




Leveraging Deep Learning Techniques to Improve P300-Based Brain Computer Interfaces

Ihsan Dağ, Linda Greta Dui , Simona Ferrante , *Member, IEEE*,
Alessandra Pedrocchi , *Senior Member, IEEE*, and Alberto Antonietti , *Member, IEEE*

Abstract—Brain-Computer Interface (BCI) has become an established technology to interconnect a human brain and an external device. One of the most popular protocols for BCI is based on the extraction of the so-called P300 wave from electroencephalography (EEG) recordings. P300 wave is an event-related potential with a latency of 300 ms after the onset of a rare stimulus. In this paper, we used deep learning architectures, namely convolutional neural networks (CNNs), to improve P300-based BCIs. We propose a novel BCI classifier, called P3CNET, that improved P300 classification accuracy performances of the best state-of-the-art classifier. In addition, we explored pre-processing and training choices that improved the usability of BCI systems. For the pre-processing of EEG data, we explored the optimal signal interval that would improve classification accuracies. Then, we explored the minimum number of calibration sessions to balance higher accuracy and shorter calibration time. To improve the explainability of deep learning architectures, we analyzed the saliency maps of the input EEG signal leading to a correct P300 classification, and we observed that the elimination of less informative electrode channels from the data did not result in better accuracy. All the methodologies and explorations were performed and validated on two different CNN classifiers, demonstrating the generalizability of the obtained results. Finally, we showed the advantages given by transfer learning when using the proposed novel architecture on other P300 datasets. The presented architectures and practical suggestions can be used by BCI practitioners to improve its effectiveness.

Index Terms—Biomedical engineering, brain-computer interfaces, deep learning, neural implants, neurotechnology.

I. INTRODUCTION

BRAIN-COMPUTER Interfaces (BCIs) aim to construct solid communication bridges between the brain of a human being and an external device. They can assist people with motor function deficits, who can gain the ability to act on their environment without utilizing peripheral nerves and muscles, but their

brain signals [1]. BCI helps in the training of lacking skills, as well, thanks to the feedback it can give on the correct execution of a mental task [2]. To evaluate the accuracy of a mental task, (1) brain signals must be recorded, mainly with non-invasive electroencephalography (EEG) electrodes; (2) task-related features must be detected in brain signals, thanks to artificial intelligence (AI) techniques [3].

P300, a visually evoked event-related potential, is one of these distinctive features, and many BCI protocols are based on the recognition of this wave in EEG signals [4]–[7]. P300 arises in the brain signals as a result of the perception of a rare stimulus in a visually concentrated state. A large positive deflection appears 300 ms after the detected stimulus [8], but some external factors influence the properties and appearance of this wave, making its detection non-trivial. E.g., in older people, the latency can be higher, and the deflection can be lower compared to younger people. The deflection of the wave is also determined by the characteristics of the stimulus, and it increases as the rarity of the stimulus increases. Neurophysiological studies showed that the major operating rhythms of the P300 are mainly the delta (1-3 Hz) and theta (4-7 Hz) frequencies, therefore this waveform is highly utilized as an indicator of attention on visual tasks [9]–[13].

Recently, BCI systems have been proposed to help people diagnosed with Autism Spectrum Disorder (ASD) to train and improve their joint attention, which is the ability to put attention on a particular object that is indicated by another individual [14]. This training method uses P300-based BCI since it utilizes P300 waveform as the distinctive feature, requiring a good classifier to be able to provide the patients with feedback on joint attention tasks. On this issue, Amaral and colleagues designed an experiment [15], [16], involving 15 subjects affected by ASD, which led to the construction of the BCIAUT-P300 dataset [17]. Virtual reality simulated a rare visual stimulus. Data were arranged into a training session and an unlabelled online session, and a scientific challenge (IFMBE Scientific Challenge Competition [18]) was called to find the best AI solution for P300 detection.

Bittencourt *et al.*, using linear support vector machines (linear SVM), reached 82% average classification accuracy with the help of session-specific optimized models [19]. Zhao *et al.* leveraged linear discriminant analysis (LDA), SVM, convolutional neural networks (CNN) and long-short term memory (LSTM) models; LDA achieved the highest average accuracy in online sessions with 67% [20]. Arancibia *et al.* took advantage of LDA, linear SVM and radial SVM models and

Manuscript received 3 August 2021; revised 13 April 2022; accepted 3 May 2022. Date of publication 12 May 2022; date of current version 5 October 2022. (Alessandra Pedrocchi and Alberto Antonietti are co-last authors.) (Corresponding author: Alberto Antonietti.)

The authors are with the Neuroengineering and Medical Robotics laboratory, Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20131 Milano, Italy (e-mail: ihsan.dag@mail.polimi.it; lindagreta.dui@polimi.it; simona.ferrante@polimi.it; alessandra.pedrocchi@polimi.it; alberto.antonietti@polimi.it).

Digital Object Identifier 10.1109/JBHI.2022.3174771

the highest average accuracy they could reach was 80% with LDA [21]. Krzeminski *et al.* trained LDA, Logistic Regression and Minimum Distance to Mean (MDM) algorithms and the best classification performance was 81.2% with LDA [22]. Chatterjee *et al.* employed Bayes LDA (BLDA), random under-sampling boosting (RUS-Boosting) and CNN models and the best performance was obtained by BLDA with 73% average accuracy [23]. Adama *et al.* leveraged decision trees, random forest, SVM and multi-layer perceptron, and the best performance was 70% with multi-layer perceptron [24]. Borra *et al.* employed Lightweight CNN to obtain 84.43% for within-session training and 92.27% for cross-session training [25] and became the winner of the competition [18].

Given these promising results, AI seems almost mature to be leveraged in a BCI context, even if higher accuracies should always be a target. However, BCI systems are mainly addressed to people with physical or cognitive disabilities, so accuracy alone does not guarantee successful BCI adoption [26].

The first problem that BCI users face is represented by the long calibration phase [15], [16]. On the one hand, subjects need to learn the task; on the other hand, the AI algorithm must learn to recognize the P300 waveform, which presents a high between-subjects variability. As a consequence, subjects might get tired or frustrated, resulting in a difficult adoption. Then, different techniques to reduce the burden on subjects must be explored.

A second problem is the signal acquisition setup. Non-invasive EEG electrodes are the preferred setup, but their arrangement on the scalp must follow precise rules (e.g., 10-10 or 10-20 systems) that might prevent an easy adoption and comfort. The P300 waveform is mostly obtained from the parietal lobe of the brain, so the usage of electrode positions can also be lowered and narrowed to a specific location on the head [27]. In Amara and colleagues' experiment [15] only 8 electrodes were used, but it can be hypothesized that even fewer electrodes are enough to preserve signal detection accuracy [28].

Finally, the generalizability of the proposed techniques should be assessed. In fact, it must be assured that the detecting capabilities of the AI algorithm do not vary if the acquisition system changes to avoid the need for multiple calibrations.

Furthermore, the optimization of the algorithms is not enough when AI models are supposed to interact with people. The Trustworthy AI guidelines [29] require that AI results that involve human beings must be understandable by human beings themselves. Hence, explainable AI (XAI) [30] must be applied to the proposed algorithms to provide a justification for their predictions.

All these needs can be translated into the aims of our study: (1) improving P300 detection accuracy by proposing a new AI model; (2) proposing different solutions to reduce the time required to achieve a reliable system by assessing (a) the optimal signal length around the visual event, and (b) the minimum number of calibration session needed; (3) exploring the feasibility of reducing the number of electrodes needed by discussing electrodes importance for the prediction; (4) assessing the generalizability of the proposed solutions by comparing their effect on the same dataset with different classifiers, and on different datasets. All these aims should be achieved with a particular

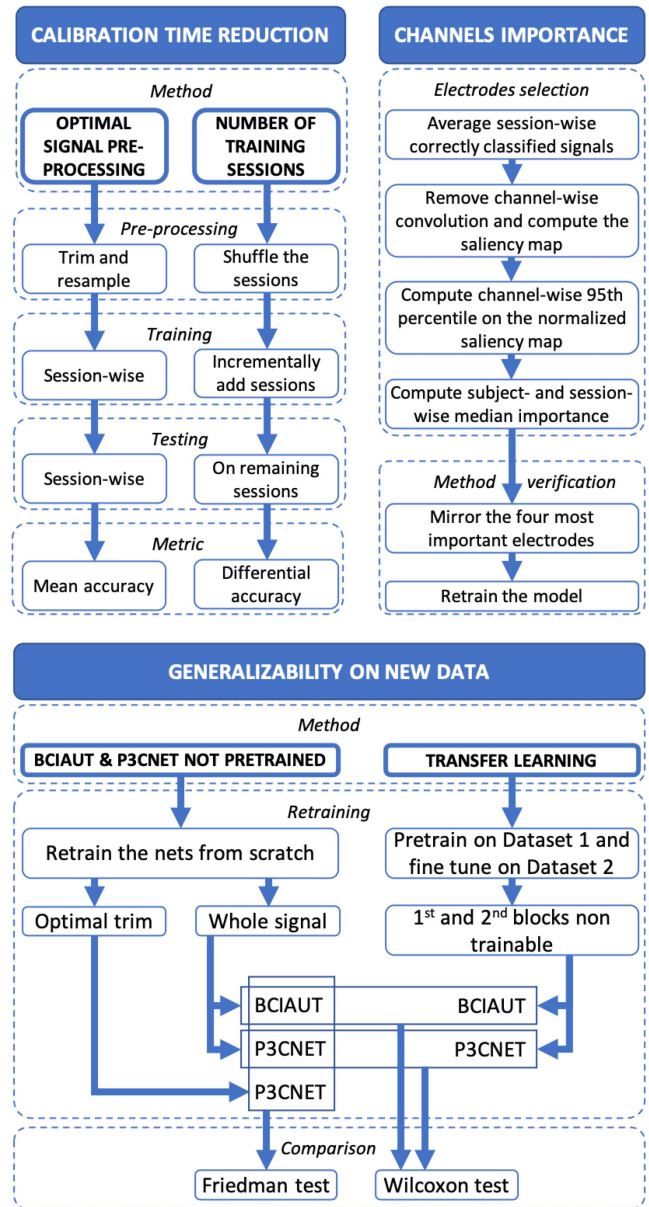


Fig. 1. Methodological workflow.

focus on models explainability. The methodological workflow followed is summarized in Fig. 1.

II. METHODS

The first aim of this work is to improve the accuracy performance of the state-of-the-art models, bench-marked using the BCIAUT-P300 dataset [18].

The BCIAUT-P300 dataset has been recorded on 15 subjects performing a joint attention task, where the participant had to pay attention to one specific object over 8 possible ones. Therefore, for every turn of 8 objects, only 1 of them was supposed to generate the P300 waveform. This feature of the experiment made the collected dataset an unbalanced one for P300 and non-P300 classes.

TABLE I
TECHNICAL DETAILS OF THE DATASETS USED

	BCIAUT-P300 [16]	GIB-UVa ERP-BCI [31]
Subjects	15 high-functioning ASD subjects	42 healthy, 31 severely disabled subjects
Age	16-38 years old subjects	19-67 years old subjects
Electrodes placement	C3, Cz, C4, CPz, P3, Pz, P4, Poz; ground at Afz	Fz, Cz, Pz, P3, P4, PO7, PO8, Oz; ground at FP
Stimulus	Flash light on 8 objects in virtual reality (joint attention)	Flash light on row-column paradigm (spelling task)
Target	Pointed by an avatar, different for each run	Command decided by the subject
Stimulus duration	100 ms	variable: 100-250 ms
Inter-stimuli interval	200 ms	variable: 62.5-75 ms
Window	[-200;1200] ms	[0;1000] ms
Sampling rate	250 Hz	256 Hz
Pre-processing	50 Hz notch-filter, 2-30 Hz 8 th order Butterworth band-pass filter	FIR bandpass filter 1000 th order [0.5;45] Hz; baseline normalization [-200;0] ms
# of trials	466 000 for all subjects	701 615 for all subjects

TABLE II
P3CNET LAYERS AND DETAILS

Layer	Details
2D Convolution	Filters = 8, Kernel Size = (8,15), Strides = (1,1), Activation = Linear, Maximum Norm = 1.0, No Bias
2D Batch Normalization	Default Parameters
Activation	Linear Exponential Unit
2D Average Pooling	Kernel Size = (1,4)
Dropout	Threshold = 0.35
2D Convolution	Filters = 16, Kernel Size = (1,5), Strides = (1,1), Activation = Linear, No Bias
2D Batch Normalization	Default Parameters
Activation	Linear Exponential Unit
2D Average Pooling	Kernel Size = (1,4)
Dropout	Threshold = 0.20
2D Convolution	Filters = 32, Kernel Size = (1,5), Strides = (1,1), Activation = Linear, No Bias
2D Batch Normalization	Default Parameters
Activation	Linear Exponential Unit
2D Average Pooling	Kernel Size = (1,2)
Dropout	Threshold = 0.40
Flatten	-
Dense	2 Classes, Activation = Softmax, Maximum Norm = 0.25

For every subject, the first 3 sessions were weekly, whereas the remaining 4 sessions (online sessions) were monthly. Each session foresaw a calibration phase and a utilization phase. The models in the IFMBE Scientific Challenge Competition were bench-marked on the (unlabeled) utilization phase of the four online sessions.

EEG signals were recorded from -200 ms to 1200 ms with respect to the stimulus onset, and the sampling rate was 250 Hz, so for each EEG signal, 350 time-samples were available in the dataset. The EEG was recorded using 8 different electrode positions (C3, Cz, C4, CPz, P3, Pz, P4, Poz) in accordance with the 10-10 system [15]. More details are reported in Table I, left column.

A. Performance Improvements

Since the BCIAUT CNN Model proposed by Borra and colleagues [25] won the IFMBE Scientific Challenge Competition [18], obtaining the highest accuracy in the online sessions, we selected it as our reference.

1) **BCIAUT CNN Model:** BCIAUT CNN model is based on a CNN classifier adapted from the EEGNet model trained to discriminate between P300 and non-P300 classes [32]. BCIAUT CNN was designed to keep the number of trainable parameters limited by means of depth-wise and point-wise convolutions, resulting in 1386 trainable parameters.

During the pre-processing, EEG signals for each trial were trimmed between -100 ms and 1000 ms with respect to the stimulus onset. Each of these signals was then resampled to 140 samples and finally standardized to have zero-mean and unit-variance. The model was implemented with the layer details described in Borra's paper [25]. Optimization and training hyper-parameters were also kept the same. The only differences in implementation have been the initialization seed, the weights assigned to different classes, and the epoch patience (the number of epochs with no improvement) used in early stopping since

these parameters were not specified in the paper. After a preliminary investigation, to closely match the results in Borra's paper, we used "1234" as initialization seed, and the epoch patience for early-stopping was defined as 55. Different weights were assigned to account for the unbalanced classes: the P300 class weighted 7 times more than the non-P300 class. In this way, having a pseudo-balanced dataset, we can use accuracy as the main target metric.

The implemented BCIAUT CNN model has been used to compare and validate the results we have obtained with the proposed BCI classifier, the P3CNET model.

2) **P3CNET Model:** We wanted to design an AI model to get possibly higher accuracy results and to provide a "second opinion" in testing the robustness of our methodologies. For this reason, we have taken inspiration from the BCIAUT CNN model, along with other deep learning networks [33]–[35]. For this purpose, we designed P3CNET with the architecture in Fig. 2 and implemented the model in line with the details reported in Table II. As the BCIAUT CNN model, 3 convolution layers were used. Further contamination came from VGG-16 [36], which is used for transfer learning on image classification, which suggested searching for higher-level features using successive convolution layers. We used 15 as kernel size for the first 2D convolution as done by Shan and colleagues [37]. The model we designed had 4338 parameters, more than the BCIAUT CNN model, but still with a low computational cost for its training. For optimization, we preferred the Adam optimizer, as in BCIAUT CNN model, and the learning rate was equalized to 0.0005. After a preliminary investigation, we set validation loss as the metric to be minimized and early-stopping epoch patience to 45. During training, the maximum epoch number was set to 1000 and the validation split to 0.2. Different weights were assigned to account for the unbalanced classes: the P300 class weighted 7 times more than the non-P300 class. For both training and testing, the batch size was set to 128.

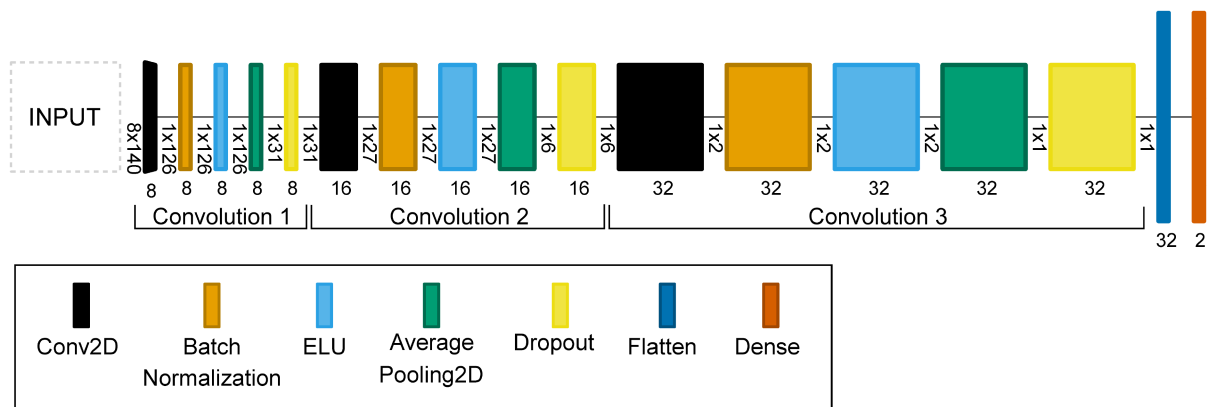


Fig. 2. P3CNET Architecture, composed of three convolution blocks, each one comprising a 2D Convolution (Conv2D), batch normalization, ELU, Average Pooling 2D, and dropout layers. Numbers close to each layer represent the shape of the output tensors.

We bench-marked the proposed P3CNET model on the four online sessions of the BCIAUT-P300 dataset and compared the results with the accuracies obtained with the BCIAUT CNN model. To verify the robustness of the model to the choice of the initialization seed, we repeated the experiment using another five initialization seeds (i.e., “1000”, “2000”, “3000”, “4000”, and “5000”).

B. Calibration Time Reduction

One of the biggest problems to face during BCI treatment stands out as the data and time required for the calibration of BCI. This can be critical for patients showing hyperactivity when the experiment time lengthens [38]. Therefore, we proposed two ways to reduce the time needed for the calibration of a P300-based classifier: (a) to facilitate the classification with an optimal selection of the EEG signal provided as input to the classifier and (b) to identify the minimum number of calibration sessions needed to obtain satisfactory performances.

1) *Optimal Signal Pre-Processing*: Two elements are crucial to obtaining high accuracy values in deep learning: an optimized and well-designed model and high-quality data pre-processing [39], [40]. For this purpose, we explored which portion of the EEG signals should be used to facilitate the detection of the P300 wave. Therefore, we developed a grid-search approach to trim the EEG signals within a specific range, selecting different initial samples (start time) and sample windows (window length). To ensure the robustness of the results, we performed the grid-search with both BCIAUT and P3CNET CNN models separately.

For each tested combination of start time and window length, the pre-processing of each session’s training and testing sets was adapted to the input tensor size of the models with appropriate resampling, and the training and testing were performed for each session separately. We computed the mean accuracy and standard deviation of all sessions and all subjects for each combination. In this way, robust results were obtained by calculating the model’s data for each combination over 105 different data points (15 subjects x 7 sessions).

Having tested all combinations for both BCIAUT and P3CNET, we primarily aimed at maximizing the mean accuracy while maintaining low standard deviations. In this way, we

TABLE III
14 SESSION ORDERS USED TO FIND THE MINIMUM NUMBER OF TRAINING SESSIONS

1,5,6,4,2,3,7	7,1,4,2,5,6,3	3,7,2,5,1,4,6	4,5,1,7,3,2,6
6,2,7,1,3,4,5	5,2,6,7,4,1,3	1,4,7,3,6,5,2	4,3,2,5,7,6,1
5,1,3,6,2,7,4	2,4,1,6,5,3,7	6,3,5,2,1,7,4	2,6,4,3,7,1,5
3,7,5,4,6,2,1	7,6,3,1,4,5,2		

identified the optimal pre-processing choices guaranteeing the best performances for both models.

2) *Number of Training Sessions*: As mentioned before, while obtaining the BCIAUT-P300 dataset, seven training sessions were performed by each subject, leading to a long experiment duration. Here, we wanted to see whether all of these training sessions were necessary and if we could eliminate some of them and shorten the overall time needed for the BCI calibration.

For each experiment, 14 different session orders were used since the existing order of the sessions may create a session-order bias. As shown in Table III, the 14 different orders were defined in such a way that each session took place 2 times at a certain position.

Each condition has been tested with 6 different initialization seeds to ensure the robustness of the findings; therefore, 84 experiments (6 seeds x 14 session orders) have been carried out for each subject.

After the session orders were defined, we included an increasing number of training sessions in the calibration data. The calibration process first started with the use of only one training set; in this case, the test sets of the remaining 6 sessions were used to measure the accuracy. Then an additional session, according to the order identified previously, was added to the calibration data, and the model was trained from scratch. In this case, the accuracy was computed on the testing sets comprising the 5 remaining sessions. This process continued until we used 6 sessions for the training and 1 session for the testing. Finally, we also tried to use all 7 sessions for the calibration. In this case, we reported the training accuracy.

To identify the minimum number of sessions needed for the training, we then looked at the difference in the median accuracies obtained by each subject in the 84 experiments when using

x or $x + 1$ training sessions. When the gain in accuracy (i.e., Delta accuracy) approaches zero, that means that the inclusion of an additional $x + 1$ session to the training did not lead to an increase in the testing accuracy.

C. Channels Importance and Electrode Selection

Deep learning models have been successfully leveraged in tasks such as image classification or pattern recognition, and they often achieved very high accuracies. Although high accuracy is the most prominent feature of a classifier, its explainability is of great importance too. Especially when used for biomedical applications, the users and the operators must have confidence in the model. This is possible if the model itself is more comprehensible. Toward this aim, one possible method is to build saliency maps, illustrating which are the features of the input (e.g., of the input image) used by the model to generate the classification. One of the main algorithms to generate saliency maps is the Gradient-weighted Class Activation Mapping (Grad-CAM) [41]. This method prepares a localization map from the last convolutional layer to the input layer. Grad-CAM uses the gradients of the target class to create a saliency map of the same size as the input image. When the input image is compared to this map pixel-by-pixel, it shows which areas or features of the input image are more significant for the model to select a certain class. The Grad-CAM algorithm can be used with a wide variety of CNN models, and we utilized it for the explainability of the two CNN models.

For both BCIAUT and P3CNET, we trained the model from scratch with the training set for each session. Then, the calibrated model was tested with the testing set of the same session, and the correctly classified input EEG signals of the P300 class were averaged. Thus, an average P300 input image was created for each session. Then these input tensors were provided to the Grad-CAM function, and a saliency map was obtained for each session.

We were particularly interested in knowing which of the 8 EEG electrodes placed at different anatomical locations on the scalp (i.e., C3, Cz, C4, CPz, P3, Pz, P4, POz) carried more information to recognize the P300 wave. Therefore, we needed to analyze each row of the input EEG signal separately.

However, since the Grad-CAM algorithm starts from the last convolutional layer while creating localization maps, it normally creates saliency maps equal to the output tensor size of that layer. Then, this map is brought to the same size as the input image tensor by applying interpolation with the same obtained values. At this point, both BCIAUT and P3CNET have a vertical dimension of 1 in the output of the last convolutional layer due to the between-channels (vertical) convolution applied in one of the previous convolutional layers. This causes Grad-CAM to interpolate from 1 to 8 and prevents analysis of the importance of different EEG channels. As a solution to this, we removed the filters that perform spatial convolution from both models. Thus, 8 separate channels were obtained in the output of the last convolutional layer for both models. Therefore, Grad-CAM provided channel-specific saliency values. Leveraging this modification, we applied Grad-CAM to all sessions and all subjects as previously described.

Since we aimed at identifying the most informative channels/electrodes, we computed the relative importance of each electrode for each session and subject. The saliency maps generated by the Grad-CAM algorithm were normalized between 0 and 1. Then, we computed for each channel (i.e., for each row of the saliency maps) the 95th percentile. Then, for each channel, we computed its importance value by looking at the median importance, considering all subjects and all sessions ($N = 105$). The channels with the highest importance values corresponded to the electrodes that brought most of the information to identify the P300 wave. Then, for each CNN model, we evaluated the importance of the electrode anatomical locations by projecting the values on the EEG topographic maps. Finally, with the hypothesis that the less informative electrodes might bring noise and worsen the performance, we re-trained the models with the four most informative electrodes only. To preserve the input size of 8 channels, we duplicated the most informative electrodes by mirroring them vertically.

D. Generalizability and Transfer Learning

As a final analysis, we aimed at verifying whether the results obtained with the two CNN models were valid when used to classify a different dataset. For this purpose, we used another EEG dataset, called GIB-UVa ERP-BCI, including P300-protocol recordings from 73 subjects (42 healthy and 31 with motor disabilities) [31], [42], details are reported in Table I, right column. EEG data were recorded with 8 channels (8 active electrodes placed at Fz, Cz, Pz, P3, P4, PO7, PO8, and Oz, according to the 10–10 system) from 0 to 1000 ms with respect to the visual stimulus onset. Using this dataset, we selected 20 random healthy subjects and 20 severely disabled patients. The patients suffered from different pathologies; we randomly selected 20 of them while keeping the relative incidence as the one in the full dataset. As a result, 4 patients had spinal cord injury, 2 had Friedrich's ataxia, 4 had cerebral palsy, 2 had polymalformative syndrome, 1 had a stroke, and 7 had multiple sclerosis. After completing the selection of the dataset and subjects, we tested different scenarios. First, we trained from scratch both CNN models to test their performances. For the BCIAUT CNN model, the original EEG data (from 0-1000 ms) was provided, while we tested the P3CNET model with both original EEG data and after the optimized pre-processing trimming identified on the BCIAUT-P300 dataset. In this way, we investigated if the optimized pre-processing is beneficial for the new dataset. Secondly, we wanted to verify if it was possible to apply transfer learning techniques [43], pre-training the CNN models with the BCIAUT-P300 dataset, and then fine-tuning the networks on the new dataset. For the transfer learning, after a preliminary exploration, we made the first two convolution blocks non-trainable. To ensure robustness, we have done 8 tests for each subject, using different initialization seeds. Therefore we performed 1920 experiments in total (40 subjects x 2 - with/without transfer learning - x 3 model conditions - BCIAUT CNN with original EEG data, P3CNET with original EEG data, P3CNET with pre-processed EEG data - x 8 initialization seeds).

In all experiments, we used 70% of the data for the training and 30% of the data for the testing.

We then compared the testing accuracies obtained by the BCIAUT CNN, P3CNET using the original EEG signal, and P3CNET using the optimized pre-processing. We employed the Friedman test for paired samples to compare the three groups, and we then performed a post hoc comparison between groups using Wilcoxon tests with continuity correction for multiple comparisons.

Then, we compared the performances obtained without the pre-training and with the pre-training (transfer learning) of the CNN models on all 7 sessions of the BCIAUT-P300 dataset. We wanted to test whether transfer learning contributed to increased accuracy. In this case, we applied the Wilcoxon test for paired samples to compare the two groups.

Finally, we stratified the results by dividing the healthy subjects from those affected by motor disability to verify if the pre-training was more beneficial for the classification of EEG data produced by healthy or pathological subjects. For the two classes, we employed the Wilcoxon tests for paired samples to compare the results with the training from scratch and with the pre-training on the BCIAUT-P300 dataset.

For all statistical tests, we set the significance level to 0.05.

E. Hardware and Software

The training and testing of the CNN models have been done exploiting the online platform Google Colab Pro, a cloud computing solution that allowed us to use a powerful GPU needed for the training of the models. We exploited the Tesla P100 hardware accelerator to decrease the time needed for the training. Even if we employed CNN with a reasonable number of free parameters, we carried out thousands of experiments, as described in the previous sections, therefore we needed a powerful platform to run them. On average, it took ~ 15 seconds to train BCIAUT and ~ 45 seconds to train P3CNET, when using the seven-sessions training set for a single subject.

We used Python language for the creation, training and testing of the CNN models and also for the subsequent analysis of the results. We used the following versions and libraries: Python 3.7.10, Tensorflow 2.4.3, Keras 2.4.3, NumPy 1.19.5, Pandas 1.1.5, Matplotlib 3.2.2, and SciPy 1.4.1.

The Tensorflow codes to create and train the two CNN models, the resulting data and the codes used to analyze the results, including a Jupyter Notebook that reproduces all the figures and results reported in this paper, are publicly available at IEEE Dataport [44].

III. RESULTS

A. Performance Improvements

The replicated BCIAUT CNN was tested on online sessions and resulted in a median accuracy of 88.00%, with a 95% confidence interval of the median [80.00; 90.00]%. The P3CNET provided a median accuracy of 92.00% [88.00; 94.00]%. These results are reported in Fig. 3.

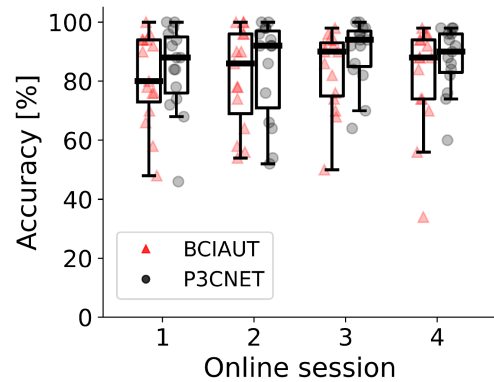


Fig. 3. Accuracy comparison on online sessions between BCIAUT and P3CNET. Red triangles (BCIAUT) and black circles (P3CNET) represent the accuracy obtained for each subject ($N = 15$) in the four online sessions. Box and whisker plots report median (thick line), quartiles (box) and the range of the data, excluding outliers (whiskers).

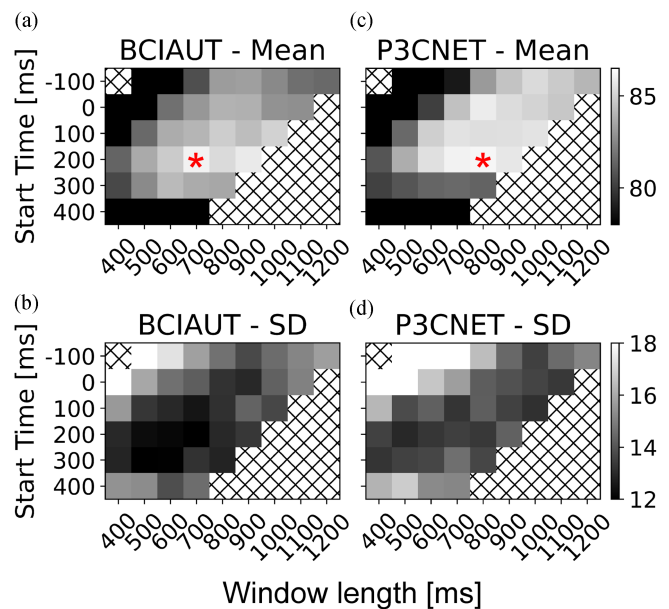


Fig. 4. Grid-search analysis to identify the optimal EEG signal selection. Panels A and C report the mean classification accuracy on online sessions obtained with BCIAUT and P3CNET, respectively. The red asterisks indicate the combinations of start time and window length that led to the highest mean accuracy. Panels B and D report the standard deviation of the classification accuracy obtained with BCIAUT and P3CNET, respectively.

With five different initialization seeds, the median accuracies obtained with P3CNET were consistently higher than BCIAUT: 86%, 87%, 88%, 86%, and 86% for BCIAUT and 91%, 90%, 90%, 92%, and 88% for P3CNET, therefore proving that P3CNET yielded to robustly better performances.

B. Calibration Time Reduction

Concerning time reduction, Fig. 4 shows the results of the search for an optimal signal trimming. Panels A and B refer to accuracy, which should be maximized, i.e., lighter colors; panels C and D refer to standard deviation, which should be

TABLE IV

STATISTICS OF TESTING ACCURACY DISTRIBUTION OBTAINED USING ONLY THREE SESSIONS FOR THE TRAINING

Accuracy Interval (%)	BCIAUT		P3CNET	
	Number of Sessions	Session Perc. (%)	Number of Sessions	Session Perc. (%)
0-10	0	0.00	0	0.00
10-20	1	1.67	1	1.67
20-30	0	0.00	0	0.00
30-40	0	0.00	0	0.00
40-50	4	6.67	2	3.33
50-60	5	8.33	3	5.00
60-70	10	16.67	8	13.33
70-80	4	6.67	8	13.33
80-90	14	23.33	17	28.33
90-100	22	36.67	21	35.00

minimized, i.e., darker colors. Panels A and C are BCIAUT results, panels B and D are P3CNET results. The best time windows, expressed as (start time, window length), for BCIAUT are (200 ms, 700 ms) and (300 ms, 500 ms), where average accuracy was $86.07\% \pm 11.71\%$; for P3CNET are (200 ms, 800 ms) and (200 ms, 500 ms), with $86.20\% \pm 12.96\%$ of accuracy. When considering both models, the (200 ms, 700 ms) combination was selected, as it resulted in an average accuracy of 86.07% for both BCIAUT (SD: 11.80%) and P3CNET (SD: 13.49%). Note that these mean accuracies have been obtained when using a single session for training and testing. These are therefore lower than the accuracies obtained in the previous section, where the training set use was much bigger. Instead, the results on the trimmed signal should be compared to the corresponding mean accuracy achieved by leveraging the (-100 ms, 1000 ms) time window, as in Borra and colleagues [25], that are 81.7% and 84.7% for BCIAUT CNN and P3CNET, respectively. Therefore, the selected time window spanned between 200 ms and 900 ms, which was half of the one used in the Amaral and colleagues' experiment [16], and 400 ms shorter than the one used by Borra and colleagues [25].

The second approach to reducing the calibration time was to lessen the number of calibration sessions. In Fig. 5, Panel A reports the accuracy distribution when the number of calibration sessions is increased; Panel B shows the improvement (Delta accuracy) caused by the addition of a new calibration session. It can be noticed that both accuracy and its improvement tend to converge. A single session largely underestimates the potential performances of the models, the second session helps boost the performance, but adding more than three sessions seems to bring an irrelevant improvement. Therefore, four out of seven sessions could be discarded from the training process. Then, we applied three-session calibration on both BCIAUT and P3CNET. Table IV summarizes the results of these tests. Averaging over the 60 sessions, accuracies of 78.94% and 81.04% were obtained for BCIAUT and P3CNET, respectively.

C. Channels Importance and Electrode Selection

Concerning the feasibility of reducing the number of acquisition channels, the accuracy results given by the elimination of the channels convolutions were 66.83%, 67.87%,

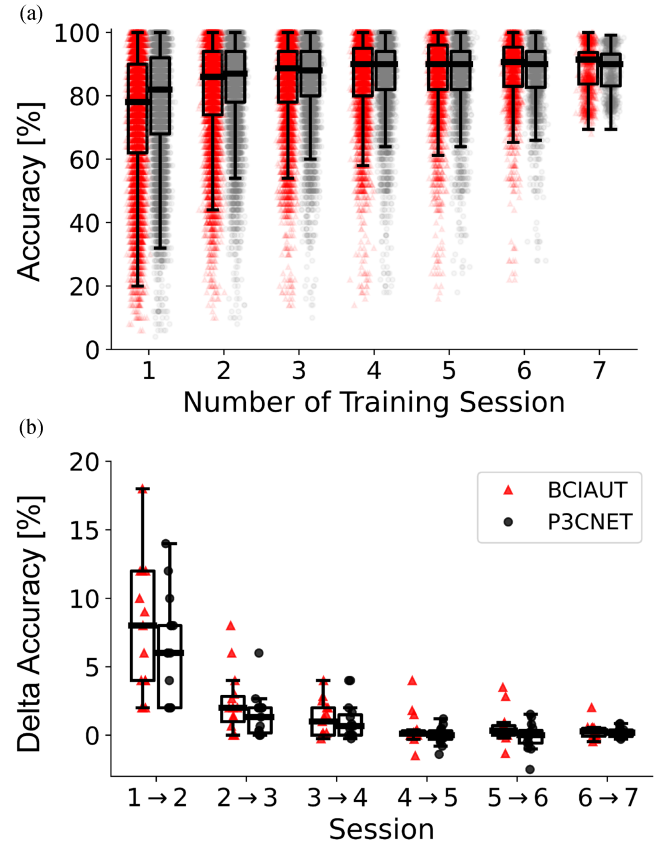


Fig. 5. Accuracy with different sizes of the training set. Panel A shows the testing accuracies obtained with a different number of training sessions (from 1 to 7) for all subjects. Each point (red triangles for BCIAUT, black circles for P3CNET) represents the accuracy obtained by one subject in one experiment. Each condition has been tested with 6 different initialization seeds and 14 combinations of session order, i.e., 84 experiments for each subject. N.B. the accuracy displayed when using all seven sessions for the training is the training accuracy since no sessions are left for the testing. Box and whiskers plots report median (thick line), quartiles (box) and the range of the data, excluding outliers (whiskers). Panel B shows the difference in accuracy between the median accuracies obtained by each subject in the 84 experiments when using x or $x + 1$ training sessions. Each point ($N = 15$, red triangles for BCIAUT, black circles for P3CNET) represents the gain (or loss, when the delta accuracy is negative) in accuracy obtained when adding a session to the training set.

67.87%, 68.00%, 67.40% for BCIAUT on online sessions, whilst P3CNET achieved 72.56%, 72.87%, 72.40%, 71.53%, 72.80%. As the accuracy was still satisfactory, we proceeded with the analysis.

Fig. 6 C shows the averaged channel importance for BCIAUT (left) and P3CNET (right), represented as topographical maps. It must be noticed that P electrodes (channels 5, 6, 7, 8) had higher values. When considering P electrodes only in the models training, accuracy results were 78.23% and 80.83% for BCIAUT and P3CNET, respectively.

D. Generalizability and Transfer Learning

The results of the application of the CNNs described in this paper to the new dataset are reported in Fig. 7A. BCIAUT

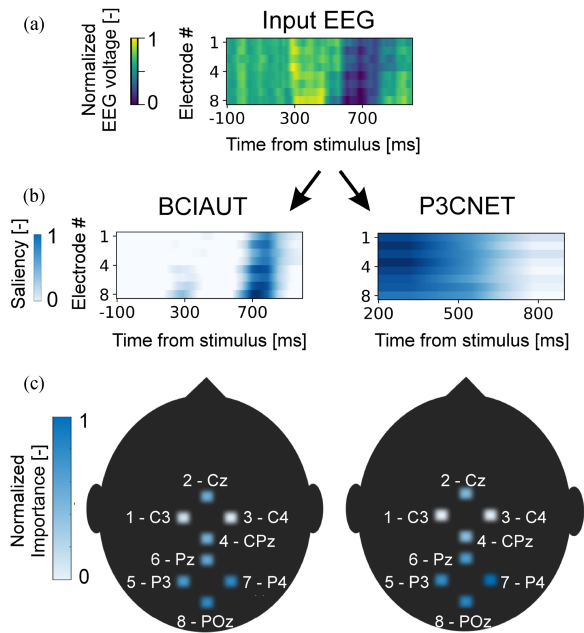


Fig. 6. Saliency maps of the input EEG signals and electrodes importance. Panel A shows an example of the normalized average EEG signal recorded during P300 trials from one subject during one session (namely, subject 12, session 1). Each row represents one EEG channel, each column represents one time-sample. Yellow and blue pixels represent positive and negative deflections, respectively. It is possible to notice the P300 wave between 300 and 500 ms and a subsequent negative deflection around 700 ms. Panel B represents the saliency maps corresponding to the input EEG for BCIAUT (left panel) and P3CNET (right panel), considering their modified versions without the channels convolution as explained in the Methods. Dark blue pixels represent regions of the input image that are considered salient for the classification of the P300 wave. Panel C shows the EEG topographical maps of electrodes normalized importance for BCIAUT (left) and P3CNET (right). Channels from 1 to 8 were represented at the anatomical locations on the scalp of the respective electrodes: C3, Cz, C4, CPz, P3, Pz, P4, POz. Dark blue and white regions represent high and low importance, respectively.

resulted in a median accuracy of 69.08%, with a 95% confidence interval of the median of [67.92; 70.47]%. P3CNET applied to the untrimmed (original) signal resulted in a median accuracy of 72.44% [71.08; 73.61]%. P3CNET applied on the time window that was considered optimal for the BCIAUT-P300 dataset resulted in a median accuracy of 71.40% [69.81; 73.32]%. The Friedman test revealed significant differences between the groups (p -value = 0.01). The post hoc analysis confirmed that the second group accuracy (“P3CNET original”, i.e., untrimmed) was significantly higher both than the BCIAUT results (p -value = $6.44 \cdot 10^{-3}$) and the trimmed P3CNET results (p -value = 0.02). On the other hand, BCIAUT and the trimmed P3CNET results did not differ significantly (p -value = 0.44).

Fig. 7B shows the results of training from scratch, with respect to performing transfer learning with both CNNs. The median accuracy achieved when training from scratch was 67.75% [66.83; 68.99]%. The median accuracy achieved when applying transfer learning was 73.50% [72.59; 74.35]%. As confidence intervals are disjointed and the Wilcoxon test rejected the null hypothesis of equal medians (p -value = $1.07 \cdot 10^{-10}$), it is possible to affirm that performance is generally higher when applying

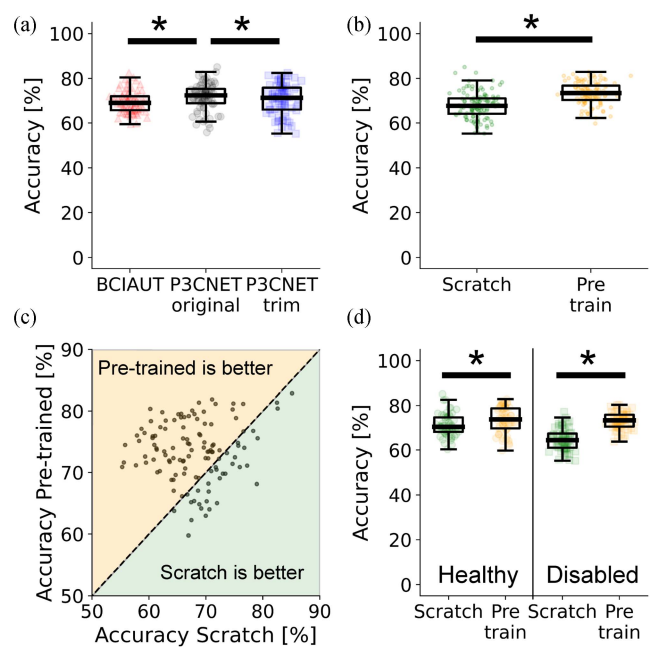


Fig. 7. Generalizability and transfer learning. Panel A shows the accuracy obtained on the GIB-UVa ERP-BCI dataset by the BCIAUT CNN model with red triangles, the P3CNET model with the original EEG data (from 0 to 1000 ms after the visual stimulus) with black circles, and the P3CNET with the pre-processed EEG data (from 200 to 900 ms after the visual stimulus) with blue squares. Box and whiskers plots report median (thick line), quartiles (box) and the range of the data, excluding outliers (whiskers). Asterisks indicate a significant difference between two groups ($N = 80$, Wilcoxon post hoc analysis with correction for multiple comparisons). Panel B shows the accuracies obtained when training both CNNs from scratch (green circles, $N = 120$) or when fine-tuning the pre-trained networks (orange circles, $N = 120$). Panel C shows the results obtained with and without the pre-training, i.e., the same data points of panel B ($N = 120$). For each experiment, accuracies without the pre-training are the abscissae, while the accuracies obtained with transfer learning are the ordinates. The data points placed in the orange area above the main diagonal (dashed black line) are experiments where the pre-training granted a higher accuracy. Data points placed below the main diagonal are experiments where the CNN trained from scratch performed better. Panel D shows the stratification of the experiments divided into healthy (left) and pathological (right) conditions ($N = 60$ for each group). For both healthy participants and severely disabled patients, transfer learning significantly improved classification accuracy.

transfer learning. The same result can be inferred from **Fig. 7C**, where data are scattered according to the accuracy achieved with (ordinate) and without (abscissa) pre-training. In this figure, it is possible to observe that transfer learning is particularly helpful when the classification accuracies are lower than 65% for the CNNs trained from scratch. For those experiments where it is difficult to identify the P300 wave, the pre-training on the BCIAUT-P300 dataset granted a higher success.

Fig. 7D shows the stratification of the latter results according to the health condition of the subjects. P300 detection accuracy on healthy subjects was slightly higher (Wilcoxon test for paired samples, p -value = 0.02) when tested with or without a pre-trained network, as training from scratch provided a median accuracy of 70.44% [69.63; 71.08] and transfer learning provided a median accuracy of 73.81% [72.34; 75.52]%. More importantly, P300 detection accuracy for disabled subjects

found a significant benefit in the application of transfer learning (Wilcoxon test for paired samples, p -value = $8.7 \cdot 10^{-10}$). In fact, median accuracy when training from scratch was 64.38% [63.54; 66.19]%, whilst it was 73.38% [72.09; 74.46]% when transfer learning was applied.

IV. DISCUSSION

In this work, we proposed deep learning methods to improve brain-computer interfaces through the detection of the P300 waveform in EEG recordings. Differently from standard approaches, which usually aim at reaching better performance only, we paid special attention to BCI users' and practitioners' needs as well. In fact, BCI users are often subjects with neurological or physical impairments that can hardly bear long preparatory phases. Hence, we proposed methods to reduce the calibration time and we explored the feasibility of reducing the number of electrodes needed in the setup. To assure that our methods can be applied in standard BCI settings, we focused on models' trustworthiness, both in terms of explainability and generalizability of the proposed solutions. To achieve these goals, we leveraged two models based on convolutional neural networks. First, we reproduced a state-of-the-art CNN architecture, the BCIAUT [25], then we proposed an advancement, the P3CNET model.

As an overall result, it is worth noting that the novel architecture we are proposing (the P3CNET) overpasses the state-of-the-art BCIAUT CNN both on the dataset it was optimized for (the BCIAUT-P300), and on a new dataset (the GIB-UVa ERP-BCI). Therefore we can affirm that we built a robust and generalizable model for better P300 detection.

The first method to reduce P300 calibration time acts on reducing the signal-to-noise ratio (SNR) of the EEG recordings, finding the most informative time window around the presumed P300 occurrence. We found that the time window which maximized accuracy and minimized its standard deviation was between 200 ms and 900 ms after the occurrence of the stimulus, i.e., 400 ms less than previous experiments [25]. This is also consistent with the physiology of this evoked potential. In fact, the P300 wave is expected to appear approximately 300 ms after the occurrence of a rare stimulus [13], followed by a deflection, called N400, of a variable length. The time interval we are proposing likely captures both these events, excluding other parts of the signal that might introduce noise. In addition to the improved SNR, a shorter input signal length would result in a simpler classifier model, with fewer parameters, thus reducing the computational cost. Additionally, this approach brought improvements in P300 detection accuracy of 1.5% points for P3CNET and 4.4% points for BCIAUT, with respect to using the untrimmed EEG signal, reaching a mean accuracy of 86% for both models in the single-session training-testing. On this topic, it is worth stressing the importance of signal pre-processing driven by domain knowledge, which is crucial even in AI applications. However, the accuracy improvement observed with the signal trimming is no more valid when a different acquisition setup is leveraged, as in the case of the GIB-UVa ERP-BCI dataset. In fact, higher accuracy levels were reached when an

untrimmed signal was fed to the P3CNET model. On this topic, we can suggest that a preliminary data exploration should be performed to identify the setup-specific optimal window.

The second method to reduce calibration time focused on reducing the number of calibration sessions. Starting from the assumption that the maximum reachable accuracy is subject-dependent, a model trained on more data from the same subject would converge to that value. When the additional calibration sessions do not cause meaningful accuracy improvements, calibration can stop, and the subject is ready for using the BCI system. Both the BCIAUT and the P3CNET models suggest that the calibration should stop after three sessions. This is an improvement with respect to Amaral and colleagues' experiments [15], [16], where each session foresaw a specific training. This time gain was paid in terms of overall accuracy, but the performance seems to remain satisfactory, even with three-sessions-only calibration. Indeed, the majority of test accuracies were higher than 70% (66.67% sessions for BCIAUT, 76.67% for P3CNET), and some sessions resulted in accuracies even higher than 90% (36.67% sessions for BCIAUT, 35.00% for P3CNET). Real-world trials would help in defining the optimal balance between shorter calibration time and better P300 detection accuracy. In fact, poor detection performance would bring a worsening in BCI experience, which could become frustrating from a different point of view.

Then, we addressed the feasibility of simplifying the acquisition setup by reducing the number of EEG electrodes. To analyze the importance of each channel, it was necessary to modify the architectures by eliminating between-channel convolution. This step brought a reduction in accuracy for both BCIAUT and P3CNET, which needs to be taken into account when examining the following steps. Nonetheless, the analysis shows that parieto-occipitals electrodes (i.e., P3, Pz, P4, POz) were usually more informative than the others, in accordance with the physiology of P300 production. Indeed, the parietal and occipital lobes are responsible for the elaboration of visual stimuli [27], [45], [46].

A small reduction in accuracy could be observed when the less important channels were discarded from the training: from 81.83% to 78.23% for BCIAUT, and from 86.33% to 80.83% for P3CNET. Such an accuracy reduction, though, is quite low, confirming once more that the parietal electrodes were still providing the most critical information for P300 classification. However, we can notice high inter-subject variability in channels importance, and discarding some electrodes might unpredictably affect the performance. Therefore, we would suggest that the original configuration should be preserved unless subject-specific problems or hardware limitations would arise.

Another interesting finding was that applying transfer learning on a network pre-trained on subjects affected by Autism Spectrum Disorder (the BCIAUT-P300 dataset) substantially helped improve P300 detection accuracy for patients with motor disabilities (almost +10% accuracy). We can hypothesize that subjects suffering from a pathological condition might have some difficulties in complying with the protocol and that a pre-training is somehow helpful in catching even non-standard patterns, such as smaller or delayed P300 deflections. This is consistent with the result on healthy subjects that seem to find no

benefit from networks pre-training. Their EEG patterns are likely to show a clearer P300 wave than pathological subjects [47], [48]. Further experiments would shed light on this finding. However, considering that the focus of our work was to ameliorate the BCI process, especially for severely disabled patients, which are the typical BCI users, we can conclude that also this method was successful.

A. Limitations

A first limitation is that our bench-mark, the BCIAUT model, could not be exactly replicated, as some information was missing from the original paper. However, the superiority of the P3CNET model was consistent in all our analyses, thus allowing us to confirm that our model better suits the P300 identification task.

A second limitation is that the two CNN models, i.e., BCIAUT and P3CNET, share part of the architecture. Therefore, it is possible that similar behaviours are not due to the robustness of the methodologies but to the similarity of the networks.

Another limitation is that we aimed at simplifying the calibration procedure to improve the overall BCI user experience. However, our focus was mainly on the optimization of the training pipeline and not directly on the time needed to perform the EEG experimental procedure. For example, the optimal classification pipeline with a complex model could take more time than a sub-optimal one without significant accuracy loss.

Last, some of the observations and comparisons are mainly qualitative since the sample size was not adequate to perform robust statistical tests. In fact, the original experimental data was limited, and the number of CNN training that we could perform was constrained by the available computational resources. In future works, our approach could be translated to different BCI protocols and datasets, further extending the domain of application of the present findings.

V. CONCLUSION

In this work, we proposed different methods to improve a P300-based BCI application. We devised an architecture, P3CNET, that overpasses state-of-the-art accuracy in P300 detection from EEG signals. However, our contribution goes beyond the simple accuracy improvement. First, we stressed the importance of EEG signal pre-processing to optimize models training by indicating the best trimming for the signal. Second, we proposed a simplification of the acquisition procedure by avoiding several unnecessary re-calibrations. Third, we provide a method to potentially remove some electrodes from the acquisition setup by leveraging explainable artificial intelligence techniques. Finally, we proved the generalizability of such methods and their domain adaptation capabilities. Therefore, we proposed reliable methods that can be applied in different BCI settings to improve the overall experience, with significant benefits for users and operators.

REFERENCES

- [1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.
- [2] D. Zapala et al., "The impact of different visual feedbacks in user training on motor imagery control in BCI," *Appl. Psychophysiology Biofeedback*, vol. 43, no. 1, pp. 23–35, Oct. 2017.
- [3] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, Jan. 2012.
- [4] F. Cavrini, L. Bianchi, L. R. Quitadamo, and G. Saggio, "A fuzzy integral ensemble method in visual P300 brain-computer interface," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–9, 2016, doi: [10.1155/2016/9845980](https://doi.org/10.1155/2016/9845980).
- [5] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2image," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, doi: [10.1145/3123266.3127907](https://doi.org/10.1145/3123266.3127907).
- [6] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6809–6817.
- [7] S. Aydin, "Deep learning classification of neuro-emotional phase domain complexity levels induced by affective video film clips," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1695–1702, Jun. 2020, doi: [10.1109/jbhi.2019.2959843](https://doi.org/10.1109/jbhi.2019.2959843).
- [8] C. Guger, C. Kapeller, H. Ogawa, R. Prückl, J. Grünwald, and K. Kamada, "Electrocorticogram based braincomputer interfaces," in *Smart Wheelchairs and Brain-Computer Interfaces*, Amsterdam, The Netherlands: Elsevier, 2018, pp. 197–227.
- [9] T. W. Picton, "The P300 wave of the human event-related potential," *J. Clin. Neurophysiol.*, vol. 9, no. 4, pp. 456–479, Oct. 1992.
- [10] A. Haider and R. Fazel-Rezai, "Application of P300 event-related potential in brain-computer interface," in *Event-Related Potentials and Evoked Potentials*, London, U.K.: InTech, 2017.
- [11] L. Bianchi, C. Liti, G. Liuzzi, V. Piccialli, and C. Salvatore, "Improving P300 speller performance by means of optimization and machine learning," *Ann. Operations Res.*, Jan. 2021, doi: [10.1007/s10479-020-03921-0](https://doi.org/10.1007/s10479-020-03921-0).
- [12] S. Aliakbarhosseinabadi, E. N. Kamavuako, N. Jiang, D. Farina, and N. Mrachacz-Kersting, "Classification of EEG signals to identify variations in attention during motor task execution," *J. Neurosci. Methods*, vol. 284, pp. 27–34, Jun. 2017, doi: [10.1016/j.jneumeth.2017.04.008](https://doi.org/10.1016/j.jneumeth.2017.04.008).
- [13] E. A. Aydin, O. F. Bay, and I. Guler, "P300-based asynchronous brain computer interface for environmental control system," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 3, pp. 653–663, May 2018.
- [14] R. Bakeman and L. B. Adamson, "Coordinating attention to people and objects in mother-infant and peer-infant interaction," *Child Develop.*, vol. 55, no. 4, Aug. 1984, Art. no. 1278.
- [15] C. P. Amaral, M. A. Simões, S. Mouga, J. Andrade, and M. Castelo-Branco, "A novel brain computer interface for classification of social joint attention in autism and comparison of 3 experimental setups: A feasibility study," *J. Neurosci. Methods*, vol. 290, pp. 105–115, Oct. 2017.
- [16] C. Amaral et al., "A feasibility clinical trial to improve social attention in autistic spectrum disorder (ASD) using a brain computer interface," *Front. Neurosci.*, vol. 12, Jul. 2018, Art. no. 477.
- [17] M. Simões et al., "BCIAUT-P300: A multi-session and multi-subject benchmark dataset on autism for P300-based brain-computer-interfaces," *Front. Neurosci.*, vol. 14, Sep. 2020, Art. no. 568104.
- [18] IFMBE scientific challenge, Oct. 2019. [Online]. Available: <https://www.medicin2019.org/scientific-challenge/>
- [19] M. Bittencourt-Villalpando and N. M. Maurits, "Linear SVM algorithm optimization for an EEG-based brain-computer interface used by high functioning autism spectrum disorder participants," in *Proc. IFMBE*, Berlin, Germany: Springer, 2019, pp. 1875–1884.
- [20] H. Zhao, S. Yu, J. Prinable, A. McEwan, and P. Karlsson, "A feasible classification algorithm for event-related potential (ERP) based brain-computer-interface (BCI) from IFMBE scientific challenge dataset," in *Proc. IFMBE*, Berlin, Germany: Springer, Sep. 2019, pp. 1861–1868.
- [21] L. de Arancibia, P. Sánchez-González, E. J. Gómez, M. E. Hernando, and I. Oropesa, "Linear vs nonlinear classification of social joint attention in autism using VR p300-based brain computer interfaces," in *Proc. IFMBE*, Berlin, Germany: Springer, 2019, pp. 1869–1874.
- [22] D. Krzemiński, S. Michelmann, M. Treder, and L. Santamaria, "Classification of P300 component using a riemannian ensemble approach," in *Proc. IFMBE*, Berlin, Germany: Springer, Sep. 2019, pp. 1885–1889.
- [23] B. Chatterjee, R. Palaniappan, and C. N. Gupta, "Performance evaluation of manifold algorithms on a P300 paradigm based online BCI dataset," in *Proc. IFMBE*, Berlin, Germany: Springer, Sep. 2019, pp. 1894–1898.
- [24] V. S. Adama, B. Schindler, and T. Schmid, "Using time domain and pearson's correlation to predict attention focus in autistic spectrum disorder from EEG p300 components," in *Proc. Int. Federation Med. Biol. Eng. (IFMBE)*, Berlin, Germany: Springer, Sep. 2019, pp. 1890–1893.

- [25] D. Borra, S. Fantozzi, and E. Magosso, "Convolutional Neural Network for a P300 Brain-Computer Interface to Improve Social Attention in Autistic Spectrum Disorder," in *Proc. IFMBE*, Berlin, Germany: Springer, Sep. 2019, pp. 1837–1843.
- [26] D. Valeriani, R. Poli, and C. Cinel, "A collaborative brain-computer interface for improving group detection of visual targets in complex natural environments," in *Proc. 7th Int. IEEE/EMBS Conf. Neural Eng.*, 2015, pp. 25–28, doi: [10.1109/ner.2015.7146551](https://doi.org/10.1109/ner.2015.7146551).
- [27] C. Babiloni, F. Vecchio, M. Miriello, G. L. Romani, and P. M. Rossini, "Visuo-spatial consciousness and parieto-occipital areas: A high-resolution EEG study," *Cereb. Cortex*, vol. 16, no. 1, pp. 37–46, 2005.
- [28] S. Katsigiannis, P. Arnau-Gonzalez, M. Arevalillo-Herraez, and N. Ramzan, "Single-channel EEG-based subject identification using visual stimuli," in *Proc. IEEE Int. Conf. Biomed. Health Inform. Conf.*, 2021, pp. 1–4.
- [29] L. Floridi, "Establishing the rules for building trustworthy AI," *Nature Mach. Intell.*, vol. 1, no. 6, pp. 261–262, 2019.
- [30] A. Holzinger, "From machine learning to explainable AI," in *Proc. IEEE World Symp. Digit. Intell. Syst. Machines*, 2018, pp. 55–66.
- [31] E. Santamaría-Vázquez, V. Martínez-Cagigal, F. Vaquerizo-Villar, and R. Hornero, "EEG-inception: A novel deep convolutional neural network for assistive ERP-based brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2773–2782, Dec. 2020.
- [32] V. J. Lawhern, J. S. Amelia, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, pp. 2–5, Nov. 2016.
- [33] B. Richhariya and M. Tanveer, "EEG signal classification using universum support vector machine," *Expert Syst. Appl.*, vol. 106, pp. 169–182, Sep. 2018, doi: [10.1016/j.eswa.2018.03.053](https://doi.org/10.1016/j.eswa.2018.03.053).
- [34] S. K. R. Singanamalla and C.-T. Lin, "Spiking neural network for augmenting electroencephalographic data for brain computer interfaces," *Front. Neurosci.*, vol. 15, Apr. 2021, Art. no. 651762, doi: [10.3389/fnins.2021.651762](https://doi.org/10.3389/fnins.2021.651762).
- [35] R. Xu, N. Jiang, C. Lin, N. Mrchacz-Kersting, K. Dremstrup, and D. Farina, "Enhanced low-latency detection of motor intention from EEG for closed-loop brain-computer interface applications," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, pp. 288–296, Feb. 2014, doi: [10.1109/tbme.2013.2294203](https://doi.org/10.1109/tbme.2013.2294203).
- [36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations (ICLR 2015)*, *Comput. Biol. Learn. Soc.*, 2019, pp. 1–14.
- [37] H. Shan, Y. Liu, and T. Stefanov, "A simple convolutional neural network for accurate P300 detection and character spelling in brain computer interface," in *Proc. 27th Int. Joint Conf. Artif. Intell. Int. Joint Conf. Artif. Intell. Org.*, Jul. 2018, pp. 1604–1610.
- [38] B. E. Yerys, G. L. Wallace, J. L. Sokoloff, D. A. Shook, J. D. James, and L. Kenworthy, "Attention deficit/hyperactivity disorder symptoms moderate cognition and behavior in children with autism spectrum disorders," *Autism Res.*, vol. 2, no. 6, pp. 322–333, 2009.
- [39] D. Ravi *et al.*, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.
- [40] B. Abibullaev and A. Zollanvari, "Learning discriminative spatio-spectral features of ERPs for accurate brain-computer interfaces," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2009–2020, May 2019.
- [41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [42] E. Santamaría-Vázquez, V. Martínez-Cagigal, and R. Hornero, "Gib-UVA ERP-BCI dataset," 2020. [Online]. Available: <https://dx.doi.org/10.21227/6bdr-4w65>
- [43] R. Singh, T. Ahmed, A. Kumar, A. K. Singh, A. K. Pandey, and S. K. Singh, "Imbalanced breast cancer classification using transfer learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 83–93, Jan./Feb. 2020. doi: [10.1109/tcbb.2020.2980831](https://doi.org/10.1109/tcbb.2020.2980831).
- [44] A. Antonietti, I. Dag, L. G. Dui, S. Ferrante, and A. Pedrocchi, "Data for: Leveraging deep learning techniques to improve P300-based brain computer interfaces," 2021. [Online]. Available: <https://dx.doi.org/10.21227/y0t4-hr09>
- [45] D. E. J. Linden, "The P300: Where in the brain is it produced and what does it tell us," *Neuroscientist*, vol. 11, no. 6, pp. 563–576, Dec. 2005.
- [46] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clin. Neurophysiol.*, vol. 118, no. 10, pp. 2128–2148, Oct. 2007.
- [47] M. T. Medina-Julia, A. Fernandez-Rodríguez, F. Velasco-Alvarez, and R. Ron-Angevin, "P300-based brain-computer interface speller: Usability evaluation of three speller sizes by severely motor-disabled patients," *Front. Hum. Neurosci.*, vol. 14, 2020, Art. no. 433.
- [48] A. Thavasimuthu, N. Shanthi, R. Sathiyasheelan, G. Emayavaramban, and T. Rajendran, "Brain-computer interface for persons with motor disabilities - a review," *Open Biomed. Eng. J.*, vol. 13, pp. 127–133, Dec. 2019. [Online]. Available: <https://openbiomedicalengineeringjournal.com/VOLUME/13/PAGE/127/>