Accepted version of an article published in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28 (12), pp. 2773-2782
DOI: https://doi.org/10.1109/TNSRE.2020.3048106

EEG-Inception: A Novel Deep Convolutional Neural Network for Assistive ERP-based Brain-Computer Interfaces

Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Fernando Vaquerizo-Villar, and Roberto Hornero, *Senior Member, IEEE*

***Abstract*—In recent years, deep-learning models gained attention for electroencephalography (EEG) classification tasks due to their excellent performance and ability to extract complex features from raw data. In particular, convolutional neural networks (CNN) showed adequate results in brain-computer interfaces (BCI) based on different control signals, including event-related potentials (ERP). In this study, we propose a novel CNN, called EEG-Inception, that improves the accuracy and calibration time of assistive ERP-based BCIs. To the best of our knowledge, EEG-Inception is the first model to integrate Inception modules for ERP detection, which combined efficiently with other structures in a light architecture, improved the performance of our approach. The model was validated in a population of 73 subjects, of which 31 present motor disabilities. Results show that EEG-Inception outperforms 5 previous approaches, yielding significant improvements for command decoding accuracy up to 16.0%, 10.7%, 7.2%, 5.7% and 5.1% in comparison to rLDA, xDAWN + Riemannian geometry, CNN-BLSTM, DeepConvNet and EEGNet, respectively. Moreover, EEG-Inception requires very few calibration trials to achieve state-of-the-art performances taking advantage of a novel training strategy that combines cross-subject transfer learning and fine-tuning to increase the feasibility of this approach for practical use in assistive applications.**

***Index Terms*—Brain–computer interfaces, event-related potentials, P300, deep learning, convolutional neural networks, Inception, transfer learning.**

I. Introduction

BRAIN–computer interfaces (BCI) enable direct communication between humans and external devices through the analysis of neural signals [**?**]. These systems have a wide range of applications, including neurorehabilitation, cognitive training or entertainment[**?**]. However, the most extended use is the control of assistive technologies to improve the autonomy and quality of life of people with severe motor disabilities, caused by spinal cord injury, head trauma, stroke, multiple sclerosis, lateral amyotrophic sclerosis or cerebral palsy, among others [**?**]. In practice, most BCIs use the electroencephalography (EEG) to record brain activity for its relative low cost, ease of use and noninvasiveness [**?**]. This technique uses electrodes placed over the scalp to register the electrical activity of superficial neurons, reflecting the ongoing brain processes with exceptional temporal resolution [**?**]. Nonetheless, EEG also has important drawbacks. For instance, it presents poor spatial resolution, recording activity from volumes between 1 and 10 cm^3 around each electrode [**?**]. Moreover, it is very sensitive to a wide range of noisy artefacts, such as muscular movements, corneo-retinal standing potentials or cardiac activity [**?**]. Therefore, EEG is an extremely complex and noisy signal, making direct decoding of individual brain processes almost impossible.

In order to overcome this limitation, BCIs usually rely on control signals with higher signal-to-noise ratio (SNR) to extract discriminative information from the EEG. In particular, visual event-related potentials (ERP) are waveforms that reflect the ongoing brain activity happening just after the perception of an external visual stimuli [**?**]. ERP-based BCIs, also known as spellers, were first proposed by Farwell and Donchin [**?**] in 1988 as a communication aid for people suffering from severe physical disability. These systems allow to select commands from a predefined set of options by detecting ERPs in the EEG of the user. The main advantages of these systems in comparison with other BCIs are the very low mental workload needed to use them, the high number of commands that can be discriminated and their versatility for a wide range of applications. However, the accuracy and maximum amount of information conveyed by ERP-based spellers per unit of time is limited due to the low SNR of ERPs, non-stationary properties of the EEG over time, and strong inter-subject variability [**?**].

A large number of studies addressed this issue by improving pattern recognition algorithms for ERP detection [**?**]. Methods based on linear discriminant analysis (LDA) and support-vector machine are the most classical approaches [**?**], [**?**], [**?**], [**?**], [**?**]. They achieve reasonable performance, especially when a large amount of training examples is available. However, these methods are noise-sensitive and show poor inter-

This work was supported by 'Ministerio de Ciencia, Innovación y Universidades' and 'European Regional Development Fund (FEDER)' under project DPI2017-84280-R, by 'European Commission' and 'FEDER' under projects 'Análisis y correlación entre el genoma completo y la actividad cerebral para la ayuda en el diagnóstico de la enfermedad de Alzheimer' and 'Análisis y correlación entre la epigenética y la actividad cerebral para evaluar el riesgo de migraña crónica y episódica en mujeres' ('Cooperation Programme Interreg V-A Spain-Portugal POCTEP 2014–2020'), and by 'CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)' through 'Instituto de Salud Carlos III (ISCIII)' co-funded with FEDER funds. V. Martínez-Cagigal was in receipt of a PIF grant from the University of Valladolid. E. Santamaría-Vázquez was in a receipt of a PIF grant from the 'Consejería de Educación de la Junta de Castilla y León', and the European Social Fund. F. Vaquerizo-Villar was in receipt of a 'Ayuda para contratos predoctorales para la Formación de Profesorado Universitario (FPU)' grant from the Ministerio de Educación, Cultura y Deporte (FPU16/02938).

The authors are with the Biomedical Engineering Group (GIB), E.T.S Ingenieros de Telecomunicación, University of Valladolid, Paseo de Belén 15, 47011, Valladolid, Spain and CIBER-BBN (ISCIII), Spain (e-mail: eduardo.santamaria@gib.tel.uva.es; victor.martinez@gib.tel.uva.es; fernando.vaquerizo@gib.tel.uva.es; robhor@tel.uva.es)

subject and inter-session generalization [?]. Moreover, several studies have demonstrated that their performance is reduced drastically in real environments with severely disabled populations [?], [?]. Additionally, novel tensor classifiers based on Riemannian geometry (RG) and adaptive methodologies have led to important improvements in BCIs in recent years [?], [?].

Nonetheless, the field of pattern recognition is living breakthrough advances thanks to the development of deep learning [?]. Deep neural networks are able to automatically extract complex features from raw data, learning hierarchical representations of the input in different levels of abstraction [?]. This ability has revolutionized fields such as computer vision, natural language processing, genomics or drug-discovery [?]. In the EEG domain, deep neural networks have been used for ERP, sensorymotor rhythms (SMR) and steady-state visual evoked potentials (SSVEP) classification, seizure detection and prediction, sleep stage scoring, mental workload detection, data augmentation or emotion recognition, among others [?]. However, there are still relatively few studies of deep-learning models for EEG analysis [?]. In fact, whether they truly present advantages over classical methods remain as an open question [?].

Since the application of deep learning to EEG is relatively new, most studies explored simple architectures based on convolutional neural networks (CNN) and recurrent neural networks (RNN) [?]. Nevertheless, there is still room for improvement, and more complex architectures specifically designed for EEG analysis could enhance the performance of CNNs in this domain. For instance, Inception modules proposed by Szegedy *et al.* [?] for computer vision allow a multiscale analysis of the input data by using convolutional layers in parallel with different kernel sizes. Empirically, this structure has proved to extract richer feature maps with less computational cost, boosting performance while keeping a reasonable training and evaluation time [?]. In this regard, EEG is made up of transient and oscillatory patterns of different temporal lengths that reflect the ongoing activity of the brain, making multiscale analysis through Inception modules especially suitable for this signal. However, there are still very few studies that explored Inception-based architectures for EEG processing. Qiao *et al.* [?] and Lee *et al.* [?] proposed Inception-based architectures to decode movements from motor imagery and motor execution tasks, respectively. In both cases, their models yielded higher performance than other state-of-the-art methods. Additionally, Yue *et al.* [?] designed a novel method to evaluate visual fatigue using a 3D Inception CNN. Nevertheless, it has not yet been studied whether Inception modules are useful for ERP-based spellers. ERPs are generally considered to have most of their power in delta (0-4 Hz) and theta (4-7 Hz) bands [?]. Thus, the majority of algorithms for ERP detection do not take into account the upper bands (i.e., alpha, beta and gamma), most even filtering frequencies above 10 Hz [?], [?], [?]. However, several studies found that these bands contain discriminative information for target vs non-target conditions, especially in the gamma range [?]. Therefore, algorithms for ERP detection should be able to process patterns at very different temporal scales in order to take full advantage of all the discriminative information available. In this respect, previous CNNs for ERP detection used one kernel size per layer, which may limit their capability to learn features at different temporal scales [?], [?], [?], [?], [?], [?], [?]. Additionally, there are other open challenges for deep learning in the EEG domain. Especially, the lack of large datasets has hindered the development of novel training strategies with great impact in other fields, such as transfer learning and fine-tuning [?]. These techniques have great potential to reduce the calibration time required by deep-learning models in BCI applications. However, to our knowledge, they have not been explored for this purpose in ERP-based spellers yet.

The main goal of this study is to design, develop and test a novel CNN in order to improve the accuracy and calibration time of ERP-based spellers for practical use. To this end, our model, called EEG-Inception, efficiently integrates Inception modules and other structures optimized for EEG processing. Additionally, the dataset used in this study had a total of 701615 observations from 73 subjects (42 controls, 31 severely disabled), the largest sample among related studies. Taking advantage from this dataset, we designed a novel training strategy using cross-subject transfer learning and fine-tuning to reduce the calibration time required for test subjects. Finally, we provide a direct and fair comparison with the most successful previous approaches in ERP-based spellers. Noteworthy, most studies do not evaluate novel classification algorithms with severely disabled subjects, the end users of these systems. This is a common limitation in the BCI literature, likely caused by the lack of public datasets with disabled subjects [?]. In order to address this issue, we released the code and database used in this study in http://dx.doi.org/10.21227/6bdr-4w65 [?], providing a new public benchmark for ERP-based spellers.

II. Related work

This section presents an overview of previous studies that proposed deep-learning models for ERP detection in a BCI framework [?], [?], [?], [?], [?], [?], [?]. Table I summarizes these approaches, highlighting their main contributions, evaluation approach, subjects and results. As can be seen, CNNs are the most popular approach. Among them, the model proposed by Lawhern *et al.* [?], called EEGNet, stands out as one of the most successful, using depthwise and separable convolutions to provide a robust and compact architecture. In fact, this model won the scientific challenge for ERP detection launched by International Federation of Medical and Biological Engineering (IFMBE) in 2019 [?]. Noteworthy, the second position was achieved by other deep-learning model, CNN-BLSTM, which combines a convolutional layer to extract spatial patterns with 2 recurrent layers based on bidirectional long-short term memory units (BLSTM) to learn temporal patterns [?]. The use of RNNs for EEG processing is scarce, especially for ERP detection [?]. This is probably because these architectures are very expensive in computational terms, taking considerably longer times than CNNs to train. Nevertheless, RNNs are specifically designed to process temporal series, which makes them a promising alternative to CNNs for EEG processing.

TABLE I
OVERVIEW OF PREVIOUS DEEP-LEARNING APPROACHES FOR ERP-BASED SPELLERS

Study	Highlights	Training/testing approach	Total subjects	Accuracy
Cecotti *et al.* 2011 [?]	First CNN for BCI classification tasks	Intra-subject	2 CS	95%
Manor *et al.* 2015 [?]	Spatio-temporal regularization	Cross-subject	15 CS	70%
Liu *et al.* 2018 [?]	Dropout and Batch normalization	Intra-subject	3 CS	97%
Lawhern *et al.* 2018 [?]	Depthwise and separable convolutions	Cross/intra-subject	18 CS	90/92%
Santamaría-Vázquez *et al.* 2019 [?]	Use of bidirectional LSTM layers	Hybrid	15 ADHD	84%
Borra *et al.* 2019 [?]	Depthwise and separable convolutions	Intra-subject	15 ADHD	92%

CS: control subjects; CNN: convolutional neural network; ADHD: subjects with attention deficit hyperactivity disorder; MDS: motor disabled subjects; LSTM: long-short term memory; Training/testing approach: intra-subject strategies train and test the models with data from the same subject, cross-subject strategies train and test the models with data from different subjects and hybrid approaches combine both techniques; Accuracy: test command decoding accuracy.

It should be noted that, to the best of our knowledge, previous studies failed to test their deep-learning models with motor-disabled subjects. Moreover, Liu *et al.* [**?**] and Cecotti *et al.* [**?**] only included 2 and 3 healthy subjects in their studies. In this regard, it is well-known that patients generally achieve lower classification accuracies due to individual aspects related to their diseases, such as neural damage, visual impairment, limited sustained attention abilities, involuntary tremors or limited cognitive performance, among others [**?**], [**?**]. Furthermore, these symptoms are highly variable between individuals, even between those with the same condition, making severely disabled subjects especially heterogeneous and challenging for ERP detection. Therefore, comprehensive evaluation in this group is required to assess the performance of new models for assistive BCI applications.

III. MATERIALS AND METHODS

A. Subjects and Signals

Seventy three subjects participated in this study: 42 healthy controls (mean age: 25.1 ± 4.3 years; 31 males) and 31 severely disabled (mean age: 44.2 ± 7.7 years; 20 males) suffering from different conditions (5 spinal cord injury; 4 Friedreich's ataxia; 4 cerebral palsy; 2 polymalformative syndrome; 1 stroke; 15 multiple sclerosis). The dataset was built on data from previous studies [**?**], [**?**], [**?**]. In all cases, subjects performed several sessions with an ERP-based speller using the row-column paradigm (RCP) [**?**]. In the RCP, the system displays a matrix of commands, whose rows and columns are highlighted randomly. Typically, rows and columns are highlighted several times (i.e., sequences) to increase the accuracy of the system. To select a command, the user has to stare at the desired option, eliciting an ERP when a stimulus is perceived. Finally, the system decodes the row and the column using signal processing algorithms to detect the ERPs and executes the corresponding command, providing feedback to the user. Table II summarizes the key aspects of the database according to the experimental protocols of each study. Noteworthy, data from different sessions were mixed to simplify the analysis. For further details, see the corresponding studies [**?**], [**?**], [**?**].

Participants were split into 3 sets: training, validation and test. Healthy subjects were randomly divided into 2 groups, assigning 80% of them to the training set and 20% to the validation set. The remaining 31 disabled subjects were assigned to the test set. This distribution is designed provide a real estimation of the performance of EEG-Inception through evaluation in people with severe disabilities, who are the end users of ERP-based spellers. Table III shows the number of subjects, trials, and observations of each set. Noteworthy, no method was applied to handle the inherent class imbalance associated to datasets from ERP-based spellers.

EEG signals were acquired with a g.USBamp (g.Tec, *Guger Technologies*, Austria) at a sampling frequency of 256 Hz using 8 active electrodes placed at Fz, Cz, Pz, P3, P4, PO7, PO8 and Oz, the ground placed at FPz and the reference in the earlobe, according to the International System 10–10 [**?**]. During the experiments, 2 different platforms were used to present the stimuli and save the signals: BCI2000 [**?**], and MEDUSA [**?**]. The experimental protocol was approved by the corresponding local ethics committee, and all participants gave their informed consent.

B. Novel CNN: EEG-Inception

EEG-Inception is a novel CNN inspired on the work of Szegedy *et al.* [21] for image classification, adapting its concepts to provide an enhanced architecture for EEG processing and ERP detection.

In order to prepare the raw EEG, a simple preprocessing pipeline was applied. First, signals were decimated to 128 Hz to reduce the computational cost of the model [**?**]. Afterward, a band-pass filter was applied between 0.5 and 45 Hz, keeping the most discriminative information and eliminating the power line frequency at 50 Hz [**?**]. Common average reference (CAR) spatial filter was also applied to improve the SNR of ERPs [**?**]. Finally, epochs were extracted from 0 to 1000 ms after the stimulus onset [**?**]. Thus, the model input is an array of shape 128×8, being the first dimension the temporal axis (i.e., samples), whereas the second dimension corresponds to the spatial axis (i.e., channels).

EEG-Inception includes several concepts adapted from the image classification domain, including Inception modules to capture dependencies between features at different scales and depthwise convolutions. Fig. 1 depicts a visual overview of the architecture, whereas Table IV details the configuration and architectural choices. It should be noted that each convolutional block (i.e., 2D convolutions and depthwise 2D convolutions) includes batch normalization to normalize the feature maps [**?**], activation function to introduce non-linearities [**?**] and dropout regularization [**?**] to prevent overfitting.

The architecture of EEG-Inception, is organized in three main blocks:

1) Inception module 1: this module processes the signal in 3 different temporal scales for each EEG channel, according

TABLE II
DESCRIPTION OF THE DATASET

Study	Subjects	Sessions	Trials/Subject	Paradigm	SD	ISI	Description
Martínez-Cagigal *et al.* 2017 [**?**]	10CS 15MD	4	87.9±7.3	RCP	62.5	125–250 †	Spelling task to control a BCI web browser. Two matrices with dimensions 5×3 and 9×5 were used.
Martínez-Cagigal *et al.* 2019 [**?**]	10CS 16MD	3	63.4±8.2	RCP	62.5	125–250 †	Spelling task for smartphone control. Two matrices with dimensions 4×4 and 6×9 were used.
Santamaría-Vázquez *et al.* 2019 [**?**]	22CS	2	60	RCP	75	100	Simple spelling task with an alphanumeric 6×6 matrix.

CS: control subjects; MD: motor-disabled; RCP: row-colmn paradigm; SD: stimulus duration in ms, ISI: inter-stimulus interval in ms. Each trial represents a command or character selection, while the number of sequences is the total stimuli per row and column of the matrix. Column "Trials/Subject" displays the total number of trials per subject mixing all sessions. † Random value in the specified range.

TABLE III
CHARACTERISTICS OF EACH SET

Subset	Subjects	Trials	Observations
Training	34 CS	2188	315159
Validation	8 CS	502	77046
Test	31 MDS	2333	309410

CS: control subjects. MDS: motor disabled subjects. Characteristics of training, validation and test sets. Trials represent a single command selection. Observations represent a single stimulus.

to the kernel sizes of the convolutional blocks C1, C2 and C3, which are 64×1, 32×1 and 16×1, respectively. Therefore, since the sampling rate of the input is 128 Hz, these sizes correspond to temporal windows of 500 ms, 250 ms and 125 ms. Following these layers, D1, D2 and D3 process the signal in the spatial domain using depthwise convolutions. Depthwise convolutions were first used in the image classification domain to factorize a convolution kernel into smaller kernels by acting on each input channel separately, reducing the total amount of parameters [**?**], [**?**]. When applied to EEG processing, they provide a method to learn optimal spatial filters (i.e., channels weights) for each temporal pattern extracted by the previous layer [**?**]. Then, the concatenation layer N1 merges the output features from D1, D2 and D3. Finally, an average pooling is applied for dimensionality reduction.

2) Inception module 2: this module is organized as the previous one. It is formed by 3 branches that process the EEG signal in 3 temporal scales of 500 ms, 250 ms and 125 ms. It should be noticed that, after the average pooling layer of the first block, these scales correspond to kernel sizes of 16×1, 8×1 and 4×1. This module extracts additional temporal features in a higher level of abstraction, considering all EEG channels. As previously, the outputs of convolutional blocks C4, C5 and C6 are concatenated. Then, an average pooling for dimensionality reduction is also applied.

3) Output module: the last 2 convolutional layers are designed to extract the most meaningful patterns for the final classification, compressing the information into few features. Noteworthy, the number of filters is decreased progressively, which, along with the average pooling layers, reduces the dimensionality in order to avoid overfitting. In fact, only 24 features are fed to the final classification layer. Finally, the softmax output estimates the probability for each class (target and non-target) [**?**].

The model was trained using the following configuration: Adam optimizer with default hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ [**?**]; categorical cross entropy loss function [**?**]; mini-batch size of 1024; and 500 epochs. To speed up training and avoid overfitting, we applied early stopping [**?**] when the loss of the validation set did not improve for 10 consecutive epochs, restoring the weights that minimized this metric.

The choice of hyperparameters in deep-learning models is crucial to achieve suitable results [**?**]. In particular, the learning rate, the activation function and the dropout rate have great impact [**?**]. In order to reach an optimal solution, they were automatically optimized on the validation set using grid search. The rest of hyperparameters (i.e., number of layers, number of branches in Inception modules, number of filters, kernel sizes and pooling sizes) were heuristically chosen.

C. Comparison models

1) Regularized LDA (rLDA): LDA-based approaches are widely used for ERP detection due to their simplicity and performance [**?**], [**?**], [**?**]. Specifically, rLDA was proposed for single-trial analysis and classification of ERP components by Blankertz *et al.* [**?**], achieving acceptable results. This model is a regularized version of LDA by means of shrinkage estimators which showed advantages over other LDA-based approaches, such as stepwise LDA (SWLDA) [**?**]. Thus, rLDA is a suitable benchmark for more advanced models. As preprocessing stage, band pass filtering between 0.5 and 10 Hz and CAR were applied. Then, epochs were extracted from 0 to 1000 ms after the stimulus onset and decimated to 20 Hz [**?**]. EEG channels were concatenated to arrange the final feature vector, which was fed to rLDA for classification. We used the scikit-learn implementation of rLDA, which implements the O. Ledoit and M. Wolf's formula to calculate the shrinkage parameter [**?**].

2) xDAWN + RG: models based on RG have gained importance for BCI applications in recent years due to their robustness and transfer learning capabilities [**?**]. Here, we used the algorithm that won the "BCI Challenge NER 2015" for comparison purposes [**?**]. This algorithm combines xDAWN spatial filtering with RG to estimate covariances matrices and project them into the tangent space, followed by a logistic regression classifier to achieve a robust ERP classification [**?**]. In this case, we used the same preprocessing pipeline as for EEG-Inception. Please refer to [**?**] for further information and an open source implementation.

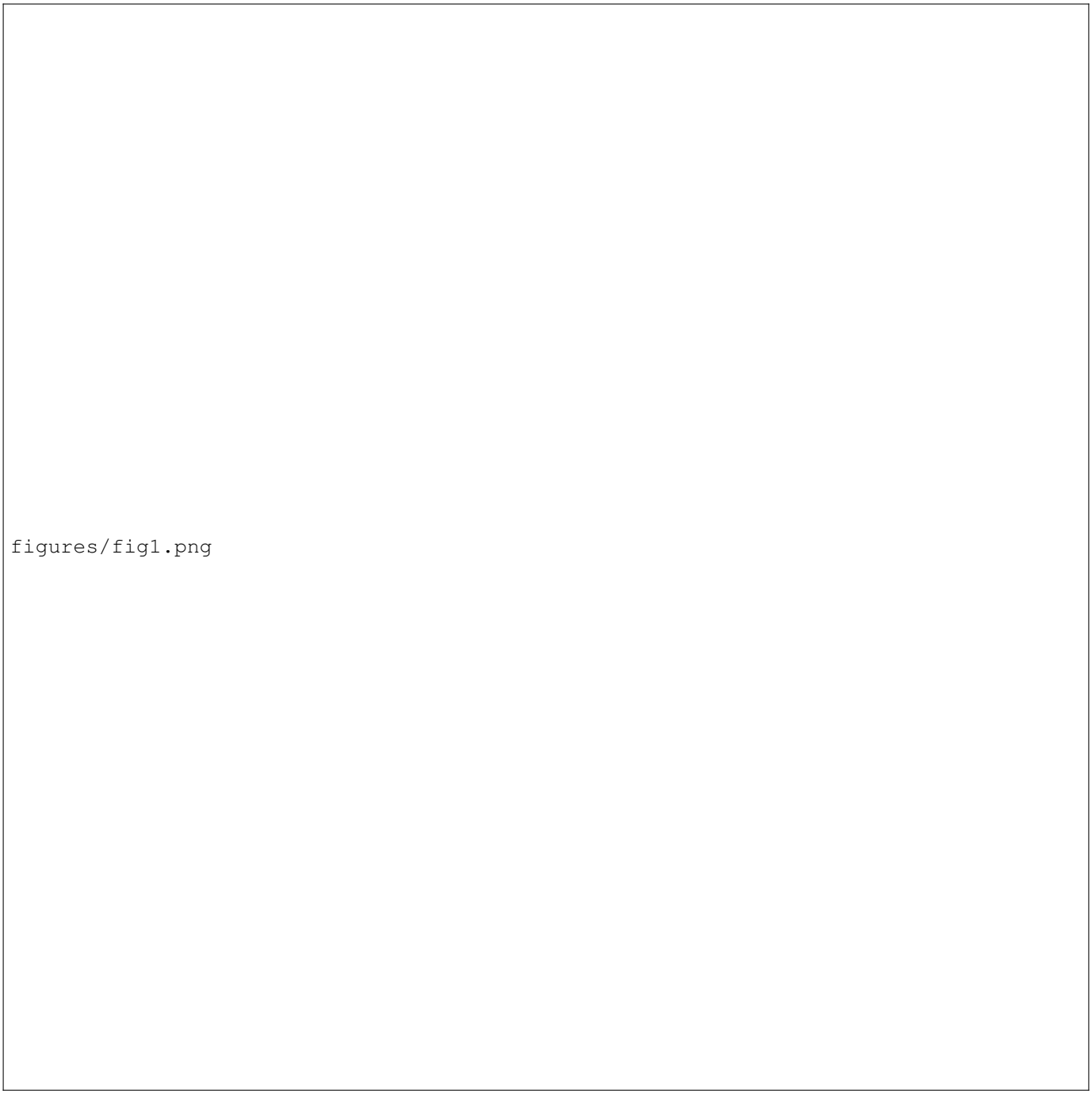

Fig. 1. Overview of EEG-Inception architecture. 2D convolution blocks and depthwise 2D convolution blocks include batch normalization, activation and dropout regularization. The kernel size is displayed for convolutional and average pooling layers.

3) CNN-BLSTM: this model was proposed to take part in the IFMBE scientific challenge launched in 2019 for ERP classification [**?**]. In this competition, CNN-BLSTM reached the second position, beating 7 other approaches [**?**]. CNN-BLSTM combines 1D convolutional layers to extract spatial features with 2 BLSTM layers to detect temporal patterns. As commented previously, the use of RNNs for EEG processing is scarce [**?**]. Therefore, CNN-BLSTM is an interesting alternative that adds variety to the comparison performed in this study. The same preprocessing was applied as for EEG-Inception.

4) DeepConvNet: this CNN was proposed by Schirrmeister *et al.* [**?**] as a generic model for EEG decoding tasks in BCI. DeepConvNet comprises 5 convolution blocks which include max-pooling layers, with a special first block designed to handle EEG input, followed by a dense softmax classification layer [**?**]. In this work, we used the implementation proposed by Lawhern *et al.* [**?**], which is publicly available. The same preprocessing was applied as for EEG-Inception.

TABLE IV
EEG-INCEPTION ARCHITECTURE DETAILS

Block	Type	Filters	Depth	Kernel	Padding	Output shape	Connected to	Role
IN	Input	–	–	–	–	$128 \times 8 \times 1$	C1, C2, C3	Input
C1	Conv2D	8	–	64×1	Same	$128 \times 8 \times 8$	D1	Temporal analysis
D1	DepthwiseConv2D	–	2	1×8	Valid	$128 \times 1 \times 16$	N1	Spatial analysis
C2	Conv2D	8	–	32×1	Same	$128 \times 8 \times 8$	D2	Temporal analysis
D2	DepthwiseConv2D	–	2	1×8	Valid	$128 \times 1 \times 16$	N1	Spatial analysis
C3	Conv2D	8	–	16×1	Same	$128 \times 8 \times 8$	D3	Temporal analysis
D3	DepthwiseConv2D	–	2	1×8	Valid	$128 \times 1 \times 16$	N1	Spatial analysis
N1	Concatenate	–	–	–	–	$128 \times 1 \times 48$	A1	Concatenation
A1	AveragePooling2D	–	–	4×1	–	$32 \times 1 \times 48$	C4, C5, C6	Concatenation
C4	Conv2D	8	–	16×1	Same	$32 \times 1 \times 8$	N2	Temporal analysis
C5	Conv2D	8	–	8×1	Same	$32 \times 1 \times 8$	N2	Temporal analysis
C6	Conv2D	8	–	4×1	Same	$32 \times 1 \times 8$	N2	Temporal analysis
N1	Concatenate	–	–	–	–	$32 \times 1 \times 24$	A2	Concatenation
A2	AveragePooling2D	–	–	2×1	–	$16 \times 1 \times 24$	C7	Dimension reduction
C7	Conv2D	12	–	8×1	Same	$16 \times 1 \times 12$	A3	Temporal analysis
A3	AveragePooling2D	–	–	2×1	–	$8 \times 1 \times 12$	C8	Dimension reduction
C8	Conv2D	6	–	4×1	Same	$8 \times 1 \times 6$	A4	Temporal analysis
A4	AveragePooling2D	–	–	2×1	–	$4 \times 1 \times 6$	C7	Dimension reduction
OUT	Dense	–	–	–	–	2	–	Softmax output

Column "Type" describes the class used to implement each block in Keras framework. It should be taken into account that this implementation may vary across different frameworks. All convolutional blocks (i.e., Conv2D and DethpwiseConv2D) include batch normalization, activation and dropout regularization. The model has 15154 parameters, of which 14926 are fitted during training.

5) EEGNET: this CNN proposed by Lawern *et al.* [**?**] is especially designed for BCI classification tasks, keeping a compact and robust architecture that has been tested with different BCI paradigms (i.e., ERP-based speller, SMR, movement-related cortical potentials and feedback error-related negativity). This network uses batch normalization and dropout to avoid overfitting and average pooling for dimensionality reduction. The same preprocessing was applied as for EEG-Inception. For training and testing the models, we used the open source implementation of EEGNet-8-2 provided by Lawern *et al.* [**?**] with the same hyperparameters for ERP detection as the original study.

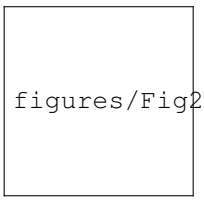

Fig. 2. Work-flow of the study. First, hyperparameter optimization process of EEG-Inception was done based on the performance on the validation set. Then, the models were trained using the training set. Finally, they were evaluated in the test set using a fine-tuning process for each subject. Fine-tuning trials were randomly selected, and the process is repeated 100 times for each N and subject.

D. Evaluation Experiment

The evaluation experiment was designed to meet real-life conditions, where the available amount of training data from a single subject is usually limited. Fig. 2 offers a graphic description of the different stages of the study, including the hyperparameter optimization process and the training and testing phases. Firstly, all models were trained using the training set. Then, we evaluated them in the test set using a fine-tuning process for each subject with $N = \{0, 5, 10, 20, 30\}$ trials. For $N = 0$, models are directly evaluated in the test set, simulating a plug & play device and thus assessing their robustness to inter-subject variability. For $N > 0$, the fine-tuning process has 3 stages: (*i*) the algorithm picks N trials from each subject randomly; (*ii*) the trained model is fine-tuned with the data from these trials (deep-learning models are initialized with the original weights after the training phase), obtaining a subject-specific model; and (*iii*) the fine-tuned model is tested with the rest of trials for each subject. This procedure was repeated 100 times for each N and subject, averaging the obtained results.

The proposed evaluation experiment reproduces a real-life setting, where the number of training trials for end users should be under 30 (approximately 20 minutes of effective training) [**?**], [**?**]. Longer calibration times would reduce the usability of an ERP-based speller radically. Therefore, the proposed experimental approach, together with the distribution of the dataset, allows to assess the effectiveness of our training strategy, which combines cross-subject transfer learning and fine-tuning to reduce the calibration time and overfitting for end users of ERP-based spellers.

All implementations were programmed in Python. Keras v2.3 framework with Tensorflow v2.0 backend was used to implement deep-learning models [**?**]. For all experiments, we used a desktop computer with the following hardware characteristics: Intel Core i9-9900 @ 3.6 GHz, 64 GB RAM, NVIDIA RTX 2080Ti 11GB.

IV. RESULTS

A. Hyperparameter optimization

Learning rate (lr), activation function (f_{act}) and dropout rate (p_{drop}) were optimized for EEG-Inception. The search space for each hyperparameter was: $lr = \{0.01, 0.001, 0.0001\}$; $f_{\text{act}} = \{\text{Sigmoid}, \text{ReLU}, \text{ELU}\}$; $p_{\text{drop}} = \{0.00 : 0.05 : 0.5\}$. Fig. 3 depicts the results of the optimization process. Noteworthy, the optimization score was computed as the averaged command decoding accuracy in the validation set after a fine-tuning process with $N = 30$ for each subject, considering

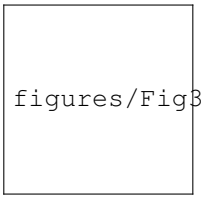

Fig. 3. Results of the optimization process of EEG-Inception in the validation set (8 healthy subjects). Each line depicts the mean command decoding accuracy after a fine-tuning process for each subject with $N = 30$, considering 5 sequences of stimulation. (a) Score for each activation function and dropout rate using the best value of learning rate ($lr = 0.001$). (b) Score for each learning rate and dropout rate using the best activation function (ELU).

5 sequences of stimulation. The values that maximized this score were selected as optimal and used in the evaluation experiment. Specifically, the optimal set was $lr = 0.001$; $f_{\text{act}} = ELU$; $p_{\text{drop}} = 0.25$. The activation function and learning rate choices were especially important. On the other hand, dropout rates between 0.1 and 0.4 were acceptable, while the performance was reduced in the edges of the search space.

B. Performance evaluation

Table V and Fig. 4 summarize the results of the evaluation experiment. Table V displays the command decoding accuracy of the models in the test set for each number of fine-tuning trials and sequences considered to make the prediction. It should be noted that models were trained and tested using observations (i.e., EEG epochs) corresponding to a single stimulus. Therefore, the command decoding process followed a single-stimulus approach: the probability of single epochs was predicted and then averaged per row and column. The command that corresponds to the row and column with higher probability was then selected [**?**]. The last column compares EEG-Inception with the rest of models. Concretely, this column shows the performance improvement of EEG-Inception compared to the other 5 models, calculated as the mean difference in command decoding accuracy. Wilcoxon Signed-Rank Test was applied to evaluate statistical differences between the accuracy of EEG-Inception and the other models. False Discovery Rate (FDR) for multiple comparisons was corrected with Benjamini-Hochberg approach [**?**]. Additionally, Fig. 4 shows the theoretical information transfer rate (ITR) achieved by the models. This metric is widely used to assess the performance of BCI systems [**?**]. The ITR, expressed in bits/min, measures the amount of information conveyed by a BCI system taking into account the speed of selection, the number of available commands and the accuracy of the system [**?**]. As can be seen, EEG-Inception outperformed the comparison models in these analyses, reaching the highest performance for command decoding accuracy and ITR. Moreover, the statistical analysis showed that the improvements in command decoding accuracy are significant (p-value < 0.01), regardless of the number of fine-tuning trials and sequences considered in the analysis. Results broken down by subject can be found in http://dx.doi.org/10.21227/6bdr-4w65.

V. Discussion

In this study, we propose a novel CNN architecture for ERP detection called EEG-Inception. The model was validated in a population of 73 subjects (i.e., 42 healthy, 31 severely disabled) using an ERP-based speller. Moreover, we provide direct comparison with 5 previous approaches: rLDA, xDAWN + RG, CNN-BLSTM, DeepConvNet and EEGNet.

A. Architecture design

The growth of computer vision and natural language processing have led the theoretical and practical development of deep learning in recent years [**?**]. As a consequence, many of the concepts and architectures that boosted this success are specific to these fields and do not generalize for EEG processing, hindering the development of deep-learning approaches in this context [**?**]. In this work, classic and novel concepts and techniques from image and EEG processing were smartly combined to achieve a novel CNN that outperformed previous approaches for ERP detection. The main contributions of the proposed architecture are detailed below.

The first highlight is the inclusion of Inception modules specifically designed for EEG processing. Inception modules were first proposed by Szegedy *et al.* [**?**] for image processing, providing significant improvements in this domain. However, there are still very few studies that explored this architecture for EEG processing, and to our knowledge, none for ERP detection [**?**], [**?**], [**?**]. EEG-Inception integrates 2 Inception modules with 3 branches that process the signal in different temporal scales (i.e., 500ms, 250ms, and 125ms), depending on the kernel size of the temporal filters (C1, C2, C3, C4, C5 and C6 as displayed in Fig. 1). In our experiments, the inclusion of these modules allowed significant improvements in comparison to single-scale approaches. Nevertheless, the performance of the model remained comparable for different kernel sizes, and the optimal values may depend on the specific context. In this study, the scales were chosen to maximize the performance for ERP detection.

The novel architecture design, along with the efficient integration of structures that proved useful in previous studies, such as depthwise convolutions to extract independent spatial filters for each temporal pattern, dropout regularization, batch normalization and average pooling also increased the accuracy of the model [**?**], [**?**], [**?**]. In this regard, the architecture of EEG-Inception was especially designed to avoid overfitting, including an output block that synthesizes the information extracted by Inception modules in very few, high-level features. In fact, only 24 features are fed to the final classification layer (see Table III). This approach maximized the gain of the fine-tuning process, which used very few calibration trials to adapt the model to new subjects, and proved especially important in our experiments.

Finally, the hyperparameter optimization also constitutes an important contribution and could help to design new approaches in the future. As can be seen in Fig. 3a, the activation function proved to be an important choice. We tested Sigmoid, ReLU and ELU functions for their extensive use in deep learning. In fact, ReLU is the preferred choice for computer

TABLE V
COMMAND DECODING ACCURACY IN THE TEST SET

N	Model	No. Sequences 1	5	10	15	Imp.
0	rLDA	14.6 ± 8.6	33.1 ± 20.4	40.3 ± 24.2	43.7 ± 25.1	**14.7**
	xDAWN + RG	17.9 ± 9.2	37.5 ± 19.6	46.2 ± 22.8	49.7 ± 23.4	**9.6**
	CNN-BLSTM	19.6 ± 8.8	42.7 ± 19.8	53.6 ± 22.0	57.2 ± 22.5	**3.7**
	DeepConvNet	20.7 ± 9.8	42.0 ± 20.8	53.2 ± 23.4	56.9 ± 23.7	**4.0**
	EEGNet	19.5 ± 9.5	42.3 ± 21.0	51.9 ± 23.2	56.5 ± 23.3	**4.6**
	EEG-Inception	21.3 ± 9.9	46.1 ± 21.1	57.4 ± 23.7	61.4 ± 23.8	**—**
5	rLDA	18.6 ± 8.6	41.7 ± 18.7	53.4 ± 20.3	58.2 ± 20.0	**16.0**
	xDAWN + RG	20.6 ± 8.9	47.3 ± 19.6	59.0 ± 20.5	63.4 ± 19.9	**10.7**
	CNN-BLSTM	24.3 ± 10.3	52.7 ± 19.8	64.7 ± 20.2	68.5 ± 19.4	**5.6**
	DeepConvNet	23.8 ± 10.3	53.2 ± 20.5	64.3 ± 21.3	68.3 ± 20.4	**5.7**
	EEGNet	23.8 ± 9.8	53.0 ± 20.2	65.3 ± 20.8	69.4 ± 20.1	**5.1**
	EEG-Inception	26.4 ± 10.8	58.7 ± 19.4	70.7 ± 18.8	74.6 ± 17.2	**—**
10	rLDA	22.0 ± 10.0	50.4 ± 19.5	63.5 ± 20.3	68.2 ± 19.2	**11.6**
	xDAWN + RG	24.1 ± 9.8	54.9 ± 19.4	67.2 ± 19.1	71.5 ± 17.8	**7.7**
	CNN-BLSTM	25.9 ± 10.7	56.0 ± 19.9	68.3 ± 19.5	72.1 ± 18.5	**6.7**
	DeepConvNet	26.4 ± 11.0	58.4 ± 20.3	69.6 ± 20.1	73.3 ± 18.8	**5.2**
	EEGNet	26.1 ± 10.2	58.2 ± 19.8	70.6 ± 19.6	74.5 ± 18.3	**4.8**
	EEG-Inception	28.8 ± 11.1	63.6 ± 18.7	75.3 ± 17.6	78.9 ± 15.5	**—**
20	rLDA	25.1 ± 11.0	56.7 ± 19.6	69.9 ± 19.6	74.6 ± 17.8	**9.9**
	xDAWN + RG	26.4 ± 10.4	60.1 ± 18.3	72.3 ± 17.3	76.4 ± 15.6	**7.2**
	CNN-BLSTM	27.5 ± 11.2	60.1 ± 19.2	72.3 ± 18.1	75.7 ± 16.8	**7.2**
	DeepConvNet	28.9 ± 11.6	62.7 ± 19.6	73.7 ± 18.7	77.1 ± 17.0	**5.6**
	EEGNet	28.8 ± 10.7	63.2 ± 18.8	75.6 ± 17.9	79.1 ± 16.2	**4.3**
	EEG-Inception	32.0 ± 12.1	68.4 ± 17.9	79.4 ± 16.0	82.8 ± 13.7	**—**
30	rLDA	26.2 ± 11.2	59.5 ± 19.3	72.7 ± 19.0	77.3 ± 16.9	**9.4**
	xDAWN + RG	27.3 ± 10.6	62.0 ± 17.9	74.2 ± 16.5	78.2 ± 14.9	**7.5**
	CNN-BLSTM	29.0 ± 11.5	62.6 ± 18.9	74.4 ± 17.4	77.7 ± 16.0	**7.2**
	DeepConvNet	29.8 ± 11.8	64.9 ± 19.1	75.8 ± 17.9	79.0 ± 16.3	**5.7**
	EEGNet	30.4 ± 11.3	65.7 ± 18.6	77.8 ± 17.2	81.0 ± 15.5	**4.3**
	EEG-Inception	33.7 ± 12.4	70.6 ± 17.4	81.5 ± 15.3	84.6 ± 13.2	**—**

N: number of fine-tuning trials for each subject. Command decoding accuracy (%) averaged over the test set subjects (31 motor disabled). Column "Imp." shows the accuracy improvement (%) yielded by EEG-Inception compared to the other 5 models, calculated as the mean difference in command decoding accuracy for each N. Statistical differences between EEG-Inception and the other models were assessed with Wilcoxon Signed Rank Test, correcting the False Discovery Rate (FDR) with Benjamini-Hochberg approach. All comparisons were significant (p-value < 0.01), regardless of N and the number of sequences.

vision [**?**]. However, ELU achieved greater performance in our experiments. This finding is in accordance with the work of Schirrmeister *et al.* [**?**] for SMR classification. Reducing the overfitting is also crucial in order to design deep-learning models for ERP detection, even with large datasets. In our experiments, dropout regularization proved to be the most useful technique to reduce this effect [**?**], [**?**]. The dropout rate was automatically optimized, reaching its optimal value in the validation set at 0.25. Finally, the learning rate used for training and fine-tune the model also played an important role to reach an optimal solution. As can be seen in Fig. 3b, a learning rate of 0.001 maximized the command decoding accuracy in the validation set.

B. Results and advantages of the training strategy

Table V shows the command decoding accuracy for each number of fine-tuning trials, model and sequences considered. As can be seen, EEG-Inception always achieved the highest accuracy, followed by EEGNet, DeepConvNet, CNN-BLSTM, xDAWN+RG and rLDA. In fact, the comparison between EEG-Inception and the rest of models showed improvements up to 16.0% for rLDA, 10.7% for xDAWN + RG, 7.2% for CNN-BLSTM, 5.7% for DeepConvNet and 5.1% for EEGNet. Moreover, the statistical test (i.e., Wilkoxon Signed Rank Test, FDR corrected with Benjamini-Hochberg approach) showed that these differences were significant (p-value < 0.01), regardless of the number of fine-tuning trials and sequences. Of note, the standard deviation reached high values for all models due to the high inter-subject variability of the test set, which had subjects with very different pathologies. Nevertheless, it should be noted that EEG-Inception was generally the model with the lowest variability, showing greater robustness to individual differences. Additionally, Fig. 4 shows the theoretical ITR achieved by the models. As before, EEG-Inception reached the highest value, 25.64 bits/min, which is comparable to the ITR achieved by healthy subjects using the same paradigm [**?**]. This is in accordance with previous studies, which demonstrated that deep-learning approaches usually achieve higher performance for ERP detection [**?**], [**?**]. Unsurprisingly, the number of fine-tuning trials had a positive impact in the performance due to the high inter-subject variability of ERPs. A larger number of calibration trials enable classification algorithms to learn subject specific features, thus increasing the command decoding accuracy. In return, it also increases the training time before using the system, which reduces its usability. Therefore, a suitable balance should be found between performance and calibration time. In this regard, deep-learning models (i.e., CNN-BLSTM, DeepConvNet, EEGNet and EEG-Inception) showed a clear advantage in comparison to rLDA and xDAWN + RG models, reaching suitable accuracies with fewer fine-tuning trials. Another aspect to consider is the speed of selection, which

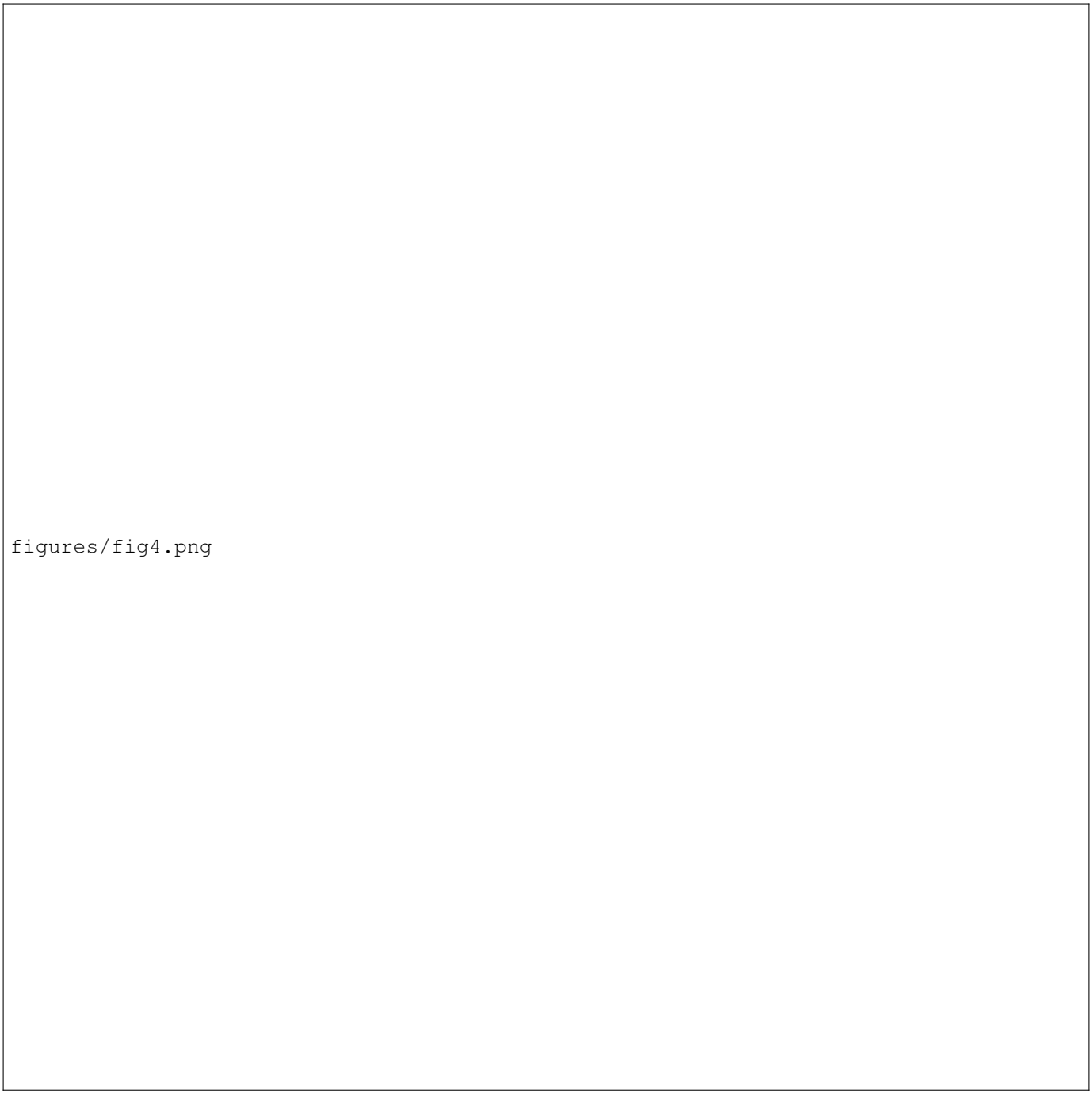

Fig. 4. ITR: information transfer rate; N: number of fine-tuning trials. ITR in bits/min in the test set (31 motor disabled subjects) as a function of N and the number of sequences.

depends on the number of sequences (i.e., stimuli per row and column). A greater number of sequences entails higher accuracy, regardless of the model, but also increases the time of selection. As before, a proper trade-off between precision and speed must be achieved for practical applications. In this regard, EEG-Inception would provide higher accuracies with less selection time.

Fundamental differences in the experimental design, stimulation paradigm and subjects make it difficult to compare the results achieved in this study with previous works in terms of performance [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**]. Especially, our test set comprised 31 severely disabled subjects, which are the end users of ERP-based spellers. On the other hand, to the best of our knowledge, none of the previous works that proposed deep-learning approaches for ERP-based spellers tested their models with target users [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**]. As stated before, severely disabled subjects pose a great challenge for their heterogeneity and special characteristics. Moreover, it has to be taken into account that, for each subject, all sessions were mixed. Therefore, results could be affected

to some extent by inter-session variability. To mitigate this problem, we opted for the random selection approach with multiple repetition, performing 100 repetitions and averaging the results. For these reasons, we implemented several of the most successful previous classification approaches for ERP-based spellers in order to assure a fair comparison. As shown, our proposal significantly outperformed all of them, demonstrating its feasibility in a practical setup with end users.

The designed training strategy also constitutes an important contribution of the study and allows to draw some insightful conclusions that should not be overlooked. Most related studies, probably due to dataset limitations, analyzed the performance of novel BCI classification methods from cross-subject (i.e., training and testing with different subjects) or intra-subject (i.e., training and testing with data from the same subject) points of view [**?**], [**?**], [**?**]. However, we applied a hybrid approach using transfer learning and fine-tuning. For $N = 0$, all models were trained using the training set and tested using the test set, following the cross-subject evaluation approach. In contrast, for $N > 0$, the model was updated with new data from each subject starting from the original weights through the fine-tuning process. This is a fundamental analysis to study the potential of deep learning for BCI, measuring the effectiveness of cross-subject transfer learning and fine-tuning. In accordance to previous studies with classical methods [**?**], our results suggest that deep-learning models can also take advantage from these techniques to reduce calibration time in ERP-based spellers without compromising the performance of the model with end users. In fact, we used a maximum of 30 calibration trials to obtain subject-specific models for the test set. In contrast, among related studies that used the same stimulation paradigm, Cecotti *et al.* [**?**] used 85 calibration trials for each subject, Liu *et al.* [**?**] included 2 databases with 85 and 42 trials, and Santamaría-Vázquez *et al.* [**?**] and Borra *et al.* [**?**] used 140 trials. Furthermore, in this study the models were initialized using signals from healthy subjects, which are easier and less expensive to acquire. While public databases of disabled subjects are scarce, there is a great amount of data available from healthy subjects. Thus, existing databases could be used to reduce the calibration time for end users. As aforementioned, EEG-Inception is designed to make the most of this training approach, providing advantages for the practical use of ERP-based spellers in real applications.

C. Limitations and future work

Despite the positive results achieved in this study, we also acknowledge several limitations that should be addressed in the future. For instance, the validation of EEG-Inception in other EEG classification tasks and public datasets is a promising research line. We only tested the performance of EEG-Inception to detect ERPs elicited by the RCP, whose main component is the P3 wave [**?**]. However, our approach has not been evaluated with ERPs elicited by other stimulation paradigms, such as miniature asymmetrical visual evoked potentials or motion visual evoked potentials [**?**]. These ERPs are characterized by different components, which affects the morphology and spatial distribution of the response. This variability makes the development of general models for ERP detection challenging [**?**]. In this regard, EEG-Inception introduces architectural advantages that could help to overcome this issue. Therefore, we think that our approach has great potential to cope with different paradigms after an appropriate fine-tuning process with enough training examples. Furthermore, we believe that, after a proper optimization of several hyper-parameters, EEG-Inception could be applied in other contexts such as BCIs based on SMR or SSVEP, sleep stage scoring, disease detection, etc. Nevertheless, additional experiments are required to corroborate these hypotheses. On the other hand, online tests would also be interesting to assess the performance of EEG-Inception in real applications. In fact, EEG-Inception could be used together with novel stimulation paradigms and interactive strategies to improve the overall performance [**?**], [**?**]. Additionally, we did not apply any method to explain the features learned by EEG-Inception. In this regard, explainable deep-learning models could help to gain insight into brain processes through EEG and optimize the architectures, being a field of research with great potential. Therefore, future endeavors are needed to address this issue. Finally, adaptive methodologies for traditional machine-learning classifiers yield improvements in accuracy and calibration time in BCI [**?**], [**?**]. However, to the best of our knowledge, adaptive approaches for deep-learning models have not been studied in this domain yet. Thus, they represent a promising research line that should be explored in the future.

VI. Conclusion

In this study, we proposed a novel CNN architecture for ERP classification called EEG-Inception. This model efficiently integrates Inception modules to facilitate the extraction of feature maps at different temporal scales with other structures optimized for practical use in ERP-based spellers. EEG-Inception showed excellent performance in this context, significantly outperforming 5 successful previous approaches such as rLDA, xDAWN + RG, CNN-BLSTM, DeepConvNet and EEGNet in an experiment that involved 73 subjects, including 31 with severe motor disabilities. To the best of our knowledge, this is the largest sample in related studies, assuring the generalization of our results. Additionally, the proposed training strategy reduced the calibration data required by deep-learning approaches to achieve a suitable accuracy with new subjects. In the future, these concepts could be applied to enhance the performance of deep-learning models in other EEG classification tasks.