the ERPs, i.e. the one between 0ms and 900ms, has been considered.

2) Temporal representation: In the temporal representation dataset, each feature contains all the samples measured at the same time instant. The number of instances is instead equal to the total number of ERPs measured, which is 1708 instances. Each instance represents the temporal trend of an ERP related to a precise stimulus in a fixed area of the scalp, i.e. the area adjacent to the considered electrode. Also for this type of representation , the same as above, two versions have been built. Therefore, in the first one, each ERP is composed of 820 samples, so the features are 820, while the second one is composed of 450.

3) Formal Representation: In the formal representation, the main morphological characteristics were extracted for each ERP. In particular, we extracted eight features that synthetically represent the entire time-series. The number of instances contained in this dataset is the same got for the temporal representation, i.e. 1708.

In particular, the features extracted for this representation are:

- Time instant of the maximum and of the first relevant peak, respectively $POS_{max_{i,el}}$ and $POS_{first_{i,el}}$;
- Height of maximum and of first relevant peak, i.e. the value assumed by the ERP in the maximum and first relevant peak positions, respectively.

$$HEIGHT_{max_{i,el}} = x_{i,el}(POS_{max_{i,el}}) \tag{5}$$

$$HEIGHT_{first_{i,el}} = x_{i,el}(POS_{first_{i,el}}) \tag{6}$$

- Area underneath the relevant portion of the ERP. It was calculated by integrating the ERP according to Simpson's rule, between 0ms and 900ms;
- Mean value and variance of the ERP.

$$mean_{i,el} = \frac{\sum_{i=1}^{450} x_{i,el}}{450} \tag{7}$$

$$var_{i,el} = \frac{\sum_{i=1}^{450} (x_{i,el} - mean_{i,el})^2}{449} \tag{8}$$

Since the features extracted for this representation are computed only on the relevant portion of the ERPs, a single version was produced.

C. Labelling Process

The labels associated with each ERP correspond to the stimulus that evoked it. These have been organized by us according to a hierarchical structure, creating intermediate classes. The final tree we produce for the labeling process has a total of 13 split nodes and it is reported in figure 1. So, a specific stimulus is identified progressively, along a path with subsequent binary divisions.

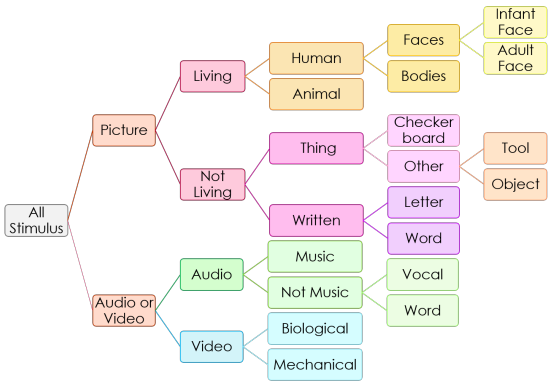

Fig. 1. Label subdivision.
This hierarchical structure has been specifically designed to facilitate the classification process.

D. Time-series Debiasing

The representations produced are therefore five: spatial and temporal, in the naive and cut versions, and formal. All have been subjected to a process of debiasing. This consists of subtracting from each ERP the average value of its first 100ms of recording (50 samples), in which the stimulus has not yet reached the cortex. This value, which in the ideal case should be zero, in real applications represents the bias due to the superimposed noise. The pseudocode of debiasing performing is reported in Algorithm 1.

Algorithm 1: Time-Series Debiasing Process

Data: Raw dataset
Result: Dataset debiased from the random noise superimposed to the ERPs

$data \leftarrow X = [x_{1,1}, x_{1,2}, ..., x_{i,el}, ..., x_{14,122}]$;
$new\_data \leftarrow []$;
for $x_{i,el}$ *in data* **do**
  $time\_series \leftarrow x_{i,el}$;
  $bias \leftarrow mean(time\_series[0:50])$;
  $time\_series \leftarrow time\_series - bias$;
  $new\_data.append(time\_series)$;
end for
return new_data;

E. Classification Methods

The classification procedures to recognize the responsible stimulus for the ERP were two: boosted trees and feed-forward neural networks. In both cases, final model is composed of 13 binary classifiers combined following the hierarchical structure produced for the labels, so that each classifier corresponds to one label split.

1) Boosted Trees-Based Classification: A decision tree is a predictive model in which the branches contain a set of rules, called split conditions, which progressively lead to the leaves, i.e. the classes to be predicted. A model based on boosted trees sequentially combines several shallow decision trees to increase the accuracy and the reliability of the prediction. This approach has been developed in response to the observation made by several researchers, such as Firedman et al in [23], who argue that, for most machine-learning problems, the output of a single classifier can not be considered reliable. In a boosted trees-based model, each tree is constructed to reduce the errors made by the previous one. The set of functions used by the model to predict the classes are calibrated during the training phase to minimize the following objective function

$$L(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} (\Omega(f(k))) \tag{9}$$

where n is equal to the number of instances contained in the training dataset, K is equal to the number of trees that compose the model and f(k) is equal to the function that represents the k-th tree in the model. The first term is the loss function, while the second one is introduced to penalize the complexity of the model and avoid over-fitting.

The specific boosting algorithm leveraged in our approach was XGBoost, considered the current state of the art because efficient [24].

Different hyper-parameters must be optimally tuned during the fine-tuning phases. According to its results, we decide to set the learning rate, i.e. the step size shrinkage to prevent over-fitting, to 0.03, the percentage of features considered by each tree to 80%, the regularization term for leaf weights, called alpha, to 10 and the number of trees in each boosted trees-based model to 13. Finally, the binary logistic is set as the objective function.

For each boosted trees-based model is also returned the features importance, which is a ranking of the features, based on their impact in determining the prediction. The metric used to compute the ranking is called F-score and is calculated as the ratio between the number of times that a feature is used as a split criterion and the total number of splits.

2) Feed-Forward Neural Networks-based Classification: Artificial neural networks are a set of algorithms able to calibrate themselves autonomously in order to map each input to the correct category. In particular, the feed-forward neural networks constitute a subclass of neural networks in which the input flows just once through the network and in one single direction.

Given in input to the network a vector X composed by n components

$$X = [x_1, x_2, ..., x_i, ..., xn]$$

the output of each node is calculated as

$$\hat{y}_i = b + \sum_{i=1}^{n} x_i * \omega_i \tag{10}$$

where ω_i is the weight associated with the i-th branch. Weights are calibrated during network training leveraging a back-propagation algorithm. The term b, the bias, is instead introduced to reduce the possibility of over-fitting, shifting the activation function to the right or left.

As for the models based on boosted trees, also neural networks-based ones require several hyper-parameters to be appropriately calibrated. Again, their value was decided after a fine-tuning phase. According to its results, the number of hidden layers in the final feed-forward neural network was set to one, and the number of hidden units to 64. As objective function for the intermediate nodes, we chose to adopt the mean squared error (MSE), while as activation function we chose the rectified linear unit (ReLU).

As we said, neural networks are known to have complex architectures, which make it difficult to reconstruct their predictive process. To overcome this problem, we leverage a criterion based on the approach proposed by Breiman in [25]. His idea was to measure the importance of a feature in the decision-making process considering the accuracy drop caused by its removal from the dataset. The more significant one feature is for prediction purposes, the greater the accuracy loss caused by its removal. However, since the original model is trained on all features, it is not possible to completely remove one from the test dataset. So, we randomly change its values, introducing a noise that makes the series no more informative. Since the importance of a feature in the prediction process is estimated by calculating the drop of the accuracy caused by its permutation, the ranking returned is called permutation importance.

VI. Experimental results

The performances obtained by the two classification models applied to the five data representations, pre and post debiasing, have been compared in terms of prediction accuracy and decision-making process interpretability.

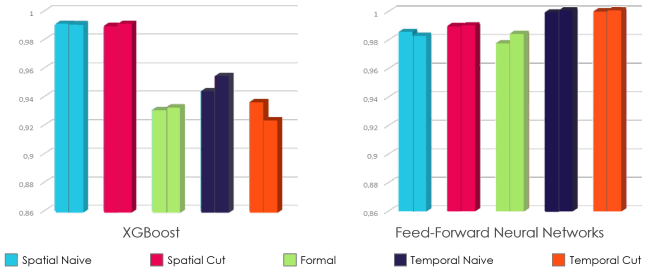

Fig. 2. Boosted trees and Feed-forward neural networks-based model's accuracy.
The figure shows the performances in terms of accuracy obtained by the two classification models on the five representations proposed, before (left column) and after (right column) the debiasing.

The average results are shown in Figure 2. It turns out, considering all the proposed representations, that ERPs debiasing process improves the accuracy of the prediction. Moreover, cutting the signal, improves the ability of the model to identify the correct classes. As for the boosted trees-based model, the best accuracy, equal to 99.02%, is obtained with the spatial representation, after ERPs debiasing and cut. Using the model based on feed-forward neural networks, on the other hand, the temporal representation appears to be the best choice to achieve the optimal accuracy performances. Operating ERPs debiasing and cut, the average accuracy for the single feed-forward neural networks-based model that composes the hierarchical classification design is 99.9%.

Models' performances has also been assessed at each different level of the hierarchical structure. Figure 3 shows the results obtained in terms of prediction accuracy for the boosted trees-based model applied to the spatial representation of ERPs, subjected to cut and debiasing. As it was reasonable to assume, the accuracy decreases in proceeding to the differentiation of the stimulus, always remaining above 95.0%. The same trend is observed for both models, regardless of the type of data representation used.

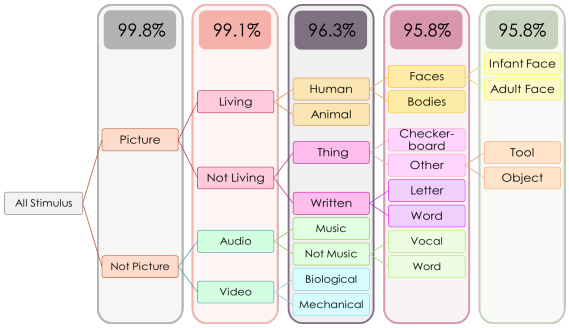

Fig. 3. Boosted trees-based model's accuracy along the hierarchical classification structure.
The figure shows the performances in terms of accuracy obtained by the boosted tree-based model on the spatial representation for the cut ERPs, after the debiasing.

In addition to the accuracy of the predictions, the interpretability of the decision-making process was also chosen as a key performance indicator. Therefore, we considered the features importance indicator for the model based on boosted trees and the permutation importance for the model based on the feed-forward neural networks.

The results produced considering the formal data representation show that there is a perfect coherence between the rankings returned by the two models.

A visual representation for the main features used in the decision-making process respect to the spatial cut representation is shown in Figure 4. Also in this case, it is possible to see that both methods rely at most on coherent electrodes.

Considering the features and permutation importance for temporal data representation, it turns out that they consider different portions of the time series given in input to learn their classification rules. In fact, while the boosted trees-based model mainly considers the initial portion of the cut signal, the feed-forward neural networks weights the whole time-series.

VII. Discussion

As explained in Section II, it is fundamental to produce models whose decision-making process is easy to be interpreted. So, we choose as the most suitable data representation the spatial one. Even though the accuracy of the boosted trees-based model for the spatial representation was comparable to the one achieved by the feed-forward neural networks-based

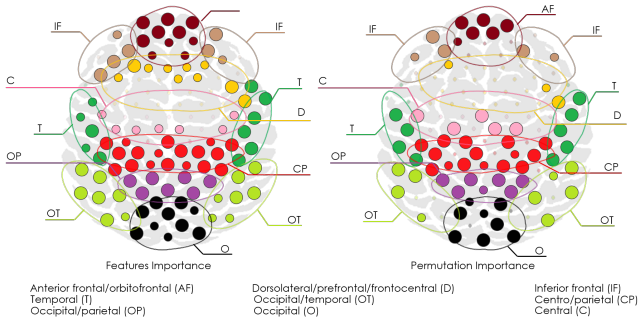

Fig. 4. Feature importance in spatial data representation.
The figure shows the ranking regarding the importance that the proposed electrodes assume in the decision-making process of the respective model. Importance of an electrode is directly proportional to the diameter of the corresponding spot. It turns out that, both of them the split rules considering the frontal, temporal, occipital parietal electrodes, and neglecting the dorsolateral, prefrontal, frontocentral and central ones.

model considering the temporal representation, the features considered in the first one are more interpretable and significant for the expert. In fact, as shown in Figure 4, they highlight the scalp topographical areas mostly involved during the decision-making process. Assuming that the electrodes in the lowest positions of the ranking can be overlooked, this ranking is a key information in the perspective of reducing the number of electrodes used. The features ranking produced considering the temporal data representation provides information about the portion of the input signal most relevant to identify the stimulus. However, the results returned by the neural networks-based model shows that the network considers the importance equally distributed along the input signal, and so it does not provide any useful information.

As far as the formal representation is concerned, it is based on compact features easy to be interpreted. In fact, they are the same used by experts to recognize the stimuli related to the ERPs. Also, the results obtained in terms of features and permutation importance are consistent with those expected, as the position and height of the maximum depolarization, the mean value and the area subtended by the ERPs are the most relevant morphological characteristics taken into account while performing the identification process. However, this representation is also the one that leads to lower levels of accuracy. This loss in terms of accuracy performances shows that, letting the classification algorithm to autonomously extract information from the proposed signals, leads to better results. However, the improvement in performance due to the cut and debiasing of the signal shows that a pre-processing phase based on prior knowledge and experience of the data analyst is still relevant to improve the overall performance. Therefore, we can conclude that both models are capable of extracting useful characteristics to optimally perform the classification task, provided that a proper signal pre-processing is performed.

A. EEG Acquisition Devices for General-Purpose Applications

As exposed in a report produced by Report and Data, retrievable at [2], in 2018 non-invasive BCI products accounts for the largest share of 38.8% of the total BCI market. The growth of this sector is due to the general-purpose solutions that are recently taking root alongside classical clinical applications. Generally, using as many electrodes as possible maximizes the spatial resolution of the measured ERPs. However, when it comes to general-purpose applications, outside of hospital or laboratory environments, a large number of electrodes applied to the subject's scalp would make the device cumbersome.

The solutions designed to extend BCI also to general-purpose applications can be divided into two categories. On the one hand, companies such as Neuralink [26] and Neuropace [27] propose to change the hardware used for acquisition. Therefore, they aim to develop neural interfaces that are easily implantable through surgical operations on any individual. Thus, they do not need to reduce the number of electrodes. On the other hand, groups of researchers try to solve this issue at an algorithmic level, aiming at obtaining high classification performances from the reduced pool of electrodes embedded in the devices already available on the market. This would allow the costumer to avoid an invasive procedure as a surgical operation, to which several risks are connected. This group of solutions includes the work we propose in this paper. After a careful analysis of the non-invasive EEG acquisition devices currently on the market, we identify two categories:

1) Soft electrode helmets. This category includes the devices belonging to the *Stat x-series*, produced by Advanced Brain Monitoring [28], and *Epocflex*, an acquisition device produced by Emotiv Systems [29]. The devices of the *Stat x-series* acquire professional and cost-effective recordings, also applicable for medical purposes. Depending on the model considered, the acquisition helmet is composed of 10 or 24 electrodes sampling at 256Hz, arranged according to the 10-20 standard system. The recorded data is stored in the memory of the device itself from where it can be downloaded for offline analysis. Epocflex, instead, is designed for general-purpose applications and consists of a soft helmet that embeds 32 electrodes sampling at 128 Hz. To be analyzed, the acquired data is sent via Bluetooth to a computer.
2) Electrode bands. This group includes devices as *Epoch+* and *Insight*, both produced by Emotiv. The first device consists of a head-band that embeds 14 electrodes, while the second one is composed of just 5 electrodes. In both devices, the sampling frequency of each electrode is 128Hz and the acquired data can be sent via Bluetooth to any electronic device, including smartphones and tablets.

First, we evaluated the prediction error's trend as a function of the number of electrodes. The results obtained for the model based on boosted trees applied to the spatial representation

of ERPs, subject to cut and debiasing, are shown in the Figure 5. From the results obtained it is possible to observe that with 100 electrodes the prediction error at the leaves is less than 10%, while up to 35 electrodes is below 20%. Besides, for some applications, a lower degree of detail in the characterization of stimuli may be sufficient. As an example, if it is enough to discriminate only the nature of the stimulus that caused the ERP, distinguishing between static picture, auditory or dynamic visual, 10 electrodes are sufficient to ensure a maximum error smaller than 5%.

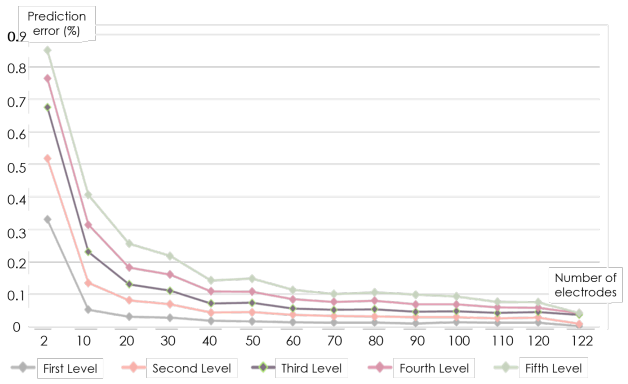

Fig. 5. Prediction error trend of boosted trees-based model varying the number of employed electrodes.

The, we simulated the performance of our algorithms considering only the electrodes embedded in the devices on the market. Besides, we also carried out the classification considering the subset of ERPs measured by the 35 electrodes whose importance was greater in the ranking produced during our models evaluation.

The results obtained in terms of prediction error for the boosted trees-based model applied to the spatial representation of the ERPs, subjected to cut and debiasing, are shown in the Figure 6. We considered six different market devices that embed different number of electrodes, placed in different positions. The prediction error increases when considering devices with a reduced number of electrodes, except for EpocFlex. In fact, although it has a higher number of electrodes than Stat X24, it has a higher prediction error at all levels. It can be assumed that this is due to the positioning of the electrodes, which is more congenial for our classification purposes in Stat X24 rather than on EpocFlex. Regardless of the device considered, the prediction error increases with the level of the hierarchical classification.

As stated in Section II, the minimum accuracy threshold to guarantee meaningful communication between man and machine is considered equal to 70%. The results obtained in our simulations show that a device with 35 electrodes in the first ranking list locations would guarantee an accuracy of more than 80% at all levels of the hierarchical structure. Moreover, also the accuracy obtained using as input for the classifier the measurements recorded by the electrodes embedded in Stat X24 and in EpocFlex is greater than the minimum threshold at all levels of the hierarchical structure. As far as Epoc+ is concerned, the accuracy of predictions that can be obtained satisfies the requirement up to the fourth level, while Stat10 and Insight up to the third. So, the achieved results shows that our algorithms are compatible with the EEG acquisition devices actually available on the market.

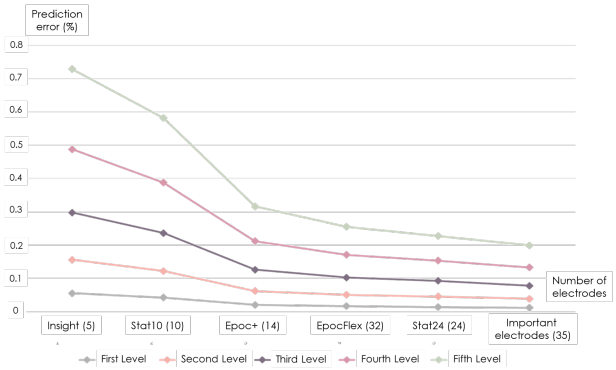

Fig. 6. Prediction error trend of boosted trees-based model varying the acquisition devices.
The figure shows the prediction error trend respect to acquisition devices considered and also considering the 35 electrodes characterized by the highest features importance.

VIII. Conclusion

In this work, we have proposed two ERP-based classification approaches to recognize one among 14 different visual and auditory stimulus. Both models have been realized by combining, in a hierarchical structure, several instances of the same classifier, either considering a boosted trees-based or feed-forward neural networks-based one. In both cases, the classifiers hyper-parameters were carefully selected according to the results of a fine-tuning phase.

As a case study, signals by 122 electrodes, operating at 512Hz, were recorded on 21 subjects. Each subject was triggered with 40 repetitions of 14 different stimuli. The collected signals were filtered with a half-amplitude bandpass of 0.016-70Hz and synchronously averaged over 40 the repetitions and over the 21 subjects. The obtained ERPs were leveraged to build five different datasets. These five datasets are different representations of the same ERPs in a spatial, temporal or morphological dimension. All the datasets were proposed to the two classification models, to identify the optimal one in terms of accuracy of the predictions and interpretability of the decision-making process.

Also, the labels related to the 14 classes were arranged according to a hierarchical structure. The design for the classification models produced is consistent with this hierarchical structure. Binary classification is performed in each node, progressively identifying the stimulus. Therefore, each stimulus is recognized, starting from the incoming ERP, through a series of binary classifications.

The performances have been evaluated also in terms of the interpretability of the decision-making process. This indicator is crucial especially in a medical context. For the boosted

trees-based model, features importance was listed based on F-Score. For feed-forward neural networks-based model, instead, a method called permutation importance was implemented, which estimates the importance of a feature in the classification process considering the accuracy loss determined by the removal of its informative contribution.

The ERP dataset representation that optimizes the trade-off between prediction accuracy and decision-making process interpretability turns out to be the spatial one, in which the time-series have been cut and debiased. The average accuracy achieved at each node by the boosted trees-based model for this representation is 99.02%. As far as the interpretability is concerned, the ranking produced by features importance reports which cortical zones are the most considered in recognizing the stimulus that caused the evoked potentials proposed. This information is relevant for a psycho-physiological analysis.

To adapt our classification models also to general-purpose studies, and not only laboratory experiment, we analyzed the prediction error trend versus the number of considered electrodes. We also evaluate our algorithms performance when considering only the electrodes available in the most used devices on the market.

Future developments include further optimization to reduce the number of electrodes and to extend the pool of recognized stimuli, without affecting the overall performances. Also, it would be interesting to monitor the behavior of the model in recognizing the ERPs of each subject.

However, the results we presented in this paper open the way to interesting developments for BCI systems. In fact, at best of our knowledge, this is the first study in the literature whose purpose is not to identify the presence of a priori known evoked potential but to recognize the stimulus that caused it. This represents a step forward towards the goal of being able to translate the user's neural commands into machine instructions, directly connecting the brain to any electronic device. This might as well enable the possible reconstruction of mental representations of specific sensory or semantic classes by classifying ongoing electrical signals of the EEG/ERPs in locked-in or comatose patients.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] "Reports and data," https://www.reportsanddata.com/, note = Accessed: 10-16-2019.

[3] J. Farquhar and N. J. Hill, "Interactions between pre-processing and classification methods for event-related-potential classification," *Neuroinformatics*, vol. 11, no. 2, pp. 175–192, 2013.

[4] J. Fan, J. W. Wade, A. P. Key, Z. E. Warren, and N. Sarkar, "Eeg-based affect and workload recognition in a virtual driving environment for asd intervention," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 1, pp. 43–51, 2017.

[5] N. K. Squires, K. C. Squires, and S. A. Hillyard, "Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man," *Electroencephalography and clinical neurophysiology*, vol. 38, no. 4, pp. 387–401, 1975.

[6] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A comparison of classification techniques for the p300 speller," *Journal of neural engineering*, vol. 3, no. 4, p. 299, 2006.

[7] G. Pfurtscheller, D. Flotzinger, and J. Kalcher, "Brain-computer interface—a new communication device for handicapped persons," *Journal of Microcomputer Applications*, vol. 16, no. 3, pp. 293–299, 1993.

[8] 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.

[9] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[10] A. Kübler, "Brain-computer interfaces for communication in paralysed patients and implications for disorders of consciousness," *The Neurology of Consciousness: Cognitive Neuroscience and Neuropathology*, pp. 217–233, 2009.

[11] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988.

[12] B. Blankertz, K.-R. Muller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlogl, G. Pfurtscheller, J. R. Millan, M. Schroder, and N. Birbaumer, "The bci competition iii: Validating alternative approaches to actual bci problems," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 14, no. 2, pp. 153–159, 2006.

[13] A. Rakotomamonjy, V. Guigue, G. Mallet, and V. Alvarado, "Ensemble of svms for improving brain computer interface p300 speller performances," in *International conference on artificial neural networks*. Springer, 2005, pp. 45–50.

[14] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter, "Bci competition 2003-data set iib: support vector machines for the p300 speller paradigm," *IEEE Transactions on biomedical Engineering*, vol. 51, no. 6, pp. 1073–1076, 2004.

[15] H. Cecotti and A. Graser, "Convolutional neural networks for p300 detection with application to brain-computer interfaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 433–445, 2010.

[16] Y. Zhang, G. Zhou, Q. Zhao, J. Jin, X. Wang, and A. Cichocki, "Spatial-temporal discriminant analysis for erp-based brain-computer interface," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 21, no. 2, pp. 233–243, 2013.

[17] M. Lopez-Gordo, F. Pelayo, A. Prieto, and E. Fernández, "An auditory brain-computer interface with accuracy prediction," *International journal of neural systems*, vol. 22, no. 03, p. 1250009, 2012.

[18] J. E. Meyers, R. J. Roberts, J. D. Bayless, K. Volkert, and P. E. Evitts, "Dichotic listening: Expanded norms and clinical application," *Archives of Clinical Neuropsychology*, vol. 17, no. 1, pp. 79–90, 2002.

[19] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[20] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in cognitive sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[21] L. Gupta, D. L. Molfese, and R. Tammana, "An artificial neural-network approach to erp classification," *Brain and cognition*, vol. 27, no. 3, pp. 311–330, 1995.

[22] 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.

[23] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[25] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[26] N. Elon Musk, "An integrated brain-machine interface platform with thousands of channels," 2019.

[27] "Neuropace," https://www.neuropace.com/, note = Accessed: 10-16-2019.

[28] "Advanced brain monitoring inc," https://www.advancedbrainmonitoring.com/, note = Accessed: 10-16-2019.

[29] "Emotiv systems," https://www.emotiv.com/, note = Accessed: 10-16-2019.