EEGSym: Overcoming Inter-subject Variability in Motor Imagery Based BCIs with Deep Learning

Sergio Pérez-Velasco, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Diego Marcos-Martínez and Roberto Hornero, *Senior Member, IEEE*

***Abstract*— In this study, we present a new Deep Learning (DL) architecture for Motor Imagery (MI) based Brain Computer Interfaces (BCIs) called *EEGSym*. Our implementation aims to improve previous state-of-the-art performances on MI classification by overcoming inter-subject variability and reducing BCI inefficiency, which has been estimated to affect 10-50% of the population. This convolutional neural network includes the use of inception modules, residual connections and a design that introduces the symmetry of the brain through the mid-sagittal plane into the network architecture. It is complemented with a data augmentation technique that improves the generalization of the model and with the use of transfer learning across different datasets. We compare *EEGSym*'s performance on inter-subject MI classification with ShallowConvNet, DeepConvNet, EEGNet and EEG-Inception. This comparison is performed on 5 publicly available datasets that include left or right hand motor imagery of 280 subjects. This population is the largest that has been evaluated in similar studies to date. *EEGSym* significantly outperforms the baseline models reaching accuracies of 88.6±9.0 on Physionet, 83.3±9.3 on OpenBMI, 85.1±9.5 on Kaya2018, 87.4±8.0 on Meng2019 and 90.2±6.5 on Stieger2021. At the same time, it allows 95.7% of the tested population (268 out of 280 users) to reach BCI control ($\geq$70% accuracy). Furthermore, these results are achieved using only 16 electrodes of the more than 60 available on some datasets. Our implementation of *EEGSym*, which includes new advances for EEG processing with DL, outperforms previous state-of-the-art approaches on inter-subject MI classification.**

***Index Terms*— Brain Computer Interface (BCI), Deep Learning (DL), Motor Imagery, Transfer Learning, Inter-subject**

This research has been developed under the grants PID2020-115468RB-I00 and RTC2019-007350-1 funded by 'Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación' and ERDF 'A way of making Europe'; under the R+D+i project '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') funded by 'European Commission' and ERDF; and by 'CIBER de Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III'. S. Pérez-Velasco, E. Santamaría- Vázquez and D. Marcos-Martínez were in receipt of a PIF grant by the 'Consejería de Educación de la Junta de Castilla y León'. 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, E. Santamaría-Vázquez, V. Martínez-Cagigal and R.Hornero are with CIBER-BBN (ISCIII), Spain (e-mail: sergio.perez@gib.tel.uva.es; eduardo.santamaria@gib.tel.uva.es; victor.martinez@gib.tel.uva.es; diego.marcos@gib.tel.uva.es; robhor@tel.uva.es)

I. Introduction

Electrical brain activity can be registered through electroencephalography (EEG), which consists of non-invasive recordings from electrodes placed on the user's scalp. EEG is characterized by its relative low cost, ease of use, high temporal resolution and portability, but also for the drawbacks of a poor spatial resolution and low signal-to-noise-ratio (SNR) [1]. Non-invasive brain-computer interface (BCI) applications make use of the EEG to enable an alternative path for the brain to communicate with the environment [2], [3]. These applications range from moving a mouse cursor through a screen [4] or command selection, [5], [6] to commanding prosthetic limbs, which are ultimately developed to assist people with severe motor disabilities [7].

In order to decode the user's intentions from the EEG, BCIs usually rely on control signals triggered through strategies known as BCI paradigms. In this work, we will focus on decoding the user's intention through their Motor Imagery (MI). For MI, the most extended protocol is to use left or right hand movement imagination. Each instance of MI is considered a trial, and the type of imagination performed can be decoded through the sensorimotor rhythms (SMR). SMR are oscillations in the electric field detected in the sensorimotor cortex of the brain. These areas are related with the preparation, control and production of voluntary movements including imaginary ones [8]. Additionally, there are other control signals related with MI like Movement Related Cortical Potentials (MRCP) [8] and Lateralized Readiness Potentials (LRP) [9]. MI is of great interest due to its great potential for rehabilitation. The use of a MI-based BCI on twelve participants has been reported to induce plasticity at the cortical level [10]. A correlation between the classification accuracy of the MI-based BCI rehabilitation and the improvement of the upper limb function was found on a population of 74 stroke patients with severe upper limb paralysis [11]. Other works studied the effect of different ways of presenting the feedback, like sensory threshold neuromuscular electrical stimulation [12] or through virtual reality [13]. The evidence found in these works has led to MI-based BCI rehabilitation to be exploited by commercial applications [14]–[16].

Nonetheless, one major drawback of BCIs is the decoding accuracy of EEG. Classical approaches of machine learning (ML) for BCIs, like common spatial patterns (CSP) with some improvements [17], [18], filter bank common spatial patterns

(FBCSP) [19], and Riemannian geometry [20] in combination with linear discriminant analysis (LDA) or support vector machines (SVM), need a tiresome calibration run from each user. This calibration run would not be a clear disadvantage if not for the intersession and inter-subject variability [21]. On one hand, the inter-subject variability does not allow a model trained in one subject to be used on another one with acceptable performance. And on the other hand, the intersession variability does not allow trials from previous sessions of the same subject to train a good performing model for the next session. Due to the combination of both, classic ML for BCIs often require a calibration run for each session, and in turn obtain not very good performances overall. However, Deep Learning (DL) models outperformed classical ML approaches, and at the same time reduced the impact of inter-subject and intersession variability due to the ability that DL has for transfer learning [22], [23].

Schirrmeister *et al.* [22] and Lawhern *et al.* [23] proved the ability of Convolutional Neural Networks (CNN) architectures for EEG decodification across different paradigms. Dose *et. al* [24] and Zhang *et al.* [25] implemented an adaptation of the CNN proposed by Schirrmeister *et al.* [22] to Physionet [26] and OpenBMI [27] datasets, respectively. These two works tried to reach higher accuracies in MI-based BCIs by providing an increase of training trials compared to the dataset used in the original work for MI [22]. There have been works that have tried to improve these performances with new DL techniques from the computer vision field. Santamaría-Vázquez *et al.* [5] already proved the improvement that inception modules [28] have on CNNs accuracy for EEG decoding in an event related potentials (ERP) based speller. Fan *et al.* [29] tackled inter-subject variability in MI with an improved CNN that included residual connections [30] and an attention mechanism [31]. Kostas *et al.* [32] adapted the DenseNet DL Network [33] from the field of computer vision to EEG decoding of MI, and Kwon *et al.* [34] applied feature engineering to the input of their proposed CNN by creating a spectro-spatial feature representation from the EEG.

Despite the advances of DL in the field of BCIs, there are several limitations that have not been addressed. Firstly, in spite of the success of Lawhern *et al.* [23] and Schirrmeister *et al.* [22] on EEG decodification at the time, there has been a surge of improved DL techniques in the field of computer vision that had yet to be adapted for EEG decoding networks. Secondly, previous CNNs extract spatial features with a single convolution along the spatial dimension in the first layers of the network [5], [22]–[24], [32], which limits the spatial relationships discovered to this first convolution. The extraction of spatial features could be enhanced by introducing the known structure of the brain into the CNN architecture or by using residual connections [30] to maintain the structure of the EEG data. Thirdly, the studies in the area of MI decoding did not fully take advantage of the power that DL has for transfer learning. They validated the results on datasets with a large amount of subjects and trials but did not try to extend its procedures on more than one dataset. At the same time, they lost the opportunity to improve their models' performance with the increased training data that including other datasets offer. Fourthly, a reduced number of electrodes facilitates real world applications by reducing the set up duration, and by decreasing the cost of the EEG recording system needed. For reference, placing an EEG cap of 64 electrodes can take up to 1 hour [35], but only Dose *et al.* [24] and Fan *et al.* [29] studied the effect that reducing the number of electrodes had on their DL model's performance for inter-subject MI classification. Finally, despite using all available electrodes and having calibration runs, current approaches still suffer from BCI inefficiency (also known as BCI illiteracy). This is the inability of BCI applications to extract discernible features from an user, which is estimated to affect 10-50% of potential users [36] in MI-based BCIs. Previous studies consider that a user attains BCI control if he reaches accuracies higher than 70% in MI binary classification [27], [37].

To overcome the above limitations, this study aims to design a novel CNN called *EEGSym* outperforming previous state-of-the-art DL architectures. To this end, we compare our model on 280 subjects from 5 different datasets against 4 state-of-the-art CNN based models. To the best of our knowledge, this population is the largest used in compared studies to date. Our approach takes advantage of transfer learning through several datasets to overcome inter-subject variability with only 8 or 16 electrodes. The novelties that this study introduces are summarized in the following points:

- A data augmentation (DA) technique that includes patch perturbation, hemisphere perturbation, and a random shift of the onset.
- An improved extraction of features through residual connections that tries to keep the spatio-temporal structure of the signal through several layers of the network.
- A siamese-network approach to exploit the symmetry of the brain along the mid-sagittal plane.

An open source implementation of the architecture and DA can be found in https://github.com/Serpeve/EEGSym

II. Methods

A. Datasets

Five datasets were used to evaluate the baseline models and *EEGSym*: Physionet [26], OpenBMI [27], Kaya2018 [38], Meng2019 [37], and Stieger2021 [39]. We selected these datasets due to the amount of subjects they include (i.e., 109, 54, 13, 42, and 62, respectively), the amount of trials, and for their shared type of movement imagined. The imagination consisted of opening/closing either the left or right hand. The shared imagination paradigm should be key for the transfer learning between datasets and subjects. All datasets except Physionet include sessions where feedback of their EEG was presented to the participants. Furthermore, Kaya2018, Meng2019 and Stieger2021 only consist of trials from feedback sessions [37]–[39]. MI duration of Stieger2021's trials vary due to the subjects reaching the target presented [39]. The summarized characteristics of each dataset are detailed on table I.

The experimental protocol share the same structure for every dataset. Starts with a resting period from 1 to 4 seconds where a fixation cue is presented to prepare the subjects for the

TABLE I
DETAILS OF THE DATASETS

Dataset	Subjects	#	MI duration (s)	Trials/ session	Sessions	Sampling Frequency (Hz)
Physionet [26]	109	64	3	45	1	128/160
OpenBMI [27]	54	62	4	100	4	1000
Kaya2018 [38]	13	38	1	900	5	200
Meng2019 [37]	42	64	6	250	3	1000/1024
Stieger2021 [39]	62	62	2-8	450	7-11	1000

Subjects: number of healthy subjects. #: number of electrodes. MI: motor imagery.

imagination period. It is followed by a MI period of different duration where a cue is presented to perform either left or right hand MI. This varying MI duration implies that when extracting the same time window length, some trials will include only part of the imagination period while others will also include the following resting period or even the start of the following trial on Kaya2018 [38]. Ends with a final resting period of 2-6 seconds of relaxation previous to the next trial.

B. Preprocessing

The raw signal of each dataset was processed as follows:

1) Extraction of channels 'F3', 'C3', 'P3', 'Cz', 'Pz', 'F4', 'C4', and 'P4' from the available channels in each dataset for the 8 electrodes configuration. The 16 electrodes configuration includes also the 'F7', 'T7', 'P7', 'O1', 'F8', 'T8', 'P8', and 'O2' channels from the 10/20 system. The amount of electrodes in these two configurations are widely used in relatively low-cost EEG-caps, and provide a reduced set-up duration.
2) Application of a fourth-order infinite impulse response (IIR) notch filter to eliminate power line signal at 50/60 Hz of each dataset that did not have it removed by hardware [26], [27].
3) Application of common average reference (CAR) spatial filtering to these 8/16 channels.
4) Resampling to 128 Hz to homogenize the datasets.
5) Extraction of the trials with a time window length of 3 seconds after the onset. This 3 second time window was the largest possible to extract over all datasets without having to discard trials without enough samples or having to artificially pad the signal.
6) Application to each trial of a channel-wise z-score standardization. Each channel signal in a trial ends with zero mean and unit variance. This operation removes the continuous component of the signal and accommodates the data to be fed to a DL neural network.

C. Data Augmentation

DA is applied to generate new training examples from existing data. This technique reduces over-fitting and enables the training of bigger models that offer better generalization on new data [40]. When applying DA, a uniform random selection between the following four options was applied for each trial differently in each pass through the whole training data: patch perturbation, hemisphere perturbation, random shift or no augmentation. Therefore, the training set would be unique for each training epoch and it would be very unlikely for a model to be trained on the same composition of trials twice.

The DA in this work was composed of 3 different ideas:

1) Patch perturbation. We adapted a DA technique from computer vision called random erasing [41] because its principles could be extrapolated to EEG data. First, we select a time window duration to be modified. Similar to random erasing, the aim of patch perturbation is to make the model robust to the presence of noise on the EEG data. Like dropout, randomly perturbing different time sections or channels of the signal will force the model to learn relations from non perturbed sections of EEG to make up for the perturbed data. At the same time, it will make the model less reliant on specific time segments or channels and generalize better. The duration is selected from an uniform distribution between 0.6 to all 3 seconds of each trial to be distorted. Secondly, a position where to place this time window is randomly selected. Thirdly, a number of channels in which this time window will be distorted is randomly selected. Always at least one channel will be left unmodified to preserve the information of that time window. Finally, the distortion consists of either changing the affected patch by 0s (erased) or by adding noise. The added noise follows a Gaussian distribution with 0 mean and with a standard deviation that varies uniformly from 0.01 up to 2 times the standard deviation of the signal.
2) Hemisphere perturbation. We hypothesize that the difference between the control signals (i.e., SMR, LRP, MCP) of left/right hand MI can be decoded from EEG changes in one hemisphere. With this in mind, the electrodes corresponding either to the left or right hemisphere are perturbed. This perturbation consist of either altering its positions in a random order or replacing all hemisphere data by Gaussian noise with 0 mean and 1 standard deviation. This technique aims for the model to learn a clear and discernible pattern of MI in either hemisphere. This perturbation also has a regularization effect, but in this case it is restricted to the spatial dimension of the signal.
3) Random shift. In MI, we know the exact time when the onset cue is presented to the users, but not the reaction time that they have for each trial. The reaction time varies its distribution for each user. We also want to consider distracted or tired subjects which will exhibit a slower reaction time in some trials. To account for this variability, the data is also augmented by shifting forward the trials onset as much as half a second. This value was set to consider the slowest tail of the two-choice reaction time distribution in humans [42]. The exact amount of time is extracted from an uniform distribution from 1 to 64 samples (corresponding to half a second with a sampling frequency of 128 Hz).

D. EEGSym

EEGSym includes previous techniques that have been proven to work for EEG decodification. One of them is the

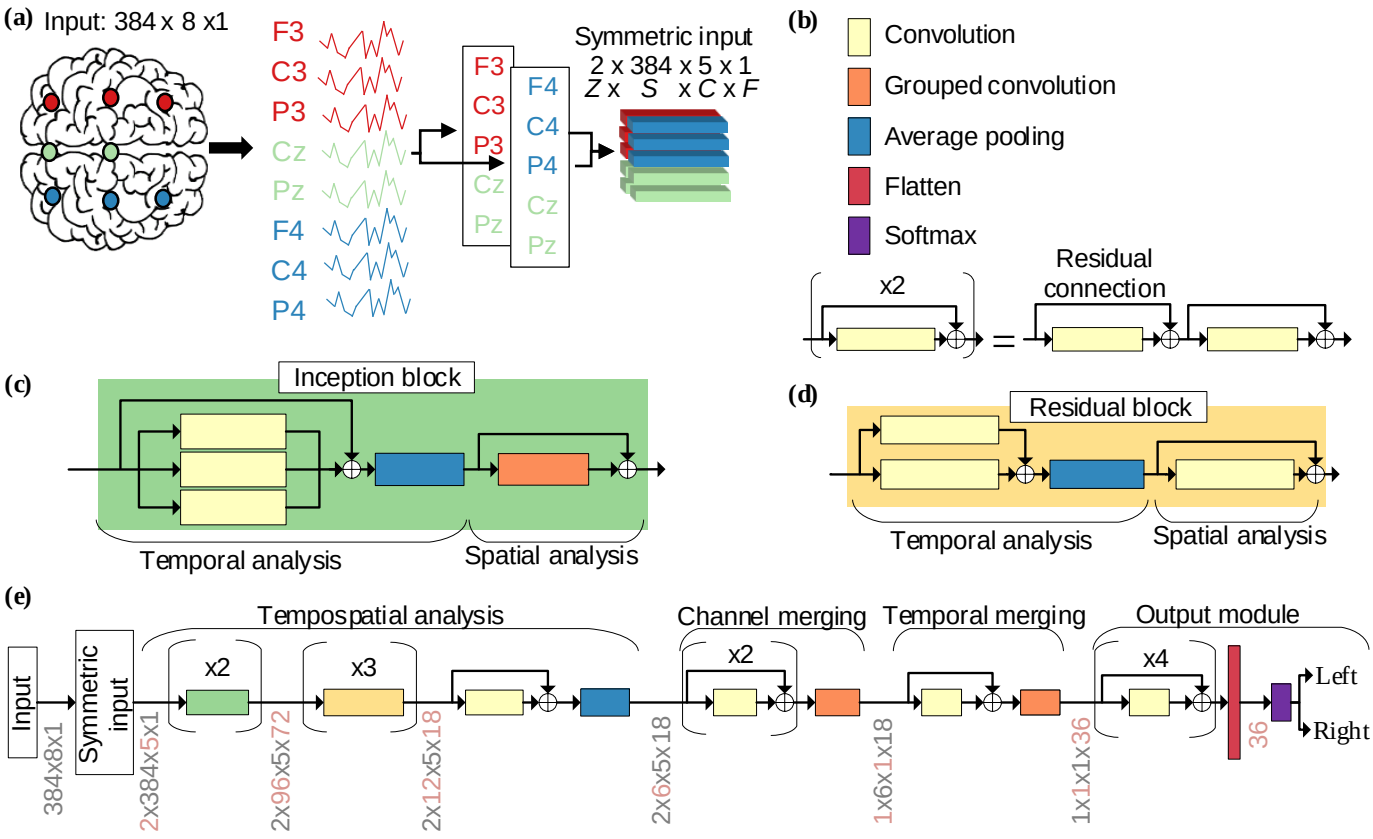

Fig. 1. Overview of *EEGSym* architecture. (a) Schematic of the division of input electrodes for an 8 electrode configuration $\boldsymbol{Z}$: hemispheres (i.e., 2), $\boldsymbol{S}$: samples (i.e., 384), $\boldsymbol{C}$: electrodes per hemisphere (i.e., 5), $\boldsymbol{F}$: number of filters. (b) Legend of the architecture overview. (c) Inception block. (d) Residual block. (e) *EEGSym* architecture. All convolution and grouped convolution operations are followed by batch normalization, 'elu' activation and dropout regularization in this order. The output sizes of each operation are indicated in gray, whereas the dimension that is affected after each stage is indicated in red. Detailed tables of 8 and 16 electrode configurations that include the details (i.e. kernel sizes, number of filters, etc.) of each operation are present in the supplementary material, and in the open implementation that can be found in https://github.com/Serpeve/EEGSym.

use of inception modules [28] in the first operations of the architecture as in EEG-Inception [5]. Another one is the use of grouped convolutions [43] to emulate the success that EEGNet [23] and EEG-Inception [5] had applying depthwise convolutions. Depthwise convolutions are a particular case of grouped convolutions when the number of groups is the same as the number of filters. Every convolution operation is followed by batch normalization, 'elu' activation and dropout regularization. The dropout rate (dr), number of filters in inception modules (N) and learning rate (lr), were determined through grid search on the validation set. The search spaces for these hypeparameters were: $dr = [0.2 : 0.1 : 0.5]$; $N = [8 : 8 : 32]$; and $lr = [0.01, 0.001, 0.0001]$. The values selected were 0.4, 24, and 0.001, respectively.

An overview of *EEGSym*'s architecture is presented in Fig. 1. Detailed tables of 8 and 16 electrode configurations that include the details (i.e. kernel sizes, number of filters, etc.) of each operation are available in the open implementation, and in the supplementary material. The architecture of *EEGSym* can be separated in 5 stages:

1) Symmetric division. Symmetric division. It creates the virtual division represented in Fig. 1.a that is performed inside the model. Hence, no redundant information is fed into the DL arquitecture. The symmetric division of the electrodes also helps to reduce the number of parameters in the spatial filters present in the following tempospatial analysis stage.
2) Tempospatial analysis. It captures the most detailed temporal relationships in the architecture. It is composed of two instances of inception blocks and three of residual blocks. The number and kernel sizes of the inception modules in the first inception block (i.e., 3 modules of size 64, 32, and 16) was selected to replicate the ones chosen in EEG-Inception [5]. These sizes correspond to temporal windows of duration 500 ms, 250 ms and 125 ms. The result of the signal processed by each convolution in the inception module is concatenated and added to the input through residual connections [30]. Afterwards, an average pooling layer reduces dimensionality in the temporal (i.e., S) dimension to prevent overfitting and reduce computation time. Finally the spatial extraction is designed with a grouped convolution that spans all hemisphere's channels (i.e., C), reducing its channels dimension to 1, and then adds the result to every channel with residual connections. These grouped convolutions are designed with the same number of groups and input filters to reproduce the function of depthwise convolutions. The residual block has as well a

temporal analysis followed by dimensionality reduction through an average pooling operation and a spatial analysis performed this time with a convolution instead of a grouped convolution, which will mix the information of all previous temporal filters extracted. After leaving the last residual block, there is a convolution with residual connections to capture temporal relations after the last spatial operation followed by an average pooling operation.

3) Channel merging. In this stage, the signal's spatial dimensionality is reduced to 1 (i.e., Z and C). It is composed of two convolutions with residual connections in the spatial dimension to capture the last distinguishable spatial features extracted. The merging of the channels dimension is performed by a grouped convolution. All convolutions and gropued convolutions in this stage are performed on both hemispheres and all channels at the same time (i.e., kernel size of $2\times1\times5$).
4) Temporal merging. After this stage, the temporal dimensionality is reduced to 1 (i.e., S). It has a convolution with residual connections followed by a grouped convolution. Both operations has a kernel size the same as the temporal dimension that enters this stage.
5) Output module. After the temporal merging, we only remain with a number of features that depends on the number of filters per branch in the inception modules (i.e., for 24 filters per branch 36 features enter this stage). This stage performs 4 convolutions with residual connections, flatten the features, and perform a softmax classification over the two classes of MI.

Furthermore, *EEGSym* includes 2 novel ideas that take advantage of the spatial characteristics of the brain and the EEG:

1) Residual connections. Our network includes an extraction of spatial features, spatial correlations between the signal of different electrodes, with residual connections that are present at every instance of the tempospatial analysis until the channel merging stage. Residual connections are a solution that allows training deeper models without reducing performance [30]. It creates shortcuts for the information leaving the previous layer to skip the transformation of the current layer. The inclusion of residual connections also allows for some layers to be skipped by pushing the weight values of a residual layer to 0. Meanwhile, the information will travel to the next layer through the shortcut. This way, it is easier for the input information to travel unmodified through the whole architecture. The reasoning behind this design is that the spatial correlations of the signal would be different in further stages of the temporal processing of the signal.
2) Symmetry. The symmetry of the brain through the mid-sagittal plane is implicitly introduced in *EEGSym* architecture. This idea takes inspiration from a paper about gaze recognition in which the authors take into account the symmetry of both eyes in the first layers of the network [44]. In a similar fashion, *EEGSym* first extracts common spatial characteristics from both hemispheres in the tempospatial analysis stage. In the channel merging stage, it extracts complex relationships between channels of both hemispheres. An scheme of the division of the input for an 8 electrode configuration can be found in Fig. 1.a.

The contribution of the two novelties introduced in EEGSym architecture is evaluated with an ablation study presented in III-B.

E. Baseline models

For comparison purposes, we used ShallowConvNet and DeepConvNet [22], EEGNet [23], and EEG-Inception [5] applying the hyperparameters described in their original publications.

1) ShallowConvNet/DeepConvNet: The work of Schirrmeister *et al.* [22] focused on showing how to design and train CNNs to decode task-related information from the raw EEG without handcrafted features [22]. They proposed two CNN architectures, ShallowConvNet and DeepConvNet, which were compared with FBCSP showing similar and even better performance in some cases. Here, we use the reproduction of the models made by Lawhern *et al.* [23] on TensorFlow. The details of its implementation can be found in [22].

2) EEGNet: Lawhern *et al.* [23] introduced EEGNet, a compact CNN for EEG-based BCIs, and compared its performance for intra-subject and inter-subject classification. They showed that it generalized across different BCI paradigms, and achieving comparably higher performances than other state-of-the-art algorithms when limited training data is available. We used the implementation released by the author whose details can be found in [23].

3) EEG-Inception: Santamaría-Vázquez *et al.* [5] were the firsts to introduce a CNN model for EEG decodification that integrated inception modules. This network improved the performance of EEGNet and DeepConvNet, as well as other traditional approaches in ERP detection. The model in TensorFlow and their specific architecture details can be found in [5].

F. Cross-validation analysis

All models were trained on a NVIDIA 3080Ti GPU, with CUDA 11.2 and cuDNN 8.1.0, in Tensorflow 2.5. An scheme of the cross-validation analysis is presented in Fig. 2. The trials are splitted into pre-training, fine-tuning and test:

1) We select a target dataset for which we are going to obtain the inter-subject MI prediction accuracy, and use every other dataset as pre-training (Fig. 2.b). From the pre-training operation we obtain an initialization of the weights' values that will be the same for every following fine-tuning operation on the target dataset. From each subject of the pre-training datasets, 10 trials of each class are selected to be part of the validation split, and the rest will be part of the training split.
2) Every subjects' trials present in the target dataset except for the one we will user for testing (Fig. 2.c) will be part

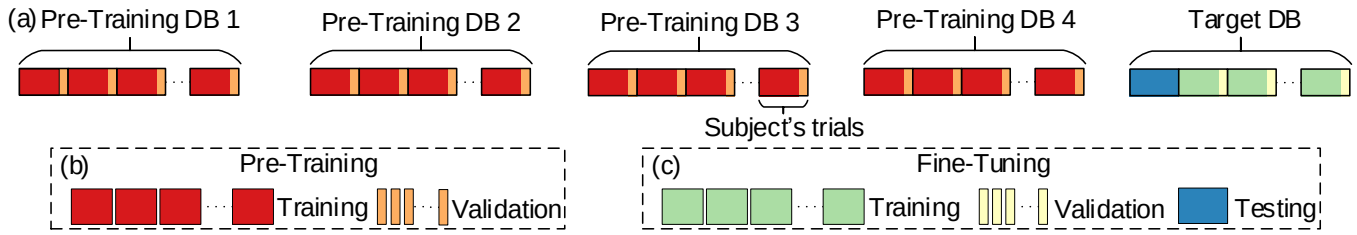

Fig. 2. Cross-validation analysis performed over each dataset as Target DB. Pre-Training DB: datasets used for pre-training the model. Target DB: dataset in which leave one subject out (LOSO) is performed, testing on one subject and fine-tuning the model on the remaining subjects. (a) Scheme of the cross-validation analysis. (b) Pre-training dataset. (c) Fine-tuning dataset and testing subject.

of the fine-tuning. Each fine-tuning subject's trials are splitted into training and validation with a 9 to 1 ratio, respectively.

3) After the fine-tuning operation, we use the trials of each independent subject as test following a leave one subject out (LOSO) scheme (Fig. 2.c). This means that, for each dataset, the fine-tuning and testing operation is performed as many times as independent subjects are in the target dataset to obtain the inter-subject MI prediction accuracy.

For each CNN, we performed the preprocessing as described in subsection II-B and implemented the following DL techniques:

- Early stopping on pre-training and fine-tuning that halts the training when validation loss does not improve for 25 consecutive iterations.
- Pre-training of the models on all datasets excluding the target dataset. The DA described in subsection II-C was only applied in this stage of the process. The learning rate used is the same for all models (i.e., 1e-2). This value is the one present in the open implementation of Lawhern *et al.* [23] for ShallowConvNet, DeepConvNet and EEGNet, and also in the open implementation of Santamaría-Vázquez *et al.* [5] for EEG-Inception.
- Fine-tuning on the target dataset without DA. The full architecture is freezed (its parameters will not be updated during training) apart from the last softmax layer. It is trained with a very low learning rate (i.e., 1e-4) until the early stopping is triggered. Finally, the full architecture is allowed to update all of its parameters with this low learning rate, until the early stopping activates. The first fine-tuning process aims to maintain the knowledge extracted in the pre-training by only adjusting the importance of the features in the softmax classification layer. On the other hand, the second fine-tuning process will further adapt the feature extraction process when the target dataset is very diffferent to the ones present in the pretraining. This procedure is adapted from the indications for fine-tuning a model present in [45].

III. Results

A. Comparison with baseline models

Following the preprocessing and cross-validation analysis described before, we tested the 8 and 16 electrode configurations with the new *EEGSym* and the baseline models. The mean accuracy obtained between all subjects with its standard deviation (σ), and the number of users that achieve BCI control (users that reach 70% accuracy) for each dataset evaluated are presented in Table II.

As can be seen in Table II, *EEGSym* always obtains significantly (p-value < 0.05) higher mean accuracies than the baseline models according to Wilcoxon signed rank test [46], with the false discovery rate (FDR) corrected with Benjamini-Hochberg approach [47]. This occurs for both electrode configurations and all datasets.

EEGSym enabled 268 users out of 280 tested users to achieve BCI control. EEG-Inception follows with 264 users, next is DeepConvNet with 260, ShallowConvNet with 258 and last is EEGNet with 252. Regardless of the architecture, it is worth noting that with our pre-training pipeline every architecture achieves $\geq$90% users with BCI control with only 16 electrodes in a calibrationless application.

B. Ablation study

An ablation study to give insight into the usefulness of the strengths of EEGSym is presented below. On the one hand, we analyzed the effect of introducing residual connections to extract spatial features at different stages of the processed information inside the DL architecture. On the other hand, the introduction of brain's symmetry inside the architecture. Both contributions have been evaluated separately for 8 and 16 electrode configurations over the Physionet [26] dataset. This dataset was selected for this comparison for being the one with the largest number of subjects. The results are summarized in Table III.

As can be observed in the 16 electrode configuration, applying each one of the novelties achieves significantly (p-value < 0.05) greater performances than the base model without symmetry or residual connections, according to Wilcoxon signed rank test [46], with the false discovery rate (FDR) corrected with Benjamini-Hochberg approach [47]. Although performances also increased for the 8 electrode configuration when applying both contributions separately, only the symmetric approach yielded a significant improvement. Nevertheless, the result of jointly using both approaches gives the best performances in both electrode configurations.

Additionally, the evolution of the training and validation losses during the pre-training on the target dataset Physionet [26], and during one instance of fine-tuning of *EEGSym* can be observed in Fig. 3. These results are for the 8 electrode configuration.

TABLE II
COMPARISON OF ACCURACIES ON TARGET DATASETS FOR 8 AND 16 ELECTRODE CONFIGURATIONS

#	Architecture	Physionet [26] $\mu\pm\sigma$	Physionet [26] BCI Control		OpenBMI [27] $\mu\pm\sigma$	OpenBMI [27] BCI Control		Kaya2018 [38] $\mu\pm\sigma$	Kaya2018 [38] BCI Control		Meng2019 [37] $\mu\pm\sigma$	Meng2019 [37] BCI Control		Stieger2021 [39] $\mu\pm\sigma$	Stieger2021 [39] BCI Control	
8 electrodes	ShallowConvNet	82.0±10.9	92/109	*	79.1±9.9	45/54	*	82.3±9.4	11/13	*	82.8±9.3	38/42	*	85.4±7.4	**60/62**	*
	DeepConvNet	82.8±10.7	95/109	*	80.5±9.3	**47/54**	*	82.4±9.2	**12/13**	*	83.6±9.1	39/42	*	86.2±7.2	**60/62**	*
	EEGNet	81.6±11.2	92/109	*	78.6±9.5	46/54	*	80.9±8.6	11/13	*	82.2±9.6	36/42	*	84.5±7.8	59/62	*
	EEG-Inception	82.7±10.8	92/109	*	80.3±9.4	**47/54**	*	81.7±9.1	**12/13**	*	84.4±8.4	**42/42**	*	87.3±7.0	**60/62**	*
	EEGSym	**84.5±9.7**	**99/109**		**82.0±9.6**	46/54		**84.7±9.1**	**12/13**		**85.2±8.3**	41/42		**88.4±6.5**	**60/62**	
16 electrodes	ShallowConvNet	86.2±9.6	100/109	*	80.0±9.7	46/54	*	83.1±9.6	11/13	*	85.5±8.6	40/42	*	87.5±7.3	59/62	*
	DeepConvNet	85.9±10.6	101/109	*	80.9±9.7	46/54	*	82.9±9.7	**12/13**	*	85.9±8.0	**41/42**	*	88.2±7.3	60/62	*
	EEGNet	83.1±10.7	97/109	*	79.6±9.8	45/54	*	82.1±9.1	11/13	*	82.4±9.4	39/42	*	85.7±8.2	58/62	*
	EEG-Inception	87.5±9.3	104/109	*	81.6±9.4	**48/54**	*	82.6±9.9	11/13	*	86.3±8.1	**41/42**	*	89.4±6.8	60/62	*
	EEGSym	**88.6±9.0**	**108/109**		**83.3±9.3**	46/54		**85.1±9.5**	**12/13**		**87.4±8.0**	**41/42**		**90.2±6.5**	**61/62**	

#: number of electrodes used. $\mu\pm\sigma$: mean accuracy and standard deviation obtained across all subjects using leave one subject out (LOSO). BCI Control: users that reach brain computer interface (BCI) control ($\geq$70% accuracy). The best results for each dataset and electrode configuration are marked in **bold**. Statistical differences between the mean accuracies of *EEGSym* and the other models were assessed with Wilcoxon signed rank test, correcting the false discovery rate (FDR) with Benjamini-Hochberg approach. Obtaining significant differences is marked with * (p-value < 0.05).

TABLE III
CONTRIBUTION OF EACH NOVELTY ON PHYSIONET

Res	Sym	8 electrodes $\mu\pm\sigma$	8 electrodes BCI Control		16 electrodes $\mu\pm\sigma$	16 electrodes BCI Control	
		82.9±10.4	90/109		87.2±9.4	104/109	
X		83.2±10.3	94/109		87.8±9.1	105/109	*
	X	84.1±10.1	96/109	*	87.7±9.8	103/109	*
X	X	84.5±9.7	99/109	*	88.6±9.0	108/109	*

Res: If there are residual connections implemented that allows the extraction of spatial features through the whole architecture. Sym: If the electrodes are introduced as described in II-D. $\mu\pm\sigma$: Mean accuracy and standard deviation obtained across all subjects using LOSO. BCI Control: Users that reach brain computer interface (BCI) control ($\geq$70% accuracy). Statistical differences, between current model and the baseline, were assessed with Wilcoxon signed rank test, correcting the false discovery rate (FDR) with Benjamini-Hochberg approach..Obtaining significant differences is marked with * (p-value < 0.05).

IV. DISCUSSION

In this study, we propose a novel CNN architecture called *EEGSym*. It takes advantage of a brain-inspired configuration, a new extraction of spatial features from the EEG based on residual connections across all CNN stages, and transfer learning across subjects. This model was also complemented by DA techniques called patch perturbation, hemisphere perturbation and random shift. It was validated with 5 datasets including a total of 280 subjects, the largest subject evaluation of related studies. A direct comparison with 4 baseline models ShallowConvNet and DeepConvNet [22], EEGNet [23] and EEG-Inception [5] was presented on those datasets.

A. Advantages of EEGSym

EEGSym allowed on 268 out of 280 subjects to achieve BCI control ($\geq$70% accuracy) in a completely inter-subject pipeline, without calibration on test subjects. In other words, 95.7% users reached BCI control in an inter-subject classification, suggesting that transfer learning has the potential to solve

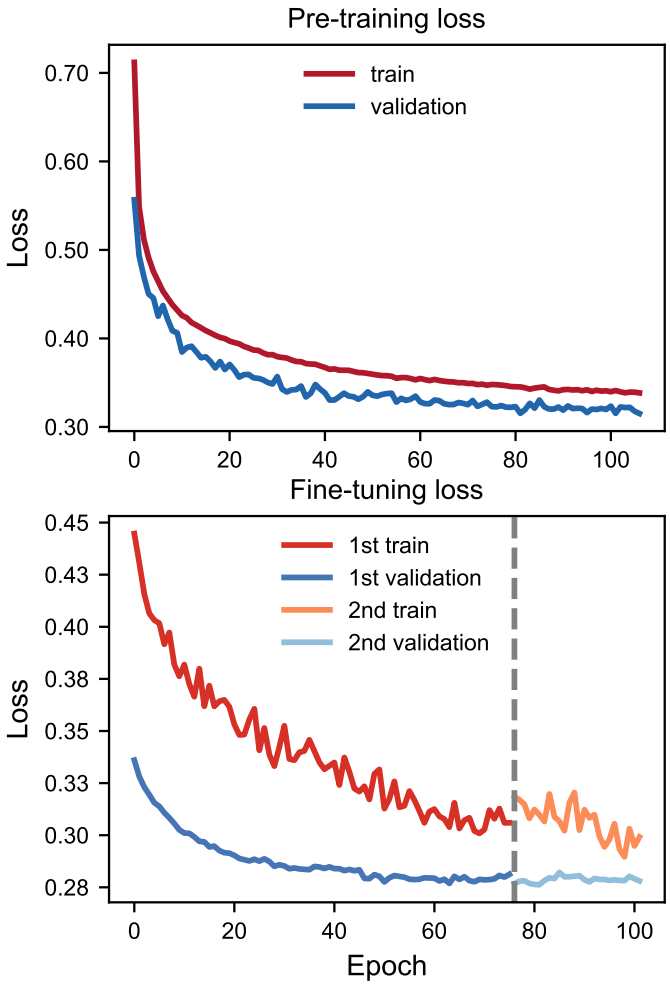

Fig. 3. Loss and validation loss of pre-training on target dataset Physionet [26] and fine-tuning on all dataset subjects except for subject 2 for an 8 electrode configuration. Dotted line in fine-tuning marks the early stopping of the first stage of fine-tuning.

BCI inefficiency. BCI inefficiency was previously estimated to affect 10-50% of potential BCI users [36]. This achievement is even more remarkable since BCI inefficiency seems to affect less than 5% of the population in inter-subject calssification, which is a more challenging problem than the usual intra-subject classification with calibration runs from the end user.

As shown in Table II, we reached accuracies of 88.6±9.0 on Physionet [26], 83.3±9.3 on OpenBMI [27], 85.1±9.5 on Kaya2018 [38], 87.4±8.0 on Meng2019 [37], and 90.2±6.5 on Stieger2021 [39]. A comparison with the mean accuracies of the baseline models was performed with Wilcoxon signed rank test, correcting the FDR with Benjamini-Hochberg approach. *EEGSym* significantly (p-value < 0.05) outperformed ShallowConvNet and DeepConvNet [22], EEGNet [23] and EEG-Inception [5] in this binary MI classification.

Furthermore, DL networks have a clear advantage in other areas like computer vision and natural language processing when large amounts of data are available. In this work, we further exploit the transfer learning capabilities of DL in the field of BCIs, by using all datasets publicly available that share the same imagination paradigm. Our results suggest that the combination of the pipeline described in subsection II-B with the new architecture, enables a plug-and-play application of MI-based BCIs. It does not need calibration trials from the end user using only 8 or 16 electrodes to reach these new state-of-the-art accuracies. Of note, motivation through rehabilitation is a key aspect for the treatment's success [48]. The reduced set-up duration and calibrationless system achievable with *EEGSym* could be key in promoting user's motivation when using MI-based BCIs for rehabilitation.

The contribution of EEGSym's designing novelties present in the implementation of this new architecture are evaluated in the ablation study. It showed that jointly applying them offered significantly better performances for both electrode configurations. However, each one of them separately showed improvements that were not always significant. The residual connections offered an improved performance for an 8 electrode configuration but it was not statistically significant. On the other hand, the symmetric approach always offered significantly higher performances.

As shown in Fig. 3, the transfer learning produced by the 36 features extracted by *EEGSym* between the pre-training and fine-tuning process is appropriate, since the starting point of the fine-tuning is similar to the ending of the pre-training. This is also shown by focusing in the first stage of the fine-tuning. In this stage only the last softmax is allowed to be fitted, so the model is being optimized over the 36 features extracted during the pre-training. Despite only tuning this last operation of the model, we reach a better fit than in the pre-training. What is more, the second stage only improves the validation loss by a minimum amount before overfitting and triggering the early-stopping mechanism.

The pre-training for Physionet [26] dataset in a 8 or 16 electrode configuration required a computation time of 4 hours and 18 minutes or 6 hours and 25 minutes, respectively. For a new application, only one pre-training opeartion is needed, and can be skipped if the pre-trained weight values present in our open implementation are used. The fine-tuning process in an 8 or 16 electrode configuration required a computation time of 7 and 12 minutes, respectively. This fine-tuning only needs to be performed the first time it is adapted to the desired MI-application, or any time there is a substantial increase of recorded trials over the first fine-tuning dataset. On inference mode, i.e. predicting a single trial, the model required 30 ms in both configurations running on a GPU. The 30 ms needed for a prediction make this DL approach also suitable for online decoding.

TABLE IV
COMPARISON WITH BINARY CLASSIFICATION OF PREVIOUS LITERATURE

	Study	TW (s)	#	$\mu\pm\sigma$
Physionet [26]	Dose et al. (2018) [24]	3	64	80.38±12.54
			16	78.03
			9	75.85
	Kostas et al. (2020) [32]	3	64	82.84
	Fan et al. (2021) [29]	3	64	82.88
			14	78.98
	Varsehi et al. (2021) [49]	3	14	83.63
			9	81.26
	Ours	3	16	**88.56**±8.96
			8	**84.45**±9.70
OpenBMI [27]	Kwon et al. (2020) [34]	4	19	74.15±15.83
	Zhang et al. (2021) [25]	4	62	84.19±9.98
	Ours	3	16	**84.72**±11.73 *
			8	**82.93**±12.10 *

#: number of electrodes. TW: time window duration used for classification. $\mu\pm\sigma$: mean accuracy and standard deviation obtained between all subjects in a subject-independent scheme. *Results are different from Table II to mimic the compared works [25], [34] where only the trials from the last test run were used for reporting accuracies.

B. Comparison with previous works

A comparison with previous studies can be found in Table IV. Physionet [26] dataset includes data from 109 subjects, but the works that we use for comparison excluded from their analysis the data of 4 subjects. Dose *et al.* [24] did not specify which subjects they exclude from their study. Furthermore, they extracted 42 trials from each user's 45 available trials, without specifying which ones to select. Fan *et al.* [29] and Varsehi *et al.* [49] removed subjects S088, S092, S100, and S104 for being damaged. However, Kostas *et al.* [32] excluded S088, S090, S092 and S100. In this work, since all subjects could be used, and noticing the disparity of excluded users in previous works, we decided to include every subject and all available trials.

The studies that addressed inter-subject classification with DL have partially exploited the ability that DL networks present for transfer learning [24], [25], [29], [32], [34]. They use the data of other subjects from the same dataset to train the network and evaluate on the rest of subjects or fine-tune the model to a specific subject of the same dataset. We believe that one of the clear advantages of our approach has

been to use data from multiple publicly available datasets that share an imagination paradigm. They were used for pre-training the network to initialize the weights of the models evaluated. This improved use of transfer learning is made clear when comparing the inter-subject accuracies on Physionet [26] dataset. All baseline models and *EEGSym* outperform previous DL approaches that used all 64 electrodes [24], [29], [32] available with the information of only 16 electrodes. Furthermore, *EEGSym* only needs 8 electrodes to overcome previous studies in this particular dataset. In OpenBMI [27] dataset *EEGSym* also obtains similar results as previous studies with only 16 out of the 62 electrodes of the dataset.

EEGSym outperforms the state-of-the-art models present in the literature with only 16 electrodes of the more than 60 available. It has been proved in Physionet [26] and OpenBMI [27] which include 109 and 54 subjects, respectively. Our results suggest that the combination of our preprocessing and pre-training with DA is a tool which enhances DL performance on this task.

C. Limitations and future work

Despite the positive results of *EEGSym* achieved in this study, we also acknowledge several limitations that should be addressed in the future. The proposed method reduces its performance without fine-tuning to the target dataset (accounting for the operator, device and procedure variability). This implies that implementing this model to a custom application will need to collect data from a few subjects to reach accuracies similar to this study. Therefore, there is still room to improve the generalization of the model towards a plug-and-play system. This could be solved by collecting more data from different centers and users to increase the publicly available resources.

The idea of introducing the known symmetry of the brain through the mid-sagittal plane into the network architecture enables it to reach higher classification accuracies and improves the generalization of the model. We have focused on the ability of the network for inter-subject classification. The ability to make the most of the available data by introducing known spatial relations needs to be extended to intra-subject classification by fine-tuning the model to each user.

Also, understanding better which features the DL networks are extracting would be very beneficial for further optimization of the task. This will fall into the explainable artificial intelligence (XAI) field, a very promising research line that could include developing a model with the consideration of its explainability.

V. Conclusion

In this study, we introduce *EEGSym*, a new CNN for binary MI classification. It includes the use of inception modules, residual connections to enhance spatial features extraction, and the incorporation of the symmetry of the brain through the mid-sagittal plane into its architecture design. It also makes use of transfer learning across subjects and datasets and of a DA technique that includes patch perturbation, hemisphere perturbation, and random shift. *EEGSym* improved state-of-the-art accuracies on inter-subject MI binary classification. These results are validated in 5 datasets with the largest amount of subjects (280) in related studies. *EEGSym* was compared to previous state-of-the-art CNNs: ShallowConvNet and DeepConvNet [22], EEGNet [23], and EEG-Inception [5]. The inter-subject scheme implemented in this study allowed *EEGSym* to be used without the need of calibration runs on new subjects and potentially solving the problem of BCI inefficiency. Furthermore, this new state-of-the-art accuracies were obtained with only 16 electrodes of the more than 60 available on some datasets. This reduced set of electrodes enables the use of more inexpensive EEG recording systems with a reduced set up duration. The combination of a reduced set up duration and the calibrationless application can boost users' motivation of MI-based BCIs, which is key for the use of this applications for rehabilitation. *EEGSym* outperforms previous state-of-the-art approaches on inter-subject MI classification reaching significantly (p-value < 0.05) higher accuracies on all 5 datasets tested and allows the higher number of users to reach BCI control.

References

[1] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain Computer Interfaces, a Review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, jan 2012.

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

[3] V. Martínez-Cagigal, E. Santamaría-Vázquez, J. Gomez-Pilar, and R. Hornero, "Towards an accessible use of smartphone-based social networks through brain-computer interfaces," *Expert Systems with Applications*, vol. 120, pp. 155–166, 2019.

[4] J. Meng, T. Streitz, N. Gulachek, D. Suma, and B. He, "Three-dimensional brain-computer interface control through simultaneous overt spatial attentional and motor imagery tasks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 11, pp. 2417–2427, 2018.

[5] E. Santamaria-Vazquez, V. Martinez-Cagigal, F. Vaquerizo-Villar, and R. Hornero, "EEG-Inception: A Novel Deep Convolutional Neural Network for Assistive ERP-based Brain-Computer Interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2773–2782, 2020.

[6] Y. Yu, Y. Liu, E. Yin, J. Jiang, Z. Zhou, and D. Hu, "An Asynchronous Hybrid Spelling Approach Based on EEG–EOG Signals for Chinese Character Input," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1292–1302, jun 2019.

[7] A. Ramos-Murguialday, M. Schürholz, V. Caggiano, M. Wildgruber, A. Caria, E. M. Hammer, S. Halder, and N. Birbaumer, "Proprioceptive Feedback and Brain Computer Interface (BCI) Based Neuroprostheses," *PLoS ONE*, vol. 7, no. 10, p. e47048, oct 2012.

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

[9] M. Carrillo-de-la Peña, S. Galdo-Álvarez, and C. Lastra-Barreira, "Equivalent is not equal: Primary motor cortex (MI) activation during motor imagery and execution of sequential movements," *Brain Research*, vol. 1226, no. Mi, pp. 134–143, aug 2008.

[10] N. Mrachacz-Kersting, M. Voigt, A. J. Stevenson, S. Aliakbaryhosseinabadi, N. Jiang, K. Dremstrup, and D. Farina, "The effect of type of afferent feedback timed with motor imagery on the induction of cortical plasticity," *Brain Research*, vol. 1674, pp. 91–100, 2017.

[11] A. A. Frolov, O. Mokienko, R. Lyukmanov, E. Biryukova, S. Kotov, L. Turbina, G. Nadareyshvily, and Y. Bushkova, "Post-stroke Rehabilitation Training with a Motor-Imagery-Based Brain-Computer Interface (BCI)-Controlled Hand Exoskeleton: A Randomized Controlled Multicenter Trial," *Frontiers in Neuroscience*, vol. 11, no. JUL, jul 2017.

[12] T. Corbet, I. Iturrate, M. Pereira, S. Perdikis, and J. d. R. Millán, "Sensory threshold neuromuscular electrical stimulation fosters motor imagery performance," *NeuroImage*, vol. 176, no. April, pp. 268–276, 2018.

[13] A. Vourvopoulos, C. Jorge, R. Abreu, P. Figueiredo, J.-C. Fernandes, and S. Bermúdez i Badia, "Efficacy and Brain Imaging Correlates of an Immersive Motor Imagery BCI-Driven VR System for Upper Limb Motor Rehabilitation: A Clinical Case Report," *Frontiers in Human Neuroscience*, vol. 13, no. July, pp. 1–17, jul 2019.

[14] D. T. Bundy, L. Souders, K. Baranyai, L. Leonard, G. Schalk, R. Coker, D. W. Moran, T. Huskey, and E. C. Leuthardt, "Contralesional Brain-Computer Interface Control of a Powered Exoskeleton for Motor Recovery in Chronic Stroke Survivors," *Stroke*, vol. 48, no. 7, pp. 1908–1915, 2017.

[15] A. Moldoveanu, O. M. Ferche, F. Moldoveanu, R. G. Lupu, D. Cinteza, D. Constantin Irimia, and C. Toader, "The TRAVEE system for a multimodal neuromotor rehabilitation," *IEEE Access*, vol. 7, pp. 8151–8171, 2019.

[16] M. Sebastián-Romagosa, W. Cho, R. Ortner, N. Murovec, T. Von Oertzen, K. Kamada, B. Z. Allison, and C. Guger, "Brain Computer Interface Treatment for Motor Rehabilitation of Upper Extremity of Stroke Patients—A Feasibility Study," *Frontiers in Neuroscience*, vol. 14, no. October, pp. 1–12, 2020.

[17] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki, "Correlation-based channel selection and regularized feature optimization for MI-based BCI," *Neural Networks*, vol. 118, pp. 262–270, oct 2019.

[18] J. Jin, R. Xiao, I. Daly, Y. Miao, X. Wang, and A. Cichocki, "Internal Feature Selection Method of CSP Based on L1-Norm and Dempster–Shafer Theory," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4814–4825, nov 2021.

[19] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in brain-computer interface," *Proceedings of the International Joint Conference on Neural Networks*, pp. 2390–2397, 2008.

[20] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, jul 2017.

[21] S. Saha and M. Baumert, "Intra- and Inter-subject Variability in EEG-Based Sensorimotor Brain Computer Interface: A Review," *Frontiers in Computational Neuroscience*, vol. 13, no. January, pp. 1–8, 2020.

[22] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[23] V. J. Lawhern, A. J. Solon, 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," *Journal of Neural Engineering*, vol. 15, no. 5, pp. 1–30, 2018.

[24] H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Expert Systems with Applications*, vol. 114, pp. 532–542, 2018.

[25] K. Zhang, N. Robinson, S. W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep Convolutional Neural Network," *Neural Networks*, vol. 136, pp. 1–10, 2021.

[26] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." *Circulation*, vol. 101, no. 23, jun 2000.

[27] M. H. Lee, O. Y. Kwon, Y. J. Kim, H. K. Kim, Y. E. Lee, J. Williamson, S. Fazli, and S. W. Lee, "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, pp. 1–16, 2019.

[28] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 07-12-June. IEEE, jun 2015, pp. 1–9.

[29] C. C. Fan, H. Yang, Z. G. Hou, Z. L. Ni, S. Chen, and Z. Fang, "Bilinear neural network with 3-D attention for brain decoding of motor imagery movements from the human EEG," *Cognitive Neurodynamics*, vol. 15, no. 1, pp. 181–189, 2021.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, 2016.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[32] D. Kostas and F. Rudzicz, "Thinker invariance: Enabling deep neural networks for BCI across more people," *Journal of Neural Engineering*, vol. 17, no. 5, 2020.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2017-Janua. IEEE, jul 2017, pp. 2261–2269.

[34] O. Y. Kwon, M. H. Lee, C. Guan, and S. W. Lee, "Subject-Independent Brain-Computer Interfaces Based on Deep Convolutional Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 3839–3852, 2020.

[35] L. Acqualagna, L. Botrel, C. Vidaurre, A. Kübler, and B. Blankertz, "Large-Scale Assessment of a Fully Automatic Co-Adaptive Motor Imagery-Based Brain Computer Interface," *PLOS ONE*, vol. 11, no. 2, p. e0148886, feb 2016.

[36] O. Alkoby, A. Abu-Rmileh, O. Shriki, and D. Todder, "Can We Predict Who Will Respond to Neurofeedback? A Review of the Inefficacy Problem and Existing Predictors for Successful EEG Neurofeedback Learning," *Neuroscience*, vol. 378, no. January, pp. 155–164, may 2018.

[37] J. Meng and B. He, "Exploring training effect in 42 human subjects using a non-invasive sensorimotor rhythm based online BCI," *Frontiers in Human Neuroscience*, vol. 13, no. April, pp. 1–19, 2019.

[38] M. Kaya, M. K. Binli, E. Ozbay, H. Yanar, and Y. Mishchenko, "A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces," *Scientific Data*, vol. 5, no. 1, p. 180211, dec 2018.

[39] J. R. Stieger, S. Engel, H. Jiang, C. C. Cline, M. J. Kreitzer, and B. He, "Mindfulness Improves Brain–Computer Interface Performance by Increasing Control Over Neural Activity in the Alpha Band," *Cerebral Cortex*, vol. 31, no. 1, pp. 426–438, jan 2021.

[40] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, 2019.

[41] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13 001–13 008, apr 2020.

[42] R. Psotta, "The visual reaction time distribution in the tasks with different demands on information processing," *Acta Gymnica*, vol. 44, no. 1, pp. 5–13, mar 2014.

[43] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2017-Janua. IEEE, jul 2017, pp. 5987–5995.

[44] N. Valliappan, N. Dai, E. Steinberg, J. He, K. Rogers, V. Ramachandran, P. Xu, M. Shojaeizadeh, L. Guo, K. Kohlhoff, and V. Navalpakkam, "Accelerating eye movement research via accurate and affordable smartphone eye tracking," *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.

[45] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[46] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, p. 80, dec 1945.

[47] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, jan 1995.

[48] C. Jeunet, B. Glize, A. McGonigal, J. M. Batail, and J. A. Micoulaud-Franchi, "Using EEG-based brain computer interface and neurofeedback targeting sensorimotor rhythms to improve motor skills: Theoretical background, applications and prospects," *Neurophysiologie Clinique*, vol. 49, no. 2, pp. 125–136, 2019.

[49] H. Varsehi and S. M. P. Firoozabadi, "An EEG channel selection method for motor imagery based brain–computer interface and neurofeedback using Granger causality," *Neural Networks*, vol. 133, pp. 193–206, 2021.