8-5-2021

An End-to-End CNN with Attentional Mechanism Applied to Raw EEG in a BCI Classification Task

Elnaz Lashgari
Chapman University, lashgari@chapman.edu

Jordan Ott
University of California, Irvine

Akima Connelly
Chapman University, conne147@mail.chapman.edu

Pierre Baldi
University of California, Irvine

Uri Maoz
Chapman University, maoz@chapman.edu

Recommended Citation

Lashgari, E., et al., *An end-to-end CNN with attentional mechanism applied to raw EEG in a BCI classification task*. Journal of Neural Engineering, 2021. https://doi.org/10.1088/1741-2552/ac1ade

An End-to-End CNN with Attentional Mechanism Applied to Raw EEG in a BCI Classification Task

Journal of Neural Engineering

An end-to-end CNN with attentional mechanism applied to raw EEG in a BCI classification task

To cite this article before publication: Elnaz Lashgari *et al* 2021 *J. Neural Eng.* in press https://doi.org/10.1088/1741-2552/ac1ade

An end-to-end CNN with attentional mechanism applied to raw EEG in a BCI classification task

Elnaz Lashgari [1,2*], Jordan Ott [3], Akima Connelly [1,2], Pierre Baldi [3,4,5], Uri Maoz [1,2,6,7,8,9]

[1] College of Science and Technology, Chapman University, Orange
[2] Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, Irvine
[3] Department of Computer Science, University of California, Irvine
[4] Center for Machine Learning and Intelligent Systems, University of California, Irvine
[5] Institute for Genomics and Bioinformatics, University of California, Irvine
[6] Computational Neuroscience and Psychology, Crean College of Health and Behavioral Sciences, Chapman University, Orange
[7] Fowler School of Engineering, Chapman University, Orange
[8] Anderson School of Management, University of California Los Angeles
9 Biology and Bioengineering, California Institute of Technology

*Corresponding author: lashgari@chapman.edu

Abstract

Objective. Motor-imagery (MI) classification base on electroencephalography (EEG) has been long studied in neuroscience and more recently widely used in healthcare applications such as mobile assistive robots and neurorehabilitation. In particular, EEG-based motor-imagery classification methods that rely on convolutional neural networks (CNNs) have achieved relatively high classification accuracy. However, naively training CNNs to classify raw EEG data from all channels, especially for high-density EEG, is computationally demanding and requires huge training sets. It often also introduces many irrelevant input features, making it difficult for the CNN to extract the informative ones. This problem is compounded by a dearth of training data, which is particularly acute for MI tasks, because these are cognitively demanding and thus fatigue inducing. ***Approach.*** To address these issues, we proposed an end-to-end CNN-based neural network with attentional mechanism together with different data augmentation (DA) techniques. We tested it on two benchmark MI datasets, Brain-Computer Interface (BCI) Competition IV 2a and 2b. In addition, we collected a new dataset, recorded using high-density EEG, and containing both MI and motor execution (ME) tasks, which we share with the community. ***Main results.*** Our proposed neural-network architecture outperformed all state-of-the-art methods that we found in the literature, with and without DA, reaching an average classification accuracy of 93.6% and 87.83% on BCI 2a and 2b, respectively. We also directly compare decoding of MI and ME tasks. Focusing on MI classification, we find optimal channel configurations and the best DA techniques as well as investigate combining data across participants and the role of transfer learning. ***Significance.*** Our proposed approach improves the classification accuracy for MI in the benchmark datasets. In addition, collecting our own dataset enables us to compare MI and ME and investigate various aspects of EEG decoding critical for neuroscience and BCI.

Keywords: EEG, motor imagery, CNN, Attentional mechanism, Data augmentation, Transfer learning

1. Introduction

Advances in brain science and computer technology in the past decade have led to exciting developments in Brain-Computer Interfaces (BCI), thereby making BCI a key research area in applied neuroscience and neuro-engineering [1]. Non-invasive BCI facilitates new methods of neurorehabilitation for physically disabled people (e.g., paralyzed patients and amputees) and patients with brain injuries (e.g., stroke patients) [1]. BCI systems utilize recorded brain activity to directly communicate between the brain and computers to control the environment in a manner compatible with the individual's intentions [2].

However, the ability to decode intentions is also an important tool for basic neuroscientific research. In particular, it strongly enhances the scientific armamentarium used to investigate volition [3, 4]. And, more specifically, decoding intention in real time would open the door to interesting experimental possibilities, such as interventions to facilitate or frustrate intentions [5-7], and intention-contingent stimulation [3]. Technological advances of recent decades—such as untethered, wireless recording, machine-learning-based analysis, and real-time analysis of raw EEG signal have increased the interest in electroencephalography (EEG) based BCI approaches [8].

EEG has proved to be the most popular brain-imaging method for BCI because it is inexpensive, noninvasive, directly measures neural activity (as opposed to fMRI for example), and can facilitate portability to clinical use [2]. EEG signals thus serve as pathways from the brain to various external devices, resulting in brain-controlled assistive devices for disabled people and brain-controlled rehabilitation devices for patients with strokes and other neurological deficits [1, 9, 10]. One of the most challenging topics in BCI is finding and analyzing the relations between recorded brain activity and underlying models of the human body, of biomechanics, and of cognitive processing. The investigation of relations between EEG signals and—real and imagined—upper limb movement has gained more attention in recent years [11, 12].

To implement an EEG-based BCI system for a particular application, a specific experimental protocol and paradigm must be chosen for all phases of the experiment. Typically, the participant first performs a particular task (e.g., a motor-imagery task, a visual task) to learn how to modulate their brain activity, while EEG signals are simultaneously recorded from their scalp. Using the recorded EEG as training data, a machine-learning-based neural decoder for the paradigm is then constructed [1]. Finally, the participant performs the task again, and the neural decoder is used for BCI control.

The process for BCI systems based on motor imagery (MI) is similar. Though, in this case, the participant imagines the movement rather than actually executing it [11]. Previous studies have confirmed that imagination activates areas of the brain that are responsible for generating actual movement [1, 13]. The most common MI paradigms reported in literature are based on sensorimotor rhythms (SMR) and imagined body kinematics. In the SMR paradigm (e.g., [14, 15] participants imagined kinesthetic movements of some body part—such as hands, feet, or tongue—which result in modulations of brain activity that are trackable using EEG [16]. Imagined movement in such SMR paradigms often causes event-related desynchronization (ERD) in mu (typically 8-12 Hz) and beta rhythms (roughly 12-30 Hz). In contrast, relaxing after MI results in event-related synchronization (ERS) [17]. The ERD and ERS modulations are most prominent in EEG signals acquired from electrode locations C3 and C4 (in the 10/20 international system); these electrodes are approximately above the motor cortices of both brain hemispheres.

MI classification is one of the most popular EEG-based BCI paradigms. EEG MI classification generally consists of four parts: signal acquisition, feature extraction, classification, and control. Most existing feature-extraction methods depend on manually designed features, based on human knowledge. Feature extraction and classification of EEG signals for MI tasks have been attempted in the time, frequency, and space (electrodes) domains—not necessarily mutually exclusively. Time-frequency feature extraction in EEG has focused mostly on short-time Fourier transform [18, 19] or wavelets [20, 21]. In the space domain, filter-bank common spatial-patterns (FBCSP) has achieved notable performance [22, 23]. However, FBCSP uses a fixed temporal duration, ignoring difference between participants. As such, it does not make full use of time-domain information. Moreover, these methods generally use handcrafted features and require heuristic parameter setting—e.g., predefined frequency bands—which often do not generalize well across tasks and participants [24]. As such, they often result in limited classification accuracy [20, 25, 26].

2. Related work

Recently, researchers have successfully used deep learning (DL) to perform automatic feature extraction [27] and classification [24, 28-30]. DL has achieved breakthrough accuracies and discovered intricate structures in various complex and high-dimensional data [31, 32]. In particular, it has provided promising results in the analysis and decoding of EEG signals [33]. Thus, NN architectures, their training procedures, regularization, optimization, and hyper-parameter settings are all active area of research in DL-based analysis of EEG, with advances often resulting in dramatic increases in decoding accuracy [33].

Recently, Zhang et al., proposed a hybrid DL architecture, which combined convolutional neural networks (CNNs) and long short-term memory (LSTM) models to handle sequential time domain data [34]. Even more recently, Dai et al., proposed an architecture composed of a CNN with a hybrid convolution scale (HS-CNN), which separates a signal into three frequency bands using bandpass filters at 4∼7 Hz, 8∼13 Hz, and 13∼32 Hz. The three frequency bands are then fed into the convolutional layers with different filter sizes [24]. The features, including different semantic information, were concatenated and then MI classification was carried out. In another study, Zhang et al., applied an attention module to LSTM to utilize long-range information for EEG-based hand-movement classification [35].

Despite their promise, these deep NN architectures are not easy to train from scratch, because they require large amounts of training data to achieve high classification accuracy. However, it is particularly challenging to obtain a large amount of training samples for MI classification. This is because gathering high-quality data requires training and experience as well as a state-of-the-art EEG machine and a noise-free environment. MI tasks are also time consuming and fatigue-inducing for the participants. For example, during the task, participants must minimize, if not altogether avoid, eye movements and other muscle contractions, especially around the head. At the same time, they typically need to employ a great deal of concentration and attention during MI tasks. Thus, participants can only produce a limited amount of data at each session and must come in for multiple sessions to construct a large dataset of EEG MI. This often results in attrition over the course of multiple sessions.

Data augmentation (DA) can lead to considerable performance gains for DL, reducing overfitting and increasing overall accuracy and stability. DA generates new samples to augment an existing dataset by transforming existing samples in some systematic manner. Exposing the classifiers to various transformations of the training samples, as DA does, makes the models more robust and

invariant to these and potentially other transformations when attempting to generalize beyond the training set [36-38].

DA is an especially important technique for EEG-based BCI because of its specific combination of two factors: the dimensionality of EEG signals tends to be high, while the number of available training samples tends to be low. In a recent systematic review on DA in EEG, Lashgari et al. collected all the papers that used DA for NN-based analysis of EEG up to and including 2019 [33]. They showed that convolutional neural networks (CNN) were the most popular NN architectures for EEG MI classification and typically resulted in accurate decoding. This is likely because CNNs are well suited to end-to-end learning, scale well to large datasets, and can exploit hierarchical structure in natural signals. The review also found that the most common input formulation for motor tasks and MI was raw EEG signals [33].

With these elements in mind, here we investigated the efficacy and generalizability of deep learning on EEG-based decoding of MI. We designed an end-to-end CNN with an attentional mechanism [39]. This is because a CNN with an attention-mechanism architecture can improve classification performance using EEG signals by focusing on essential, task-relevant features on different time-steps.

We begin by testing this architecture on 2 benchmark datasets (BCI Competition IV 2a and 2b) as well as on the dataset that we collected, which we share with the community. Then, we compare MI to ME on the dataset that we collected. Next, we tackled a common question when collecting EEG data: how many channels to record for optimal decoding accuracy? We thus compared the decoding accuracy for different numbers of channels. It has also been demonstrated that DA techniques hold promise for EEG decoding. So, we also tested how much DA can boost the accuracy of our method across the datasets. How much EEG data is needed to train deep NN is also not well understood, especially in relation to DA techniques. We therefore next investigate how the accuracy of our model depends on the amount of data on which we train and the type and amount of DA we use. Of course, structure and anatomical features vary across brains. So, we further investigated what happens to the decoding accuracy when we train and test it on EEG from single participants, on pair of participants, triplets, and so on. In the interest of understanding how well models of EEG decoding generalize to previously unseen participants, we also investigated what happens when we train the model on all but one participant and then test on that remaining participant, with and without transfer learning.

3. Methods

3.1. Proposed CNN-based neural-network architecture

Convolutional models have been successful in many signal processing applications, as they allow temporally related inputs to be processed together via a sliding-window approach (Figure 1). This produces shared weights, where the same weight kernel is applied across the temporal domain (for a 1D convolutional model over time). In our architecture (Figure 1), this reduces the number of parameters needed in such a model and enables the signal to maintain its spatial relations—across time within each electrode and across electrodes over the head. The signal from each electrode channel is fed through the same convolutional base to produce an output matrix of dimension $C \times E$, where C is the number of electrodes (or channels) and E is the size of the embedding

dimension (Figure 1). Hence, the convolutional layers in effect reduce the dimension of the input to the embedding dimension, E.

Now, in the self-attention part of the network [39, 40], we first initialize the weights for the Query (Q), Key (K), and Value (V). The magnitudes of Q, K, and V are derived by the product of the input (I) and the weights. The second step is to calculate the attentional score (S): $S = QK^T$. The shape of S will be $C \times C$. The Softmax (W) of S is calculated to return a vector of C x 1. The third step is to find the weighted values (M), $M = WV^T$. Each input's value for M is concatenated to return a shape of C x C, which will be the value for the final Attention. $Tanh$ was used to produce the alignment score. In the following, the equations show more details:

I	Input for self-attention, shape (number of channels (C) x the size of the embedding dimension (E))
Key, Value, Query	Initialize weights for key, value and query with shape of input size (C x E)
K = I x Key^T	Derive key, query, and value
V = I x $Value^T$	Shape (C x C)
Q = I x $Query^T$	
$S = Q \cdot K^T$	Calculate attention score by dot product (C x 1)
$W = Softmax(S)$	Calculate Softmax (C x 1)
$M = W \times V$	Multiply scores with value
$O = tanh\,(M \times W^T)$	Linear transformation of M

The attention layer discussed above is added after the convolutional base (Figure 1), so that each electrode channel is computed with every other channel to produce a matrix of scalar values. Summing across rows and normalizing these scalars produces a vector of attention scores. These scores are used to create a linear combination of all the electrode channel vectors, which is passed to the fully connected layers of the network for classification. A valuable part of this model is therefore its interpretability [5, 41, 42]. The attention scores for each electrode channel can be examined to determine the importance of each electrode in the model's prediction. However, in this study we were not interested in the added interpretability that the attentional mechanism affords us. Instead, we relied on the attentional mechanism to improve the prediction accuracy of our architecture. This is because a CNN with attention-mechanism architecture can improve classification performance using EEG signals by focusing on essential, task-relevant features on different time-steps, via the sliding windows. Table 1 shows the summary of the proposed NN parameter.

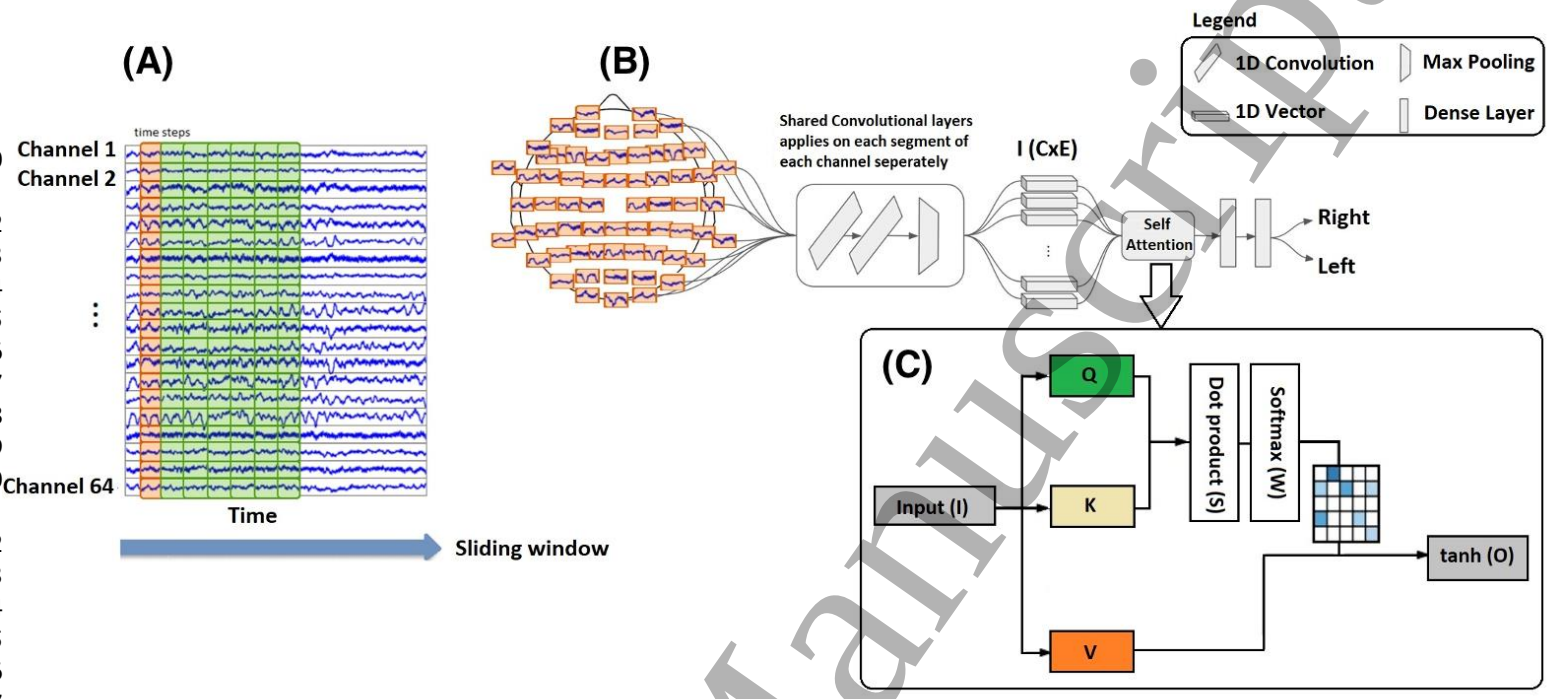


Figure 1. Our proposed CNN with attentional mechanism. (A) The sliding window (length is 1000ms and step-size is 100ms) applied to 64 EEG channels. (B) The 64 segments of raw EEG signal, depicted in orange in (A). Each time window and channel are separately sent through shared convolution layers. The embedded features I (C x E) applied to self-attention. The output of self-attention passes through 2 dense layers. (C) An expansion of the self-attention block.

Table 1 Summary of the proposed CNN with attentional mechanism parameters ("-1" represents a flexible shape, essentially the batch size)

Layer (Type)	Output Shape	Param #	Shared convolutional layer
Convolution 1D	[-1, 16, 64]	816	x64
Convolution 1D	[-1, 16, 6]	12,816	x64
Max Pooling 1D	[-1, 16, 3]	0	x64
1D Vector	[-1, 48]	2,304	0
Attention	[[-1, 64, 48], [-1, 64, 64]]	4,608	0
Dense	[-1, 32]	98,336	0
Dense	[-1, 2]	66	0
Total parameters	**977,762**		

3.2. Hyperparameter Optimization and Training

When implementing NN there are several choices (or hyperparameters) that must be set prior to training—those range from the type of architecture to the depth and width of the layers, through to the neuronal activation-function in the different layers, and so on. Choosing hyperparameters arbitrarily is likely to lead to suboptimal results. To address this, we first created a 3-way split of our data into a training, validation, and test sets to identify reasonable architectures and parameter ranges. Then, guided by those preestablished ranges, we conducted NN optimization via a Bayesian hyperparameter search using SHERPA [43], a Python library for hyperparameter tuning. The Bayesian search has the advantage of learning a distribution over the hyperparameters of the network architecture, in relation to the task to be optimized. By employing this procedure, we were able to evaluate a large space of possible models and test many configurations.

We detail the hyperparameters of interest in Table 2, as well as the range of available options during the search. The hyperparameters of interest consisted of the activation function, dropout percentage, learning rate, learning rate decay, nodes per layer, and the optimizer. Additional hyperparameters for convolutional models included the number of filters and the kernel size. We tried 250 different hyperparameter settings for each network architecture (Dense NN, Conv Net-Dense NN, Conv Net-Attention-Dense NN), for a total of 750 models over 3 different NN (Dense NN, Conv Net-Dense NN, Conv Net-Attention-Dense NN). Table 3 present the result of best hyperparameters tuning by SHERPA for the 3 datasets: BCI competition IV 2a (BCI 2a), BCI competition IV 2b (BCI 2b), and our dataset and for 3 different models (Dense, CNN-Dense, and CNN-Attention-Dense).

For the 3 datasets examined in this study, we adhered to the following procedure. For each set of hyperparameters sampled in the search, we partitioned each subject's data into a training and validation set. The proposed architecture was thus trained on each subject separately. Then, to evaluate the architecture, we averaged the validation accuracy scores across subjects. We then selected the network architecture with the highest average accuracy score across all subjects. Critically, this process ensures that we find architectures that perform well across subjects, but which are not tailored to specific subjects or tasks.

All networks were trained for 250 epochs using an early stopping condition—i.e., when the accuracy on the validation set did not improve for 25 epochs, training stopped. All models were trained using 10-fold cross-validation. The partitioning was stratified to ensure a constant ratio of representation amongst right and left examples—roughly 50/50—in keeping with the ratio in the data overall. This cross-validation procedure requires a given model to be trained 10 distinct times (re-initializing the network parameters each time) and ensures that, on the one hand, different subsets of the data are used for training and testing, while on the other hand, each datapoint serves as part of the training set (9 times) and in the test set (once). To be clear, when we performed cross validation, we used data partitions that were not used during the hyperparameter search. The accuracies reported below are therefore always the average accuracies across the 10 validation sets described above.

To double check our results, we carried one additional train/validation/test split of 75/15/10%, respectively. After this train/validation/test procedure, we ended up with neural architectures that were the same as those selected by the cross-validation procedure above—both in terms of the number of layers and the kernel size. This gave us confidence that our results are not due to some leakage between the training and test sets. Our cross-validation procedure allowed us to report

confidence scores, in the form of average accuracies and standard deviations. It also demonstrated that we did not cherry pick a data partition in which the proposed architectures happened to perform well; rather, our models were robust across partitions.

Training took place on NVIDA Titan V GPUs with 12GB of memory. Each epoch took less than a minute to complete. Training for a single fold typically completed within 30 minutes.

Table 2. The hyperparameter space

Name	**Range**	**Type**
Activation	(ReLU, ELU)	Choice
Dropout	(0, 0.9)	Continuous
Kernel Size	(25, 50, 75)	Choice
Learning Rate	(0.0001, 0.1)	Continuous
Learning Rate Decay	(0.5, 1.0)	Continuous
Number of Dense Nodes	(8, 512)	Discrete
Number of Filters	(16, 32, 64)	Choice
Optimizer	Adam, SGD, RMSProp	Choice

Table 3 Hyperparameter tuning by SHERPA for 3 the datasets (BCI 2a, BCI 2b and our experimental dataset) for 3 different models (Dense, CNN dense, and CNN attention dense)

dataset	Model	Kernel Size	Activation	Dropout	Learning Rate	Learning Rate Decay	Number of filters	Dense Nodes	Optimizer
BCI 2a	Dense NN	NAN	ReLU	0.171	0.017	1	NAN	27	Adam
	Conv Net-Dense NN	25	ELU	0.092	0.052	1	64	303	SGD
	Conv Net-Attention-Dense NN	25	ELU	0.9	0.1	1	32	91	SGD
BCI 2b	Dense NN	NAN	ReLU	0.845	0.001	0.864	NAN	289	Adam
	Conv Net-Dense NN	50	ReLU	0	0.1	1	16	15	SGD
	Conv Net-Attention-Dense NN	25	ELU	0	0.1	1	64	263	SGD
Our dataset	Dense NN	NAN	ReLU	0.687	0.037	1	NAN	369	SGD
	Conv Net-Dense NN	25	ReLU	0.68	0.034	0.989	32	196	SGD
	Conv Net-Attention-Dense NN	50	ELU	0.807	0.1	0.978	32	183	SGD

3.3. Data augmentation

Generally, in machine learning, but especially for NN, the classification accuracy tends to critically depend on the amount of training data; limited training data typically leads to low accuracy. DA comprises the systematic generation of new samples to augment an existing dataset by transforming existing samples in a manner that increases the accuracy and stability of classification [33]. Exposing the classifiers to varied representations of its training samples typically makes the model more invariant and robust to such transformations when attempting to generalize the model to new datasets. DA for the MI task fell into 5 categories in our analysis: noise addition [44, 45], GAN [46-49], sliding window [30, 50, 51], Fourier transform [38], and recombination of segmentation [24]. Table 4 shows more details about each of these methods. We evaluate all DA techniques with a magnification factor $m = (2, 5, 10, 15, 20, 30, 50)$ for our proposed CNN.

Table 4. DA techniques that are used on the MI task

DA methods	Details of the method
Sliding window [30, 50, 51]	Sliding window over the input of each trial, which leads to many more training examples for the network compared to using than the entire. More formally, given an original trial $X^j \in \mathbb{R}^{E\times T}$, with E electrodes and T timesteps, we create a set of crops with crop size T' as time slices of the trial: $C^j = (X^j_{1,\dots,E;\, t,\dots t+T'} \mid t \in 1,\dots T-T')$. All of these $T-T'$ crops then become training examples for our CNN and will get the same label, y_j, as the original trial. The best results in the BCI dataset are for 1s window length. In this study, we tried to evaluate this technique with different m and 100 ms step-size.
Noise Addition [44, 45]	We found two main categories for adding noise to the EEG signals in purpose of DA: (1) Add various types of noise such as Gaussian, Poisson, Salt and pepper noise, etc. with different parameters (for instance: mean (μ) and standard deviation (σ) to the raw signal (2) Convert EEG signals to sequences of images and add noise to the images [33]. Our proposed end-to-end CNN is for raw EEG. Therefore, we add noise just on the raw EEG signal. We add Gaussian noise with different parameters (mean = 0, standard deviation $\sigma = (0.01, 0.1, 0.2, 0.5)$ to all channels of raw EEG signal.
GANs [46-49]s	The GAN framework consists of two opposing networks trying to outplay each other [52]. The discriminator (D) is trained to distinguish between real and fake input data. The generator (G) takes a latent noise variable z as input and tries to generate fake samples that would not be recognized as fake by the discriminator. To learn a generator distribution p_g over data x, the generator builds a mapping function from a prior noise distribution $p_z(z)$ to data space as $G(z;\theta_g)$. And the discriminator, $D(x;\ \theta_d)$, outputs a single scalar representing the probability that x came from training data rather than pg. G and D are both trained simultaneously: we adjust parameters for G to minimize $log(1-D(G(z))$ and adjust parameters for D to minimize $logD(x)$ [52]. This results in a minimax game in which the generator is forced by the discriminator to produce ever better samples with value function $V(G,D)$: $$min_G max_D\, V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))].$$ GAN can be extended to a conditional model if both generator and discriminator are conditioned on some extra information such as y. In conditional generative adversarial nets (cGANs) y could be any kind of auxiliary information, such as class labels or data from other modalities. We can perform the conditioning by feeding y into the both the discriminator and generator as additional input layer. In the generator the prior input noise $p_z(z)$, and y are combined in joint hidden representation, and the adversarial training framework allows for considerable flexibly in how this hidden representation is composed. In the discriminator x and y are presented as inputs and to a discriminative function. The objective function of a two-player minmax game would be as: $$min_G max_D\, V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x|y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z|y)))].$$

Recombination of segmentation [24]	Perform segmentation on the input trials (i.e., left-/right-hand MI) with the same label. Each trial is segmented into three crops. The crops with the same labels are then recombined to generate new trials. For the same person and the same class, the crops at the same position from multiple trials are randomly swapped and recombined in the time/frequency domain to generate recombined trials [24].
Fourier Transform/Wavelet [38]	Apply the empirical mode-decomposition algorithm on the EEG frames and mixed their intrinsic mode functions to create new, artificial EEG frames [38]. The algorithm decomposes the original EEG signals into a finite number of functions called "intrinsic mode functions" (IMFs). Once the signal has been decomposed, we can recover it by adding all the IMFs and the residue without loss. To generate the new samples, we swapped the IMFs of the decompositions. Moreover, the intrinsic characteristics of each class (left/right) will be preserved because we mixed the IMFs of the same class. We randomly select the trials that contribute with their IMFs to generate samples for specific class.

3.4. Dataset and experimental protocol

We used three datasets in this study: (1) A dataset that we collected ourselves, (2) the BCI 2a dataset [53], and (3) the BCI 2b dataset [54] (Figure 2).

Our dataset: Seven healthy volunteers (3 male and 4 female) participated in the study, all were right-handed and between the ages of 23 to 30 (mean age 28). All participants gave written, informed consent to participate in the study. Participants were seated in a chair at a distance of 80 cm from an LCD screen with both hands resting on a Table. They held a tennis ball in each hand and were told to remain relaxed and strive to minimize movement and eye blinks. When required to respond, they were to squeeze the tennis ball in their hand but try to avoid tensing their arms or shoulders. Each session (ME and MI—Figure 2) was repeated twice. The whole experiment thus consisted of four sessions. Every session lasted 30–40 minutes with 10 to 15 minutes breaks between sessions. The duration of the whole experiment, including setup, was kept below 3 hours to minimize fatigue. EEG data was recorded and sampled at 250 Hz using 64 active electrodes (BrainVision actiCHamp) placed according to the 10/20 montage. Bipolar electromyography (EMG) electrodes were placed on the Brachioradialis for both hands as a sanity check for any movement in MI session.

Sessions 1 and 3 were designed to identify EEG signals related to ME. Participants were instructed to squeeze the tennis ball with their right or left hand while fixating on the cross displayed on the screen. They were encouraged to minimize all other movement and to only use the designated hand. One hundred trials were collected for each hand.

Session 2 and 4 aimed to show that a decoding model based on actual ME, derived from the first session, could be used to decode EEG activity in the absence of execution. Participants were instructed to carry out MI of the repetitive hand movement instructed in session 1 while fixating on the cross displayed on the screen. One hundred trials were collected for both left and right imagination per each session. All other aspects of the task were identical to session 1. This session also allowed us to screen participants for the presence of motor-related EEG oscillations, and at least minimal voluntary control over these oscillations. Hence, overall, we collected 200 trials of

ME and 200 trials of MI for each subject. The data underlying this study have been uploaded to figshare.

Data are available from the following link: https://doi.org/10.6084/m9.figshare.14721297.v1

BCI 2a: BCI 2a contains EEG data from 9 healthy participants [53], 2 sessions per participant. Each session is made up of 288 trials, resulting in 5184 trials overall. No feedback was provided. Twenty-two Ag/AGCL channels were used to record EEG. The signals were sampled with 250 Hz and bandpass filtered between 0.5-100 Hz. To compare our results with previous studies ([24], [55], [56] etc.)we focused on the C3, CZ, and C4 electrodes.

BCI 2b: BCI 2b contains EEG data from another 9 healthy participants [54]. For each participant, 5 sessions of data are collected. Each of the first 2 sessions has 120 trials and each of the last 3 sessions has 160 trials. The total number of trials is thus 6480. Two types of trials are included in these datasets: left- and right-hand MI. The first 2 sessions contain training data without feedback, while the last three sessions gave a smiley face as feedback. The EEG data is again collected over the C3, CZ, and C4 electrodes, which were placed following the international 10–20 system. The sampling frequency was 250 Hz. Table 5 presents the summary of three datasets.

Table 5. Summary of the 3 datasets used in this study: Our experimental dataset, BCI 2a, and BCI 2b

	Experimental dataset	**BCI 2a**	**BCI 2b**
The dataset provided by	The Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University	The Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology	The Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology
Open-source dataset	Yes	Yes	Yes
Description of dataset	2-class MI and ME (left hand and right hand). Session 1 and 3 are ME and 2and 4 MI, No feedback.	4-class MI (left hand, right hand, both feet, and tongue. No feedback	2-class MI (right hand, left hand). The first two sessions contain training data without feedback, and the last three sessions with smiley feedback.
# Channels	64 EEG channels (0.5-100Hz -BrainVision actiCHamp	22 bipolar EEG channels (0.5-100Hz; notch filtered)	3 bipolar EEG channels (0.5-100Hz; notch filtered)
Sampling frequency	250 Hz	250 Hz	250 Hz
# Subjects	7	9	9
# Sessions per subject	4	2	5
# Trials per session	100	288	120 for first 2 sessions and 160 trials for last 3 session
Total trials for each subject	400	576	720
Total trials in the dataset	2800	5184	6480

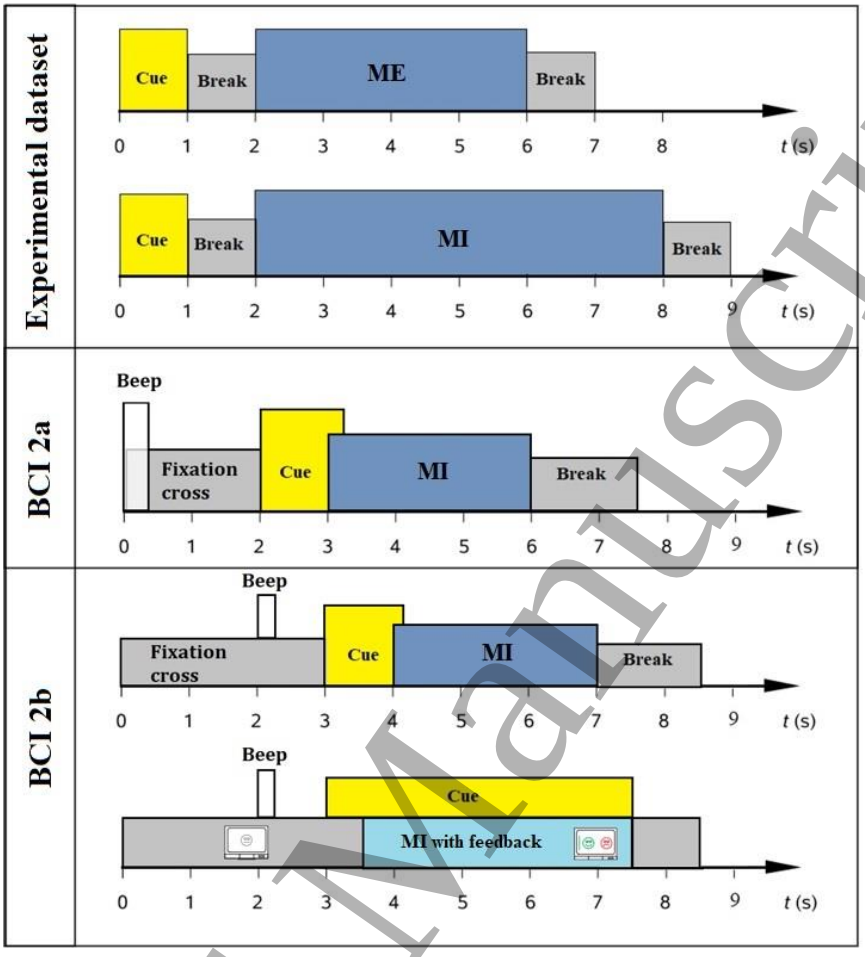


Figure 2 The experimental paradigms for our experimental dataset, BCI 2a, and BCI 2b.

3.5 Channel selection

Analyzing dense-array EEG is computationally expensive and complex; it also typically requires more expensive EEG systems than those with sparser electrodes. We therefore tested 4 different electrode configurations on our participants—which included 3, 7, 18, or all 64 electrodes (see Methods)—to further test the effect of channel selection on classification accuracy for MI in our own dataset.

Configuration (1) C3, CZ, and C4 electrodes were chosen in accordance with the 10-20 framework [57] since these electrodes have been shown to be especially discriminatory in hand and foot movements data [58]. It should be noted that right (left) hand's MI operation is usually detected above the left (right) motor cortex underneath the C3 (C4) electrode, and the foot's MI action is typically captured by the CZ electrode.

Configuration (2) The brain's frontal, central and parietal lobes are important from a neurological perspective for MI commands. We therefore also focused on these 7 electrodes (i.e. F3, F4, C3, CZ, C4, P3 and P4), which reside above these lobes of interest according to the 10-20 standard are considered in criteria 2 [57].

Configuration (3) Electrodes that are generally placed around the left and right motor cortices are included in this configuration because they are related to MI. According to 10-20 electrode montage [57], 18 electrodes lie around motor cortex. These are labelled C5, C3, C1, C2, C4, C6, CP5, CP3, CP1, CP2, CP4, CP6, P5, P3, P1, P2, P4 and P6 [59, 60].

Configuration (4) We used all 64 EEG channels.

In Figure 3, we showed these four configurations.

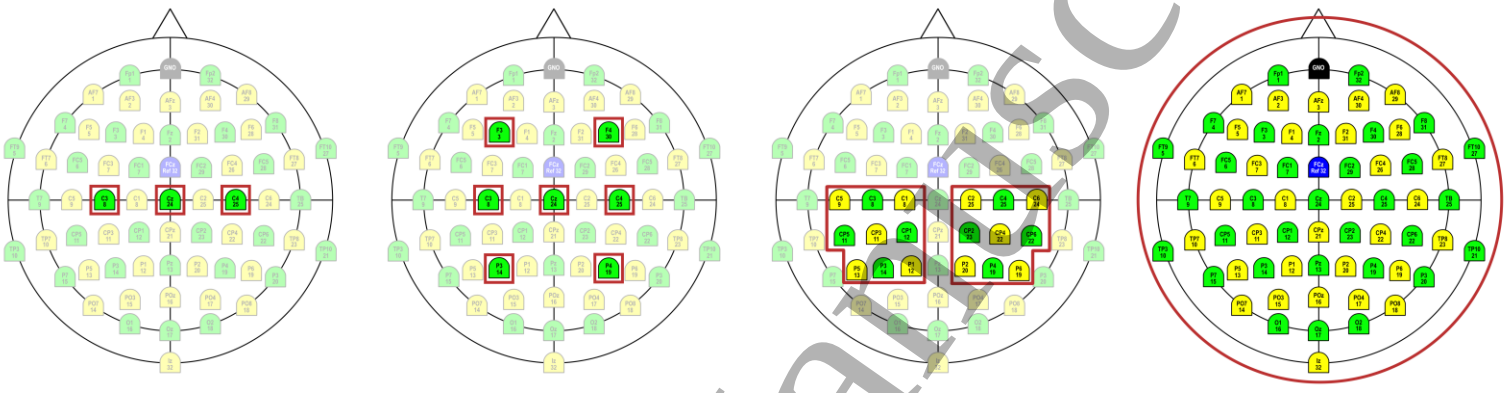


Figure 3. Four different electrode configurations on the actiCAP—which included 3, 7, 18, and all 64 electrodes

4. Results

4.1- Performance of the proposed CNN (Neural architectures vs. Neural architectures)

To evaluate the performance of our proposed CNN, we conducted comparisons between the Dense NN, Conv Net-Dense NN, Dai et al. (2020), and Conv Net-Attention-Dense NN (Figure 4). The baseline Conv Net is identical to the Conv Net-Attention-Dense NN but lacks the attention module (see Methods, Figure 1, Table 1). The dense network sends all channels through 2 dense layers, then it concatenates all the vectors into a single one and sends that through 2 more dense layers. We used SHERPA for hyperparameter optimization for all 4 types of networks [43]. We also reproduced the proposed NN in Dai et al. (2020) [24] without the use of DA to compare it with the proposed CNN with the attentional mechanism.

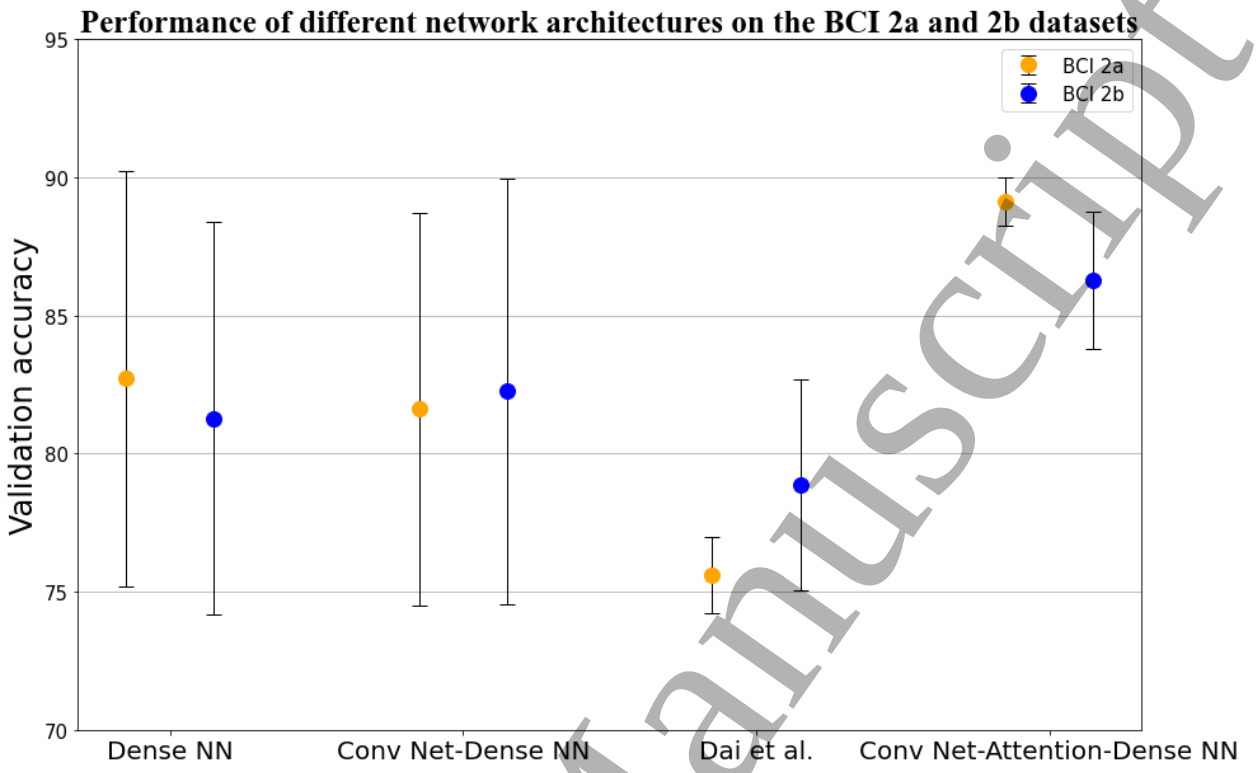


Figure 4 Comparison the average validation accuracy (±SE) on the BCI 2a and BCI 2b datasets with Dense, CNN, Dai et al. (2020), and CNN-Attention-Dense (See section 2.1 and 2.2).

Table 6 represents the classification results of our proposed CNN (with the attentional mechanism) without DA and with DA, which resulted in the highest accuracy for both datasets. Those are further compared against the results of Dai et al. [24]. All classifications were carried out on the BCI 2a and BCI 2b datasets. The average accuracy in Dai et al. (2020) for BCI 2a and BCI 2b were 91.57% (±5.73) and 87.6% (±8.48), respectively. In comparison, our proposed method with DA (GAN and m=15) achieved an average accuracy of 93.6% (±2.59) for BCI 2a and 87.83% (±6.34) for BCI 2b. Hence, our method has a higher average accuracy than Dai et al. (2020) while maintaining less variability in the accuracy across participants for both datasets. For the BCI 2a, our proposed method was 90.54% or higher for all participants while Dai et al. (2020) got this accuracy just for 5 of 9 participants (56%). Furthermore, we reproduced the NN described in [24] without the use of DA to compare with our proposed CNN with the attentional mechanism without DA. Our results on the BCI 2a and 2b datasets were 89.11% (±3.77) and 86.28% (±7.41), respectively, outperforming those of [24] at 75.61% (±14.63) and 78.88% (±11.42), respectively. Again, our results were also less variable than theirs.

Table 7 further compares our results with various other state-of-the-art methods. As is apparent from the Table, our results outperform all others, typically by a wide margin. On average, our method is 16.44 % and 7.21% more accurate than the other method for the 2a and 2b datasets, respectively. What is more, even without DA, our method has a higher average accuracy than all other methods except for Dai et al. (2020). And, with DA, our method beats all other methods, including Da. et al.'s.

Table 6. Participant-by participant comparison of the proposed CNN with attentional mechanism—with and without DA—against Dai et al. [24] results on the BCI 2a and BCI 2b datasets.

	BCI 2a				BCI 2b			
Participant	**Dai et al. (2020)[24]**	**Reproduced the result in [24] (without DA)**	**Proposed method without DA**	**Proposed method with DA (GAN m=15)**	**[24]**	**Reproduced the result (without DA)**	**Proposed method without DA**	**Proposed method with DA (sliding window m=2)**
1	90.07%	69.77%	91.58%	95.38%	80.50%	70.83%	81.64%	84.13%
2	80.28%	65.62%	89.67%	91.25%	70.60%	63.24%	73.17%	77.92%
3	97.08%	97.91%	91.89%	91.25%	85.60%	62.64%	81.50%	83.64%
4	89.66%	69.45%	90.05%	96.12%	94.60%	97.84%	98.61%	99.18%
5	97.04%	62.51%	91.28%	95.05%	98.30%	80.95%	93.83%	94.97%
6	87.04%	62.48%	90.97%	94.62%	86.60%	80.28%	85.22%	85.83%
7	92.14%	66.66%	81.38%	91.22%	89.60%	84.58%	86.57%	86.57%
8	98.51%	90.64%	91.20%	90.54%	95.60%	86.05%	89.90%	90.50%
9	92.31%	95.46%	83.95%	97.50%	87.40%	83.47%	86.05%	87.73%
AVG	91.57%	75.61%	89.11%	93.60%	87.60%	78.88%	86.28%	87.83%
S.D.	5.73	14.63	3.77	2.59	8.48	11.42	7.41	6.34
S.E.	1.91	4.87	1.26	0.87	2.83	3.81	2.47	2.11

Table 7. Comparison of our proposed method (with and without data augmentation) with other state-of-the-art methods. All methods were run on the same dataset (BCI 2a and/or BCI 2b).

	[61]	[62]	[63]	[56]	[64]	[29]	[18]	[65]	[66]	[67]	[55]	[68]	[24]	Proposed method (without DA)	Proposed method (with DA)
Dataset	**2b**	**2b**	**2b**	**2a/2b**	**2b**	**2b**	**2b**	**2b**	**2a**	**2a/2b**	**2a/2b**	**2a**	**2a/2b**	**2a/2b**	**2a/2b**
S1	77.0	70.0	80.0	63.69/73.2	84.6	81.0	76.0	72.5	88.9	90.28/70.3	66.7/62.8	91.5	90.07/80.5	91.58/81.64	95.38/84.13
S2	64.5	60.0	66.0	61.97/67.5	66.3	65.0	65.8	56.4	51.4	57.64/50.6	63.9/67.1	60.6	80.28/70.6	89.67/73.17	91.25/77.92
S3	61.0	61.0	53.0	91.09/63	62.9	66.0	75.3	55.6	96.5	95.14/52.8	77.8/98.7	94.2	97.08/85.6	91.89/81.50	91.25/83.64
S4	96.5	97.5	98.5	61.72/97.4	95.8	98.0	95.3	97.2	70.1	65.97/93.8	63.2/88.4	76.7	89.66/94.6	90.05/98.61	96.12/99.18
S5	82.0	92.8	93.5	63.41/95.5	89.2	93.0	83.0	88.4	54.9	61.11/63.8	72.2/96.3	58.5	97.04/98.3	91.28/93.83	95.05/94.97
S6	84.5	81.0	89.0	66.11/86.7	97.9	88.0	79.5	78.7	71.5	65.28/74.1	70.1/75.3	68.5	87.04/86.6	90.97/85.22	94.62/85.83
S7	75.0	77.5	81.5	59.57/84.7	82.1	82.0	74.5	77.5	81.3	61.11/61.9	64.6/72.2	78.6	92.14/89.6	81.38/86.57	91.22/86.57
S8	91.0	92.5	94.0	62.84/95.9	86.3	94.0	75.3	91.9	93.8	91.67/83.1	76.4/87.8	97.0	98.51/95.6	91.20/89.90	90.54/90.50
S9	87.0	87.2	90.5	84.46/92.6	97.1	91.0	73.3	83.4	93.8	86.11/77.2	77.1/85.3	93.9	92.31/87.4	83.95/86.05	97.50/87.73
AVG	**80**	**80**	**83**	**68.32/84.1**	**84.7**	**84**	**77.6**	**78**	**78.01**	**74.92/69.7**	**70.2/81.6**	**79.93**	**91.57/87.6**	**89.11/86.28**	**93.60/87.83**
S.D.	**1.3**	**1.5**	**1.6**	**1.3/1.5**	**1.4**	**1.3**	**0.9**	**1.6**	**1.9**	**1.7/1.6**	**0.7/1.4**	**1.7**	**0.6/0.9**	**0.4/0.8**	**0.3/0.7**

4.2- Properties of our collected dataset

There are several available BCI datasets [53, 54, 69]. However, we wanted to investigate several open questions in neuroscience and BCI that were outside the scope of the available datasets. So, we took the time and effort to collect our own dataset, which we are now sharing with the community. First, we wanted to test and directly compare the performance of our proposed attentional CNN on ME, MI, and their combination. In particular, we wanted to track the decoding accuracy over time via a sliding-window approach. We therefore increased the duration of the motor-imagination period from 2-3s to 4-6s to gain more insight and track the changes in decoding accuracy over time.

Second, BCI datasets typically instruct subjects to make trivial movements, such as pressing a button. We wanted to test our subjects on a less trivial paradigm, that requires them to exert some force. We therefore had our subjects squeeze a tennis ball (ME) or imagine doing that (MI). We expected this to make our classifier more robust against variety of MI tasks. This is vindicated by recent evidence that decoding attempted handwriting movements results in much higher accuracy than attempted typing [70].

Third, most of the BCI datasets for MI focused on electrodes above the motor region—such as C3, C4, and Cz [54]. We wanted to test to what degree general, high-density EEG recordings across the cortex (to the extent that those brain regions are accessible to EEG) contribute to the performance of an MI classifier. This also let us investigate the extent to which channel selection is useful in MI classification. Forth, an additional goal of our study was to evaluate the role of DA in MI classification. So, we needed a large enough dataset to be able to compare classification results when training our classifier on only a portion of the dataset. Altogether we recorded 400 trials pers subject (200 each for ME and MI, see Methods).

4.3- Motor Imagery vs. Motor Execution

MI could be described as kinesthetic anticipation of corresponding overt ME without producing an actual motor output. Jeannerod stated that MI is functionally equivalent to its ME counterpart [71]. More specifically, MI is related to the preparation of ME and represents meaningful neurophysiological dynamics of human motor functions [72]. Consequently, both MI and ME are accompanied by activation in common sensorimotor areas, such as the primary motor area (M1), supplementary motor area (SMA), and premotor cortex (PMC) [71, 72]. The neurophysiology underlying MI may differ in healthy people and patients with motor-impairing conditions [73]. MI-based BCI may further augment the motor learning process in healthy participants [74]. What is more, in patients with impaired motor functions, MI is often the only viable option to drive rehabilitative BCI, because these patients cannot perform overt ME [73]. The individuality and severity of motor impairments impact the underlying neurophysiology; for example, post-stroke neurophysiology relies on lesion locations [75]. Additional work is needed to further delineate the roles of MI and ME in motor learning or relearning for both healthy and impaired participants to refine the design of BCI for supplementing the motor learning process.

Our own dataset enables us to directly compare ME and MI within each participant. In our task, the participants were presented with the cue for 1 s, then saw a blank screen for 1 s, and finally began ME or MI for 4 s (see Methods). However, Dai et al. (2020), only used 2 s of MI. To better compare our results to theirs, we ran a sliding window analysis only for the first 2 s of the 4-s-long ME or MI period. We used window sizes of 100 ms, 300 ms, 500 ms, 1 s, and 2 s, with the step size fixed at 100 ms (see Figure 4 and Methods) on the data from all 64 channels. With this analysis, we would expect to see a rise in the accuracy leading up to the moment when the participants needed to begin ME or MI. Further, as participants were supposed to execute or imagine the movement for 4s, we expected the accuracy to then generally plateau over this after the above rise (similarly to Salvaris & Haggard, 2014 for example).

The left column in Figure 5 represents the average validation accuracy over all 7 participants and the right column is specifically for Participant 4. Both show the accuracy of the running-window analysis and over the first 4 s after cue onset for 3 analyses: ME only, MI only, and the combination of ME and MI trials. The window shown at the 4 s mark is from 3900 to 4000 ms for the 100 ms window, for 3700 to 4000 ms for the 300 ms window, and so on.

Our method's accuracy on ME is greater than on MI (Figure 5), which is consistent with previous findings about ME versus MI [76]. The average validation accuracy for the combination of MI and ME (All) is also greater than MI. Looking at the variability among the different window sizes, we see more variability in the ME condition than the MI or combined condition, on average. Our averaged results over all participants also align with our expectations, in that the accuracy rises from chance toward the beginning of the ME and MI periods and then generally plateaus (again, compare with Salvaris & Haggard, 2014).

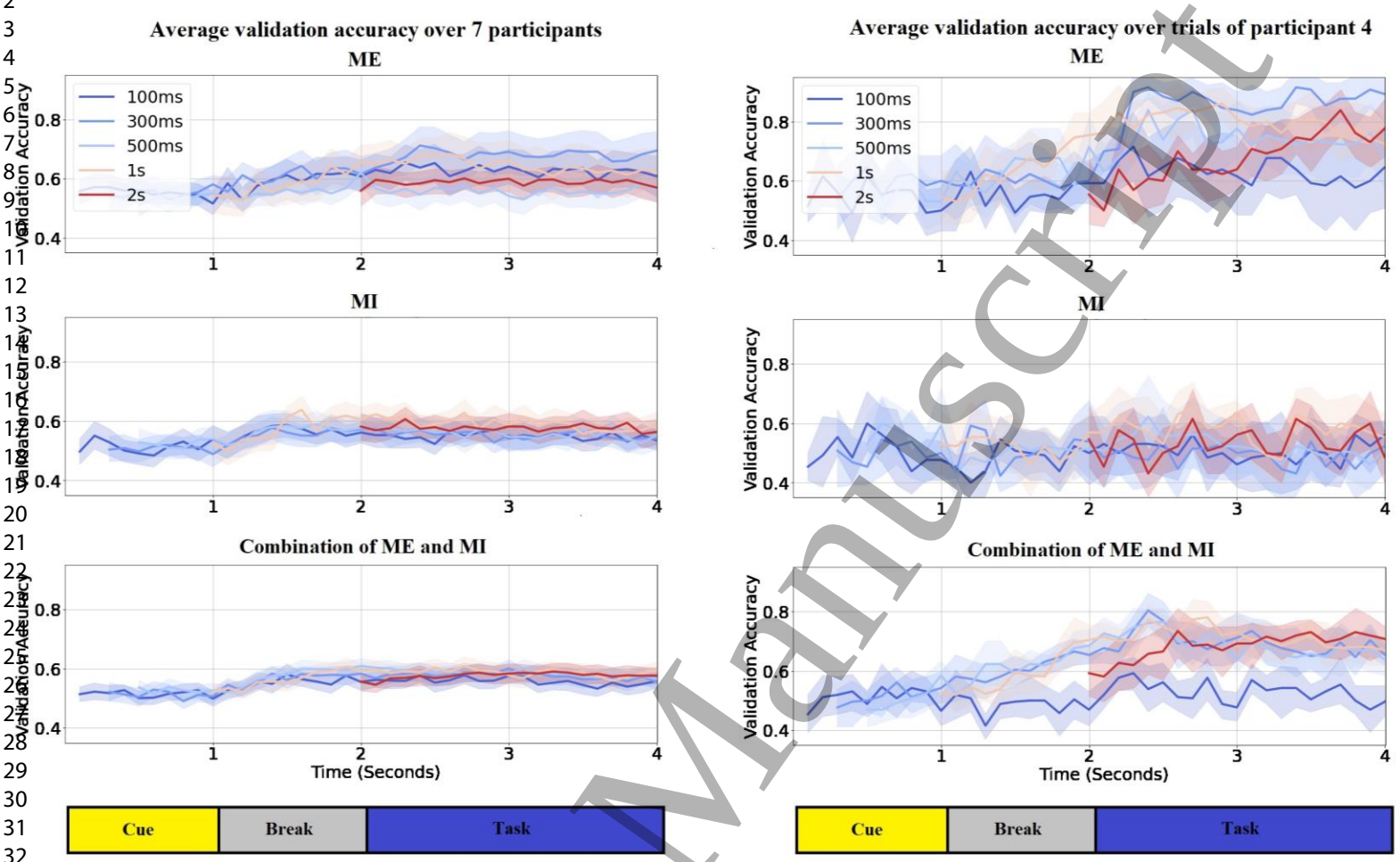


Figure 5. Validation accuracy of sliding-window analysis in ME (top), MI (middle), and ME and MI combined (bottom). The left column is the accuracy over time averaged across all 7 participants. The right column depicts the accuracy for the participant with the highest overall accuracy in the ME condition (Participant 4).

4.4- Channel selection

Analyzing dense-array EEG is computationally expensive and complex; it also typically requires more expensive EEG systems than those with sparser electrodes. Therefore, in this study we tested 4 different electrode configurations on our participants—which included 3, 7, 18, or all 64 electrodes (see Methods)—to further test the effect of channel selection on classification accuracy for MI in our own dataset.

The validation accuracy of the 7 participants for the 4 different channel-configurations are shown in Figure 6. In Table 8, the validation accuracy for each participant and the average accuracy across all participants are shown. The 18-channel layout had the highest accuracy, at 81.73% ($\pm$2.5).

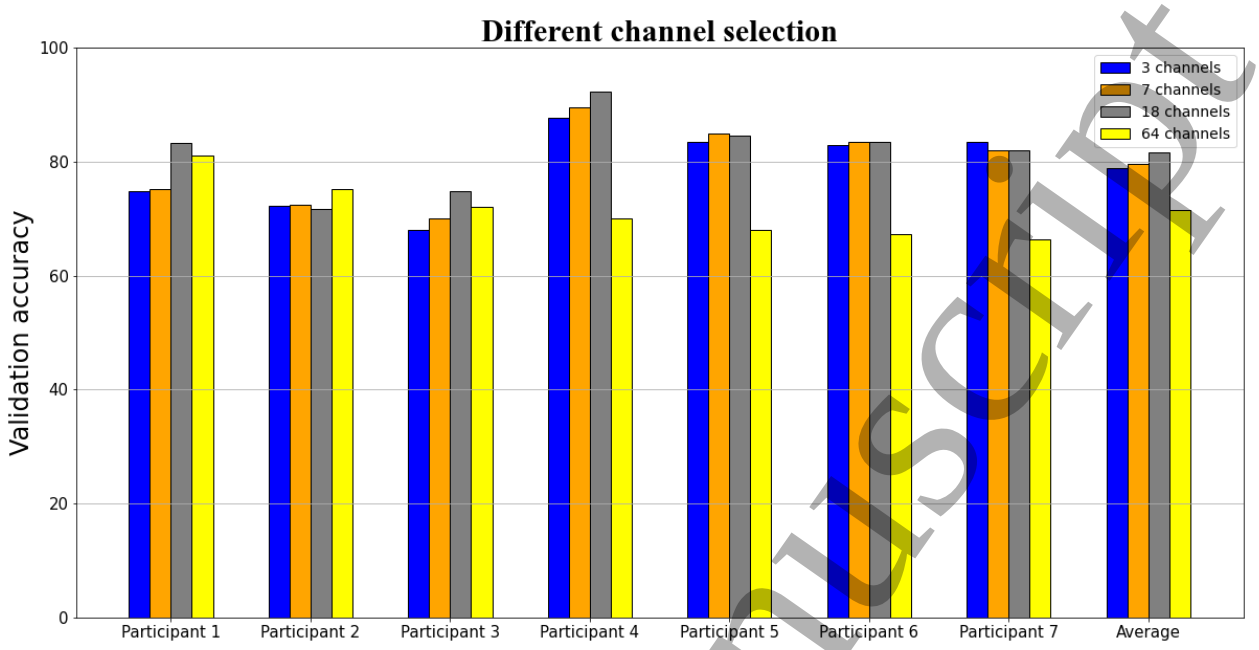


Figure 6. Validation accuracy for different channel configurations on the 7 participants of our dataset

Table 8. Validation accuracy for different channel selections on our dataset for single participants and the average over all participants. For each participant, we present mean ± SE over trials. In the bottom row, we present mean ± SE over participants.

Participant	3 channels	7 channels	18 channels	64 channels
1	74.75(±4.3)	75.25(±2.2)	83.25(±4.1)	81.18(±8.9)
2	72.25(±4.2)	72.50(±4.9)	71.75(±4.1)	75.22(±4.3)
3	68.01(±3.9)	70.01(±4.1)	74.75(±4.3)	72.05(±3.2)
4	87.69(±5.4)	89.62(±3.2)	92.31(±3.6)	70.03(±3.1)
5	83.50(±6.3)	85.01(±3.3)	84.50(±5.7)	68.08(±2.2)
6	83.00(±4.2)	83.50(±6.7)	83.51(±5.8)	67.33(±1.6)
7	83.50(±3.4)	82.01(±6.7)	82.01(±5.9)	66.41(±2.4)
AVG ($\pm S.E.$)	**78.95(±2.7)**	**79.70(±2.7)**	**81.73(±2.5)**	**71.47(±1.9)**

4.5- Data augmentation

We used 5 types of DA for the MI task: noise addition [44, 45], GAN [46-49], sliding window [30, 50, 51], Fourier transform [38], and recombination of segmentation [24]. Table 9 represents the result of different DA techniques on the BCI 2a, BCI 2b and our dataset for 64 channels and 18 channels. We evaluate all DA techniques with magnification factor $m = (2, 5, 10, 15, 20, 30, 50)$ for the proposed CNN. For Fourier transform, we used the same technique as in [38]. For noise addition, we opted for Gaussian noise with $\mu = 0, \sigma = (0.1, 0.2, 0.5)$.

cGANs allow generation based on a class assignment [43]. In this study, the GAN had 2 different conditions that were implemented: In order to provide context about the task, the first GAN model

generates a sample conditioned on the participant's decision—i.e., left vs. right. The second GAN model applies finer granularity by conditioning not only on left vs right but also the electrode channel. When generating data, the conditional inputs provide additional information and allow the model to tailor its outputs with greater detail (see Table 4). Figure 7 illustrates the architecture of cGAN in our work:

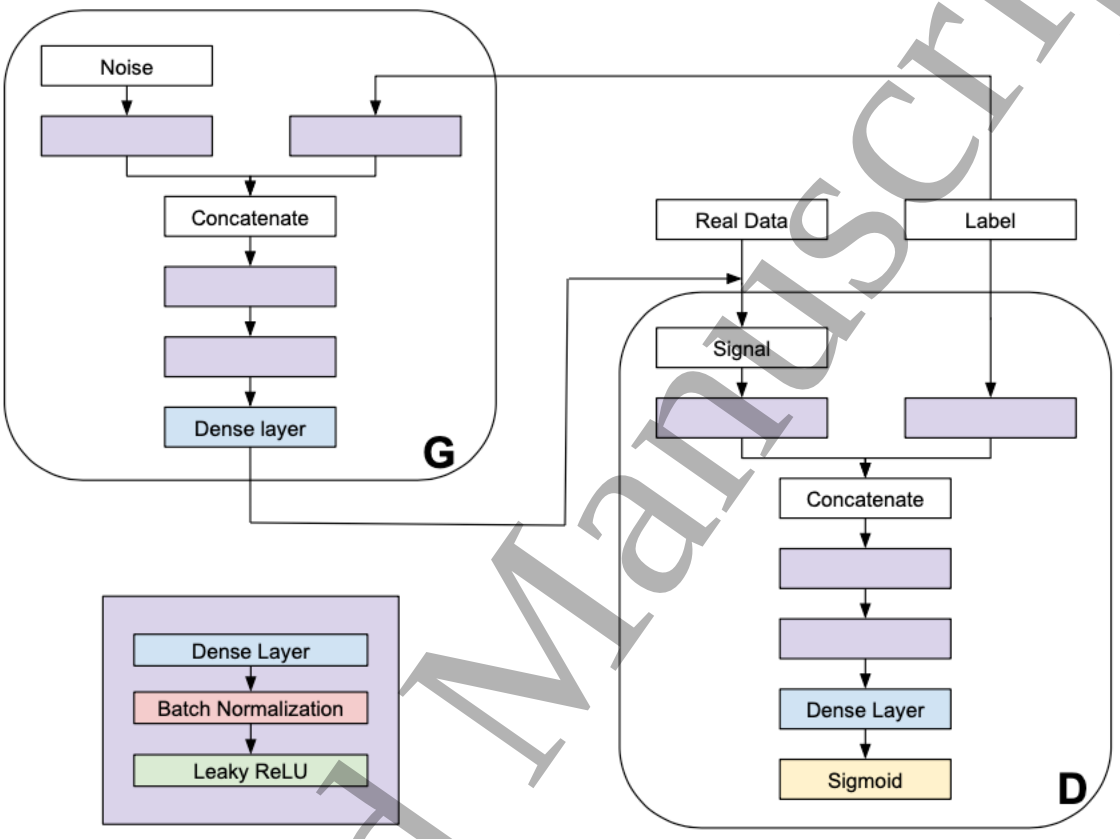


Figure 7 Our proposed cGAN model. In the generator (G), the prior input noise and label are combined into a hidden representation. In the discriminator (D), Real Data (i.e., raw EEG data) and the Label are presented as inputs to a discriminative function. The contents of all purple boxes in the architecture are the same and are expanded at the bottom left.

We also evaluated sliding-window technique (lengths $l = 1000\ ms$ with sampling frequency 250 Hz and step-size 100 ms). Table 9 demonstrated that GAN (conditional left vs. right and channels) with m=15 resulted in the best accuracy (93.6%) for BCI 2a dataset while Sliding Window (500 ms windows and 100 ms step size) with m=2 achieved the best accuracy (87.83%) for BCI 2b dataset. For our dataset, Fourier Transform with m=15 for 64 (86.61%) and 18 (83.42%) channels, respectively. The BCI 2a dataset had a magnification factor of 15 for the best result compared to a magnification factor of only 2 for BCI 2b. This might be because we did not include neurofeedback within our experimental paradigm. Decoding neurofeedback dataset has less complexity which is why BCI 2b dataset was seen to have a smaller magnification factor of 2. Our dataset did not include neurofeedback in the paradigm similarly to the BCI 2a dataset.

Table 9. Comparison of different DA techniques with different magnification factors and hyperparameters for BCI 2a, BCI 2b, and our experimental dataset (for 64 channels and 18 channels)

Dataset	DA techniques		Fourier-Transform	Noise Addition			GAN		Sliding Window
	parameter for each DA		(EMD)	σ =0.1	σ =0.2	σ =0.5	Conditional (left vs. right)	Conditional (left vs. right and channels)	Sliding window of length 1s (step-size: 100 ms)
BCI 2a	**Magnification factor**	2	0.8671	0.9056	0.8982	0.8768	0.9133	0.9025	0.8948
		5	0.8652	0.8999	0.8849	0.8908	0.9240	0.9092	0.8904
		10	0.8822	0.8902	0.8920	0.8721	0.9087	0.9217	0.8992
		15	0.8858	0.8988	0.8756	0.8750	0.9358	**0.9360**	0.8949
		20	0.8932	0.8898	0.8975	0.8904	0.9193	0.9300	0.9092
BCI 2b	**Magnification factor**	2	0.8535	0.8647	0.8614	0.8575	0.7939	0.8511	**0.8783**
		5	0.8391	0.8746	0.8696	0.8558	0.7747	0.8624	0.8747
		10	0.8339	0.8677	0.8668	0.8560	0.7733	0.8582	0.8726
		15	0.8228	0.8660	0.8717	0.8551	0.7601	0.8646	0.8749
		20	0.8217	0.8736	0.8677	0.8535	0.7611	0.8708	0.8691
Our dataset (64 channels)	**Magnification factor**	2	0.8442	0.8146	0.7548	0.7720	0.7914	0.8159	0.7904
		5	0.8305	0.7743	0.7844	0.7897	0.8377	0.7945	0.7933
		10	0.8377	0.7907	0.7885	0.7793	0.8024	0.8044	0.8033
		15	**0.8661**	0.7775	0.7541	0.7556	0.8184	0.7824	0.8362
		20	0.8560	0.7521	0.7826	0.7886	0.7994	0.8052	0.7990
Our dataset (18 channels)	**Magnification factor**	2	0.8124	0.8051	0.8056	0.8079	0.8045	0.8174	0.8190
		5	0.8010	0.8179	0.8121	0.8090	0.7969	0.8156	0.8224
		10	0.7988	0.8123	0.8162	0.8048	0.7965	0.8020	0.8312
		15	0.7954	0.8203	0.8141	0.8047	0.7842	0.8015	**0.8342**
		20	0.7963	0.8209	0.8051	0.8048	0.7875	0.8102	0.8277

4.6- Different portions of dataset

A dearth of data is a common problem when training machine-learning models on neuroimaging data. We therefore wanted to systematically test to what degree DA can compensate for the reduced availability of data. We thus randomly selected 100%, 75%, 50%, or 25% of the samples in our dataset. And we tested the accuracy of DA on these different proportions of our dataset for different DA techniques and magnification factors (Table 10). Fourier transform resulted in the best accuracy for 100%, 75%, and 50% of the data, with 86.61%, 88.26%, and 86.18% accuracy, under

magnification factors 15, 5, and 10, respectively. When using only 25% of the data, GAN (conditional left vs. right and channels) was the best DA technique in terms of accuracy, with 82.18% and a magnification factor of 15.

Table 10. Accuracies for different proportion of our dataset with different DA techniques

		DA techniques	**Fourier-Transform**	**Noise Addition**			**GAN**		**Sliding window**
Proportion of dataset		**Parameter for each DA**	**(EMD)**	**σ =0.1**	**σ =0.2**	**σ =0.5**	**Conditional (left vs. right)**	**Conditional (left vs. right and channels)**	**Sliding window of length 125**
100%	**Magnification factor**	2	0.8442	0.8146	0.7548	0.772	0.7914	0.8159	0.7904
		5	0.8305	0.7743	0.7844	0.7897	0.8377	0.7945	0.7933
		10	0.8377	0.7907	0.7885	0.7793	0.8024	0.8044	0.8033
		15	**0.8661**	0.7775	0.7541	0.7556	0.8184	0.7824	0.8362
		20	0.856	0.7521	0.7826	0.7886	0.7994	0.8052	0.799
75%	**Magnification factor**	2	0.8644	0.7975	0.7886	0.8129	0.7772	0.7927	0.7695
		5	**0.8826**	0.7856	0.7877	0.7987	0.7997	0.8045	0.7998
		10	0.8707	0.8096	0.7743	0.7921	0.804	0.795	0.798
		15	0.8732	0.7735	0.8013	0.7741	0.778	0.8057	0.8104
		20	0.8625	0.8066	0.7838	0.7814	0.8223	0.8159	0.8158
50%	**Magnification factor**	2	0.8346	0.8116	0.7957	0.7909	0.7743	0.756	0.7669
		5	0.8536	0.7672	0.7687	0.782	0.7754	0.8063	0.7656
		10	**0.8618**	0.8067	0.8222	0.7695	0.8034	0.7943	0.7503
		15	0.8474	0.8037	0.7969	0.7687	0.7671	0.8151	0.7426
		20	0.8128	0.756	0.801	0.7539	0.8247	0.8069	0.8039
25%	**Magnification factor**	2	0.7422	0.798	0.7868	0.8057	0.7595	0.7731	0.7387
		5	0.7683	0.8016	0.7569	0.7755	0.7714	0.7821	0.7202
		10	0.7417	0.7838	0.7767	0.8087	0.7643	0.8204	0.7256
		15	0.7909	0.7643	0.8187	0.7584	0.7737	**0.8218**	0.7138
		20	0.7826	0.7643	0.7814	0.7513	0.7731	0.7982	0.7501

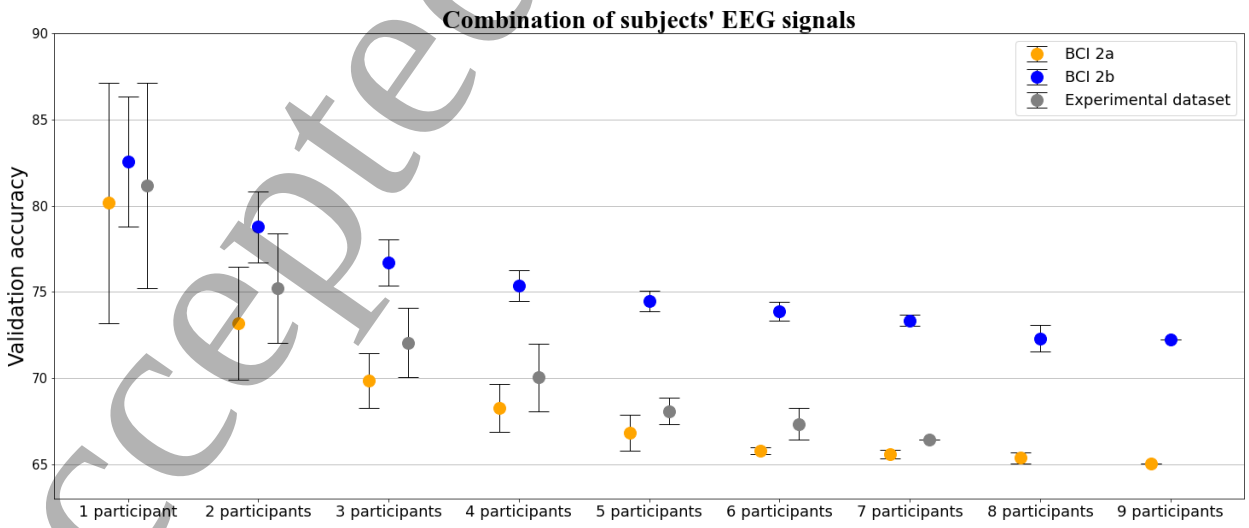
4.7- Combination of participants' EEG signals

The variability in brain anatomy and even more so functionality among different individuals is well known [e.g., 77]. Strong structure-function correspondences is therefore typically derived only at the aggregate level [78]. For example, Smith et al. delineated structural differences, suggesting that the number of folds and thickness of the cortex could be associated with whole-

brain functional network [79]. Furthermore, inter-participant variability in brain topography may also occurs due to participant-specific cognitive styles and the strategies that different participants use to perform the task [80]. This might augment the underlying learning processes—e.g., motor and perceptual learning [81]. Intra- and inter-participant variability might be explained by scale-dependent brain networks in spatial, temporal and topological domains [82].

Motor variability due to variability in human kinematic parameters—e.g., force field adaptation, speed and trajectory, and motivational factors such as level of user engagement, arousal and feelings of competence, necessary for performing a motor task—is an integral part of the motor learning process [83-85]. What is more, EEG signals are of course measured from the scalp rather than directly inside the brain, so they suffer from various signal distortions and technical limitations [86]. Given the above, the extent to which machine-learning models can be transferred between participants is not completely understood. The EEG patterns associated with motor variability could partly explain intra-individual variability in SMR-based BCI [87]. The neurophysiological processes underpinning the SMR often vary over time and across participants. Inherent intra- and inter-participant variability causes covariate shift in data distributions that impede the transferability of model parameters among sessions/participants.

Given the above, we evaluate the performance of the proposed NN on combinations of data across participants. The validation accuracy was averaged over every possible combination for each dataset—e.g., all participant pairs, all triplets, etc. After finding all the possible combinations, the data was split into training and test for each combination to compute the validation accuracy. The averages of the validation accuracy over all the states for the three datasets are reported in Figure 8 (Top) and differences between group (bottom). As we add more participants, the accuracy decreases—but the decreases become smaller. In Figure 8 (bottom), for the BCI 2a and 2b datasets, after combining 6 or more participants, we can see the curves plateau. This suggest that our proposed CNN was able to learn the important variations of the different EEG signals among the different subjects thus achieving stable accuracy.



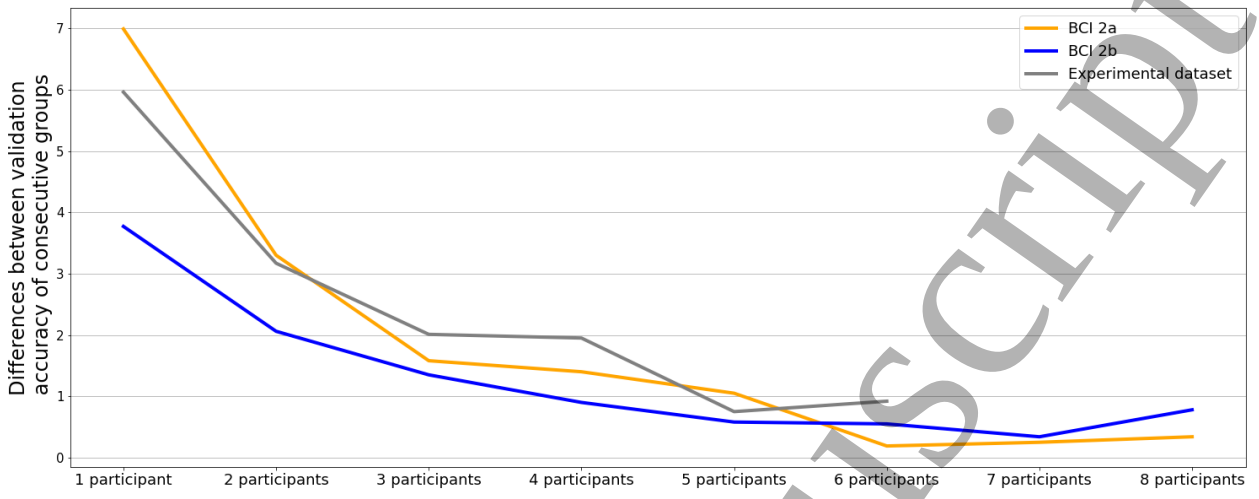


Figure 8. (Top)Validation accuracies for combinations of participants for BCI 2a, BCI 2b, and our experimental dataset. (Bottom) line plots of differences between mean validation accuracies of consecutive groups for the 3 datasets. The x axis labels are the smaller groups; so, differences between 2 participants and one are plotted above the label "1 participant", between 3 and 2 participants above "2 participants", and so on.

4.8- Leave-one-participant out and transfer learning

This subsection addresses two separates but closely related tasks. The first, leave-one-out, trains a NN on n-1 participants and tests on the remaining n^{th} participant. This task addresses the question of how information is shared between different participants' EEG signals (see section 3.7 Figure 6, on the x-axis, 8 participants for BCI 2a, BCI 2b and 6 participants for our dataset).

The second task, transfer learning, pretrains a NN on $n-1$ participants and fine-tunes to the n^{th} participant [88]. The pre-training phase orients the network weights to extract meaningful representations from the data. Then the fine-tuning, where the learning rate is decreased, adjusts to the task of interest, the n^{th} participant. For transfer learning, 10-fold cross validation over the n^{th} participant was used. Each fold fine-tunes on 9 folds and tests on the held-out 10^{th} fold. Table 11 shows the result of transfer learning on the BCI 2a, BCI 2b, and our dataset (64 channels and 18 channels). Figure 9 compared the result with and without transfer learning for all 3 datasets. For instance, the validation accuracy without transfer learning on participant n is defined by the trained model based on combination of the other $n-1$ participants and is tested on the complete dataset of participant 9. However, the validation accuracy with transfer learning on participant n is tuned to the trained model based on combination of the other $n-1$ participants based on 10% of the n^{th} participant and is tested on 90% of participant n.

Table 11. Leave-one-out and transfer-learning validation accuracy for BCI 2a, BCI 2b, and our dataset (64 and 18 channels)

Train (participants index)	Finetune (participant index)	BCI 2a (with transfer learning for different participants)	BCI 2b (with transfer learning for different participants)
2-3-4-5-6-7-8-9	1	78.12	78.75
3-4-5-6-7-8-9-1	2	76.38	71.62
4-5-6-7-8-9-1-2	3	89.53	79.17
5-6-7-8-9-1-2-3	4	77.77	97.02
6-7-8-9-1-2-3-4	5	77.41	83.10
7-8-9-1-2-3-4-5	6	78.83	81.94
8-9-1-2-3-4-5-6	7	80.58	81.67
9-1-2-3-4-5-6-7	8	81.60	87.36
1-2-3-4-5-6-7-8	9	90.63	84.44
Train (participants index)	**Finetune (participant index)**	**Our dataset, 64 channels**	**Our dataset, 18 channels**
2-3-4-5-7-6	1	83.25	83.75
3-4-5-6-7-1	2	73.01	87.25
4-5-6-7-1-2	3	76.50	77.50
5-6-7-1-2-3	4	91.15	92.70
6-7-1-2-3-4	5	91.01	85.50
1-2-3-4-5-7	6	82.10	82.50
1-2-3-4-5-6	7	84.50	86.50

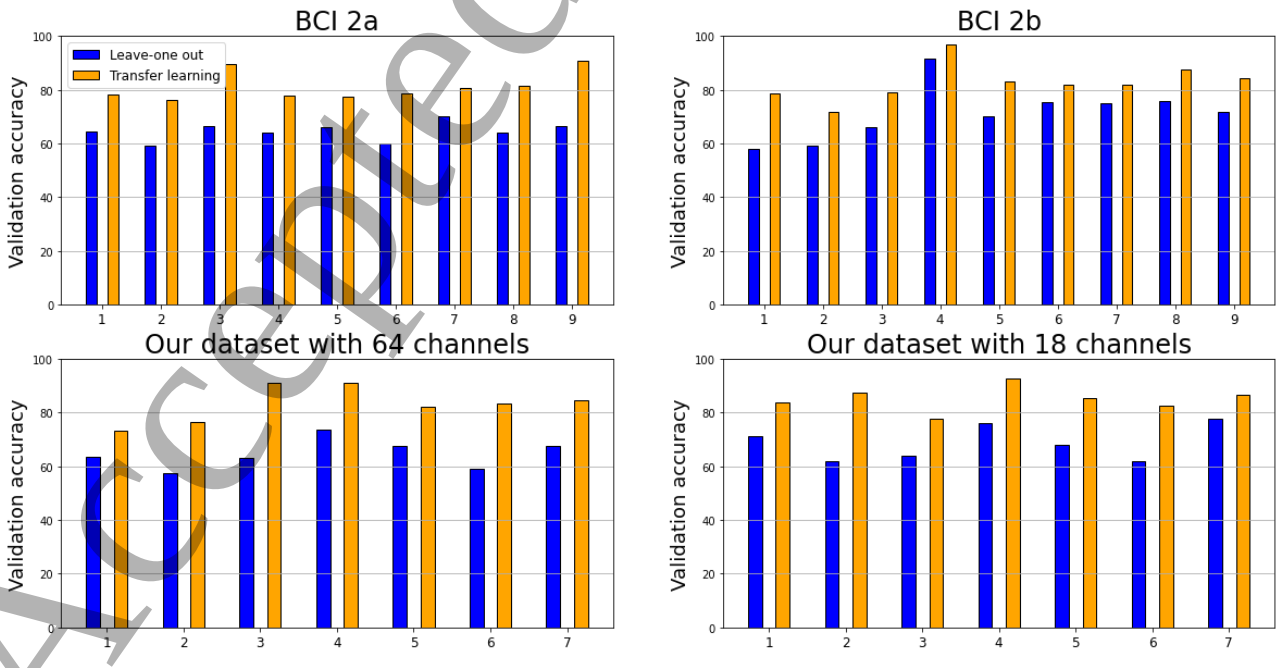


Figure 9. Validation accuracy for BCI 2a, BCI 2b, and our dataset (64 and 18 channels) with and without transfer learning

5. Discussion

In this study we proposed an end-to-end CNN architecture for EEG-based MI classification. This proposed mechanism is used to automatically extract features from raw EEG data (Figure 1 and Table 1). The NN optimization used the SHERPA Bayesian hyperparameter search on 3 datasets: the BCI Competition IV 2a and BCI Competition IV 2b, which have become benchmarks in the field, and a dataset that we collected ourselves (Figure 2; see Methods).

We began by comparing the architecture we favored, Conv Net-Attention-Dense NN, to two other baseline architectures—a Dense NN and Conv Net-Dense NN —as well as to what was, to the best of our knowledge, the top result in the field on the benchmark datasets—the architecture described in Dai et al. (2020) (see Figure 4). Our CNN-Attention-Dense achieved 93.6% (S.E.: □0.87) and 87.8% (S.E.: □2.11) accuracy over the BCI 2a and 2b datasets, respectively (Table 6). That is 6.4% to 13.5% and 4.03% to 5% better than the other architectures for BCI 2a and 2b, respectively (Figure 4). We further compared our results with all the papers we could find that classified the BCI 2a and 2b datasets and reported participant-by-participant results. For the BCI 2a dataset, our proposed EEG MI classification method achieved an improvement of 2.03% to 25.28% over all other methods (Table 7). For the BCI 2b dataset, our proposed method achieved an average improvement of 0.23% to 18.13% over previous methods (Table 7).

To the best of our knowledge, our CNN-Attention-Dense architecture achieved the highest accuracy thus far for the 2 benchmark datasets—BCI 2a and 2b. On top of that, an additional strength of our approach is its automated features extraction, directly from raw EEG. This contrasts with most methods, which tend to use handcrafted features and require heuristic parameter setting (e.g., predefined frequency bands). Automated features have the advantage of often generalizing better across tasks and participants [24]. Another potential advantage of our architecture is that the attentional mechanism could potentially lead to more interpretable results. However, we leave the explainable-AI facet of our architecture for further, future research.

The dataset that we collected for this study used 64 electrodes (according to the 10/20 montage; Figure 3). It included both ME and MI tasks and enabled us to compare the two tasks. Having all 3 datasets further enabled us to compare MI with and without neuro-feedback training (datasets 2b and 2a, respectively) as well as imagining button presses versus squeezing tennis balls (datasets 2a and 2b versus our own dataset, respectively).

A long-standing question in neuroscience and motor control is the extent of shared neural mechanisms between MI and ME [76]; though there is a general consensus that MI and ME at least share some important neural mechanisms. This similarity has been used in the MI-decoding literature, where some attempts to decode MI have relied on ME as training data [3]. Our results suggest that it is easier to decode ME than MI, at least when using EEG and relying on our decoding methods (Figure 5). Furthermore, we found that, on average, the decoding accuracy started at chance and then rose toward the time that participants were required to move or to imagine moving. After that it more or less plateaued. Interestingly, though perhaps not surprisingly, the accuracy level at the plateau, when using sliding windows, was lower than the accuracy for the full 4 s of ME (compare Figures 5 and 6). A likely contributing factor to this is that the sliding-window analysis decoded the EEG over shorter time windows than the full 4 s.

Another long-standing question when decoding EEG, and especially dense-array EEG, relates to how many and which electrodes (or channels) to use when recording the task. On the one hand, when using all channels (64, in our case), the set-up time for the task is longer, analyzing the larger dataset is more complex and computationally expensive, and brain signals unrelated to the task

and noise are perhaps more often introduced. On the other hand, using only a limited number of channels, there may not be full coverage of brain regions that may be involved in the decision-making and action-preparation processes. We therefore wanted to identify the appropriate channels relevant to the MI task. We thus selected different combinations of channels, according to 10-20 system standard, based on what is known about the neurophysiology of decision making and action formation, [3, 5, 89]. Hence we included different EEG configurations in our study (see Results), with 3, 7, or 18, channels around the motor cortex (see Methods), or with all 64 channels [57]. Our analysis suggests that, without DA, the 18-channels configuration had the best average accuracy (81.73±2.5), at least on our dataset (Figure 6, Table 8), while using all 64 channels resulted in the worse accuracy (71.47±1.9). Our results therefore suggest that, for MI decoding, it may be best to use only the 18 channels around the left and right motor region rather than all the channels. However, that result should be taken with a grain of salt, because when including DA, the tables were flipped, and it was the 64-channel configuration that did best, as described above.

One of the EEG configurations we tested included only 3 channels (C3, Cz, and C4)—this thus let us more directly compare our dataset to the two benchmark ones and the results of other studies. On those 3 electrodes, we achieved a mean accuracy of 79.95% for our dataset, while our analysis resulted in an accuracy of 89.11% and 86.28% for BCI 2a and 2b, respectively—all without DA. The higher accuracy for the benchmark datasets over our dataset might be due to the difference in tasks, the inclusion of neurofeedback (in BCI 2b), or that they perhaps ran participants who were better able to elicit good EEG data.

One general challenge of EEG decoding, especially with deep NNs, is obtaining enough data to train the numerous parameters in these large statistical models. The problem is compounded for MI tasks, because they are highly cognitively demanding. So, participants are easily fatigued and thus cannot produce a large amount of data in each experimental session. Bringing participants in for multiple sessions runs into issues of participant attrition for example. Another issue with collecting EEG over multiple session is the non-stationary nature of EEG signals [90]—i.e., the statistics of the EEG signals vary across time. As a result, a classifier trained at a specific time would tend to generalize increasingly poorly to data recorded at another time that was increasingly temporally removed—even for the same participant. This is a challenge for real-life applications of EEG, which must often work train on only limited amounts of data.

Some studies indeed strived for very lengthy data collection paradigms. One study, investigating MI control of 3D movement, had participants come back for up to 50 experimental sessions, which amounted to more than 20 hours of training per participant in some cases [91]. In another study, focusing on an EEG-based stroke-rehabilitation system [92], it took 12 weeks to collect enough data for three MI tasks, with each participant participating in 2 sessions per week [92]. While these are extreme examples, they highlight how common it is for participants to become fatigued after as little as 1 hour or less of data collection [93-95].

A promising solution to this dearth of data is to use DA, especially when using DL models on EEG data [33]. We therefore tested 5 disfferent DA techniques: sliding window, noise addition, GAN, Recombination of segmentation, and Fourier transform/wavelet. We further tested different magnification factors and hyperparameters (e.g., different window sizes for sliding window, various standard deviations for noise addition) for each technique we evaluated. Based on the guidelines in Lashgari et al. (2020) we evaluated the accuracy of the proposed method before and after DA. Our main objective was to find the best DA technique for each of the 3 datasets above. As far as we know, this is the first study to compare these various DA techniques as well as the different hyperparameters of the various techniques on benchmark datasets BCI 2a and 2b (see Table 9). We found that different techniques work best for different datasets. For BCI 2a, GAN

(conditional left vs. right and channels, m = 15) achieved the best accuracy, 93.6%. In contrast, sliding window (m = 2) gave the best accuracy for BCI 2b, at 87.83%. The DA step thus clearly boosted the performance of our proposed CNN (Table 6) as discussed below.

Interestingly, the BCI 2a dataset did not include neurofeedback training for the participants, while BCI 2b did. At the same time, the DA method that worked best for BCI 2a was a highly complex GAN with a large magnification factor, while that for BCI 2b it was a simple sliding window with a small magnification factor. So, one possible conclusion is that the neurofeedback training in BCI 2b, which effectively trained the participants to emit neural activity that would be better classified by the classifier, may have led to the superior accuracy from a simpler DA technique.

We also tested different DA techniques on our own dataset, which included 64 channels (see Methods). This achieved an accuracy of 86.61% (m = 15) with Fourier transform (Table 9). Using only 18 channels and the sliding-window DA technique (m = 15), we achieved an accuracy of 83.42%. Hence, using DA, we achieved higher accuracy with 64 channels than with 18 channels. Interestingly, without DA, the situation was flipped: the 64-channel data had lower accuracy 71.47(±1.9) than the 18-channel data 81.73(±2.5) (see Table 8). This suggests that, if one dataset has lower accuracy than another without DA, it does not necessarily mean that the first dataset would also have lower accuracy than the second after DA.

As noted above, our accuracies were higher than those of Dai et al. (2020) (Table 6)—which was the top result in the field. Besides higher accuracies on average, our accuracies for individual participants were 90.54% or higher (Table 6), while Dai et al. (2020) achieved this accuracy or higher for just for 5 of the 9 participants. Further, we were interested in the effect of DA on the accuracy of their results. But they did not report that for BCI 2a. And we were unable to obtain their code. What is more, they did not specify the details of their DA techniques. We therefore reimplemented their architecture from their paper, as per the details in their methods, without DA, to compare it with our architecture without DA. The accuracy of our proposed CNN without DA at 89.11% (±3.8; SE here and below) and 86.28% (±7.4) outperformed the NN reproduced from Dai et al. (2020) at 75.61% (±14.6) and 78.88% (±11.4) for BCI 2a and 2b datasets, respectively.

Following the above, an exciting potential use of DA is to replace lengthy, multi-session data-acquisition efforts [91, 92]. For brain-imaging studies, it would decrease the time and funds that researchers need to spend on data collection and reduce the inconvenience of participants. This is especially pertinent for situations where gathering additional data is financially, ethically, or otherwise difficult. Though DA would of course come at the expense of additional training time for the statistical models. We tested this by training on only some of the training set—25, 50, 75, or 100% (see Table 10)—while testing different DA techniques on the remaining data.

We therefore tested the extent to which data augmentation could replace gathering more data, at least for the dataset that we collected (Table 10). More specifically, we collected 400 trials from each participant (see Methods) and used different proportions of the MI dataset (100%, 75%, 50% and 25%) to train the model. We then augmented those different proportions of the dataset with various DA techniques that have different magnification factors. Our aim was to test the effects of those DA parameters on classification accuracy (Table 10). With 100%, 75% and 50% of the data, m =15, 5, and 10, using Fourier transform achieved the highest accuracies, that were overall similar, at 86.61%, 88.26%, and 86.18% accuracy. Yet, classification based on just 25% of the data, m=15 and GAN (conditional left vs. right and channels) resulted in a lower accuracy, 82.18%. It might be that the smaller dataset required a more sophisticated DA technique that for the other

proportions was needed to achieve its best accuracy. Though this accuracy was clearly lower than for the other proportions of data. This hints at the limits of DA for EEG.

It is well known that there is general anatomical similarity as well as structure-function correspondence among humans. But the anatomy of different brains also differs, at least to some extent, as does the structure-function correspondence. So, brain science typically operates at the aggregate level [78]. In particular, Smith et al. delineated structural differences, suggesting that the number of folds and thickness of the cortex could be associated with whole-brain functional networks [79]. Furthermore, inter-participant variability in topography occurs due to participant-specific cognitive style and strategy to perform a task over time [96], which could augment the underlying learning processes, e.g., motor and perceptual learning [81].

This question has clear implications for the analysis of EEG over groups of participants. We therefore wanted to investigate to what extent the number of participants over which we trained and tested our machine-learning model reduced the classification accuracy of the statistical model over that group. We thus trained and tested our model on all individual participants, on all pairs of participants, all triplets, quadruplets, and so on (Figure 8). It appears that, for all 3 datasets, the accuracy dropped most markedly between training and testing on individual participants to training and testing on pairs. Then there were diminishing decreases going from pairs to triplets, triplets to quadruplets, and so on, leading to roughly a plateau from groups of 6 participants and on. This suggests that the costs associated with inter-individual differences in brain structure and activity outweigh the benefits of the additional data when training over a group of participants. Though the decoding accuracy appeared to stop decreasing and reached somewhat of a plateau after around 6 participants. Future work, with a larger number of participants, could test the hypothesis that the accuracy would begin to rise again when training and testing over enough additional participants. One reason that this could happen is that the introduction of an ever-increasing number of additional participants might end up more than compensating for the neural variability between different brains. In other words, the advantages of the increasingly larger data available to train the model would outweigh the disadvantages of the variability across additional brains. Testing this hypothesis is left for future studies.

Following the discussion of inter-participant brain variability above, another key question in EEG analysis and especially for classification using DL is the extent to which a machine-learning model that was training on one group of participants could be generalized to new participants [97]. Put differently, we were wondering to what extent transfer learning, which has been increasingly used in the machine-learning literature, especially of late [98-100], would be useful for EEG classification using DL. We tested this by directly comparing two analyses. In the first, we trained a model on all but one participant and then tested it on that remaining participant (i.e., leave-one-participant-out classification). The second analysis comprised of again training on all but one participant, but then using transfer learning and finetuning the model on one part of the left-out participant. Finally, we tested the model on independent data from that participant (see Results 3.7). Our results clearly indicated that transfer learning led to higher accuracy than leave one out (Figure 9)—an increase in accuracy of 16.66%, 11.35%, and 18.6% for BCI 2a, BCI 2b, and for our dataset, respectively. This demonstrates the clear advantages of transfer learning for EEG analysis using DL. With DL models getting increasingly complex, the ability to finetune them for new participants rather than retrain them from scratch becomes increasingly important. In addition, our results suggest that the BCI community could use transfer learning with EEG to train a model on an existing dataset and then improve its performance for a new participant using only finetuning

of the model [98, 101]. According to our results, this could markedly improve the performance of BCI classifiers.

Due to the good classification performance of our proposed neural-network architecture and the relatively simple data processing, without prior manual feature extraction, our method holds promise for online, real-time, EEG-based classification of MI. It is left to future work to test how well the system will work in real time. Further, based on our results, it seems useful to use transfer learning between participants in a real-time paradigm. Furthermore, our neural-network architecture uses an attentional mechanism that helps identify the most salient brain regions that drive the network's classification ability. However, we leave the analysis of these brain regions for future work.

Acknowledgements

This publication was made possible in part through the support of a joint grant from the John Templeton Foundation and the Fetzer Institute, BIAL foundation and Boston Scientific Corporation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation or the Fetzer Institute. This publication was also made possible in part by the support of the Boston Scientific Investigator-Sponsored Research Program.
The authors would like to acknowledge the assistance of Natalie Nichols and Kai Lee Hague for graphics in this research.

Ethical statement

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We know of no conflicts of interest associated with this publication. The research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements. All subjects gave written, informed consent to participate in the study, which was approved by the Chapman University IRB (IRB-18-104). As Corresponding Author, I confirm that the manuscript has been read and approved for submission by all the named authors.

Reference

1. Abiri, R., et al., *A comprehensive review of EEG-based brain–computer interface paradigms.* Journal of neural engineering, 2019. **16**(1): p. 011001.
2. Nicolas-Alonso, L.F. and J. Gomez-Gil, *Brain computer interfaces, a review.* sensors, 2012. **12**(2): p. 1211-1279.
3. Salvaris, M. and P. Haggard, *Decoding intention at sensorimotor timescales.* PloS one, 2014. **9**(2): p. e85100.
4. Schneider, L., et al., *What we think before a voluntary movement.* Journal of cognitive neuroscience, 2013. **25**(6): p. 822-829.

5. Schultze-Kraft, M., et al., *The point of no return in vetoing self-initiated movements.* Proceedings of the National Academy of Sciences, 2016. **113**(4): p. 1080-1085.
6. Lashgari, E. and U. Maoz, *Electromyography Classification during Reach-to-Grasp Motion using Manifold Learning.* bioRxiv, 2020.
7. Lashgari, E., A. Pouya, and U. Maoz, *Decoding object weight from electromyography during human grasping.* BioRxiv, 2021.
8. Wolpaw, J.R., et al., *Brain–computer interfaces for communication and control.* Clinical neurophysiology, 2002. **113**(6): p. 767-791.
9. Daly, J.J. and J.R. Wolpaw, *Brain–computer interfaces in neurological rehabilitation.* The Lancet Neurology, 2008. **7**(11): p. 1032-1043.
10. Machado, S., L.F. Almada, and R.N. Annavarapu, *Progress and prospects in EEG-based brain-computer interface: clinical applications in neurorehabilitation.* Journal of Rehabilitation Robotics, 2013. **1**(1): p. 28-41.
11. Mulder, T., *Motor imagery and action observation: cognitive tools for rehabilitation.* Journal of neural transmission, 2007. **114**(10): p. 1265-1278.
12. Contreras-Vidal, J.L., et al., *Restoration of whole body movement: toward a noninvasive brain-machine interface system.* IEEE pulse, 2012. **3**(1): p. 34-37.
13. Pfurtscheller, G. and C. Neuper, *Motor imagery activates primary sensorimotor area in humans.* Neuroscience letters, 1997. **239**(2-3): p. 65-68.
14. He, B., et al., *Noninvasive brain-computer interfaces based on sensorimotor rhythms.* Proceedings of the IEEE, 2015. **103**(6): p. 907-925.
15. Yuan, H. and B. He, *Brain–computer interfaces using sensorimotor rhythms: current state and future perspectives.* IEEE Transactions on Biomedical Engineering, 2014. **61**(5): p. 1425-1435.
16. Morash, V., et al., *Classifying EEG signals preceding right hand, left hand, tongue, and right foot movements and motor imageries.* Clinical neurophysiology, 2008. **119**(11): p. 2570-2578.
17. Pfurtscheller, G. and F.L. Da Silva, *Event-related EEG/MEG synchronization and desynchronization: basic principles.* Clinical neurophysiology, 1999. **110**(11): p. 1842-1857.
18. Tabar, Y.R. and U. Halici, *A novel deep learning approach for classification of EEG motor imagery signals.* Journal of neural engineering, 2016. **14**(1): p. 016003.
19. Zabidi, A., et al. *Short-time Fourier Transform analysis of EEG signal generated during imagined writing*. in *2012 International Conference on System Engineering and Technology (ICSET)*. 2012. IEEE.
20. Amin, H.U., et al., *Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques.* Australasian physical & engineering sciences in medicine, 2015. **38**(1): p. 139-149.
21. Edelman, B.J., B. Baxter, and B. He, *EEG source imaging enhances the decoding of complex right-hand motor imagery tasks.* IEEE Transactions on Biomedical Engineering, 2015. **63**(1): p. 4-14.
22. Ang, K.K., et al. *Filter bank common spatial pattern (FBCSP) in brain-computer interface*. in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008. IEEE.
23. Meng, J., et al., *Noninvasive electroencephalogram based control of a robotic arm for reach and grasp tasks.* Scientific Reports, 2016. **6**: p. 38565.
24. Dai, G., et al., *HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification.* Journal of neural engineering, 2020. **17**(1): p. 016025.
25. Gandhi, T., B.K. Panigrahi, and S. Anand, *A comparative study of wavelet families for EEG signal classification.* Neurocomputing, 2011. **74**(17): p. 3051-3057.
26. Yang, Y., et al. *Time-frequency selection in two bipolar channels for improving the classification of motor imagery EEG*. in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2012. IEEE.

27. Baldi, P., *Deep Learning in Science: Theory, Algorithms, and Applications*. 2021, Cambridge University Press, Cambridge, UK.
28. Lawhern, V.J., et al., *EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces.* Journal of neural engineering, 2018. **15**(5): p. 056013.
29. Lu, N., et al., *A deep learning scheme for motor imagery classification based on restricted boltzmann machines.* IEEE transactions on neural systems and rehabilitation engineering, 2016. **25**(6): p. 566-576.
30. Schirrmeister, R.T., et al., *Deep learning with convolutional neural networks for EEG decoding and visualization.* Human brain mapping, 2017. **38**(11): p. 5391-5420.
31. Tortora, S., et al., *Deep learning-based BCI for gait decoding from EEG with LSTM recurrent neural network.* Journal of Neural Engineering, 2020. **17**(4): p. 046011.
32. Zhang, H., et al., *Motor imagery recognition with automatic EEG channel selection and deep learning.* Journal of Neural Engineering, 2021. **18**(1): p. 016004.
33. Lashgari, E., D. Liang, and U. Maoz, *Data Augmentation for Deep-Learning-Based Electroencephalography.* Journal of Neuroscience Methods, 2020: p. 108885.
34. Zhang, R., et al., *A novel hybrid deep learning scheme for four-class motor imagery classification.* Journal of neural engineering, 2019. **16**(6): p. 066004.
35. Zhang, G., et al., *Classification of hand movements from EEG using a deep attention-based LSTM network.* IEEE Sensors Journal, 2019. **20**(6): p. 3113-3122.
36. Wang, S., et al. *Training deep neural networks on imbalanced data sets*. in *2016 international joint conference on neural networks (IJCNN)*. 2016. IEEE.
37. Zhang, C., et al., *Understanding deep learning requires rethinking generalization.* arXiv preprint arXiv:1611.03530, 2016.
38. Zhang, Z., et al., *A novel deep learning approach with data augmentation to classify motor imagery signals.* IEEE Access, 2019. **7**: p. 15945-15954.
39. Vaswani, A., et al., *Attention is all you need.* arXiv preprint arXiv:1706.03762, 2017.
40. Shaw, P., J. Uszkoreit, and A. Vaswani, *Self-attention with relative position representations.* arXiv preprint arXiv:1803.02155, 2018.
41. Cisotto, G., et al., *Comparison of Attention-based Deep Learning Models for EEG Classification.* arXiv preprint arXiv:2012.01074, 2020.
42. Mrini, K., et al., *Rethinking self-attention: An interpretable self-attentive encoder-decoder parser.* 2019.
43. Hertel, L., et al., *Sherpa: Robust Hyperparameter Optimization for Machine Learning.* arXiv preprint arXiv:2005.04048, 2020.
44. Li, Y., et al., *A Channel-Projection Mixed-Scale Convolutional Neural Network for Motor Imagery EEG Decoding.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019. **27**(6): p. 1170-1180.
45. Parvan, M., et al. *Transfer Learning based Motor Imagery Classification using Convolutional Neural Networks*. in *2019 27th Iranian Conference on Electrical Engineering (ICEE)*. 2019. IEEE.
46. Hartmann, K.G., R.T. Schirrmeister, and T. Ball, *EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals.* arXiv preprint arXiv:1806.01875, 2018.
47. Yang, B., et al. *A Framework on Optimization Strategy for EEG Motor Imagery Recognition*. in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019. IEEE.
48. Zhang, Q. and Y. Liu, *Improving brain computer interface performance by data augmentation with conditional Deep Convolutional Generative Adversarial Networks.* arXiv preprint arXiv:1806.07108, 2018.
49. Zhang, X., et al. *Dada: Deep adversarial data augmentation for extremely low data regime classification*. in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019. IEEE.

50. Majidov, I. and T. Whangbo, *Efficient Classification of Motor Imagery Electroencephalography Signals Using Deep Learning Methods.* Sensors, 2019. **19**(7): p. 1736.
51. Tayeb, Z., et al., *Validating deep neural networks for online decoding of motor imagery movements from EEG signals.* Sensors, 2019. **19**(1): p. 210.
52. Mirza, M. and S. Osindero, *Conditional generative adversarial nets.* arXiv preprint arXiv:1411.1784, 2014.
53. Brunner, C., et al., *BCI Competition 2008–Graz data set A.* Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology, 2008. **16**: p. 1-6.
54. Leeb, R., et al., *BCI Competition 2008–Graz data set B.* Graz University of Technology, Austria, 2008: p. 1-6.
55. Gaur, P., et al. *An empirical mode decomposition based filtering method for classification of motor-imagery EEG signals for enhancing brain-computer interface*. in *2015 International Joint Conference on Neural Networks (IJCNN)*. 2015. IEEE.
56. Luo, J., et al., *Dynamic frequency feature selection based approach for classification of motor imageries.* Computers in biology and medicine, 2016. **75**: p. 45-53.
57. Jurcak, V., D. Tsuzuki, and I. Dan, *10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems.* Neuroimage, 2007. **34**(4): p. 1600-1611.
58. Kevric, J. and A. Subasi, *Comparison of signal decomposition methods in classification of EEG signals for motor-imagery BCI system.* Biomedical Signal Processing and Control, 2017. **31**: p. 398-406.
59. Sadiq, M.T., et al., *Motor imagery EEG signals classification based on mode amplitude and frequency components using empirical wavelet transform.* IEEE Access, 2019. **7**: p. 127678-127692.
60. Sadiq, M.T., et al., *Motor Imagery EEG Signals Decoding by Multivariate Empirical Wavelet Transform-Based Framework for Robust Brain–Computer Interfaces.* IEEE Access, 2019. **7**: p. 171431-171451.
61. Shahid, S. and G. Prasad, *Bispectrum-based feature extraction technique for devising a practical brain–computer interface*. Journal of neural engineering, 2011. **8**(2): p. 025014.
62. Ang, K.K., et al., *Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b.* Frontiers in neuroscience, 2012. **6**: p. 39.
63. Saa, J.F.D. and M. Çetin, *A latent discriminative model-based approach for classification of imaginary motor tasks from EEG data.* Journal of neural engineering, 2012. **9**(2): p. 026020.
64. Li, M.-a., et al., *Adaptive feature extraction of motor imagery EEG with optimal wavelet packets and SE-isomap.* Applied Sciences, 2017. **7**(4): p. 390.
65. Zheng, Q., F. Zhu, and P.-A. Heng, *Robust support matrix machine for single trial EEG classification*. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2018. **26**(3): p. 551-562.
66. Lotte, F. and C. Guan, *Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms.* IEEE Transactions on biomedical Engineering, 2010. **58**(2): p. 355-362.
67. Raza, H., et al., *Adaptive learning with covariate shift-detection for motor imagery-based brain–computer interface.* Soft Computing, 2016. **20**(8): p. 3085-3096.
68. Gaur, P., et al., *A multi-class EEG-based BCI classification using multivariate empirical mode decomposition based filtering and Riemannian geometry.* Expert Systems with Applications, 2018. **95**: p. 201-211.
69. Blankertz, B., et al., *The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects.* NeuroImage, 2007. **37**(2): p. 539-550.
70. Willett, F.R., et al., *High-performance brain-to-text communication via handwriting.* Nature, 2021. **593**(7858): p. 249-254.

71. Jeannerod, M., *Mental imagery in the motor context.* Neuropsychologia, 1995. **33**(11): p. 1419-1432.
72. Zich, C., et al., *Real-time EEG feedback during simultaneous EEG–fMRI identifies the cortical signature of motor imagery.* Neuroimage, 2015. **114**: p. 438-447.
73. Lotze, M., et al., *Phantom movements and pain An fMRI study in upper limb amputees.* Brain, 2001. **124**(11): p. 2268-2277.
74. Ruffino, C., C. Papaxanthis, and F. Lebon, *Neural plasticity during motor learning with motor imagery practice: Review and perspectives.* Neuroscience, 2017. **341**: p. 61-78.
75. Niazi, I.K., et al., *Detection of movement-related cortical potentials based on subject-independent training.* Medical & biological engineering & computing, 2013. **51**(5): p. 507-512.
76. de Lange, F.P., *Neural mechanisms of motor imagery*. 2008.
77. Mueller, S., et al., *Individual variability in functional connectivity architecture of the human brain.* Neuron, 2013. **77**(3): p. 586-595.
78. Honey, C.J., J.-P. Thivierge, and O. Sporns, *Can structure predict function in the human brain?* Neuroimage, 2010. **52**(3): p. 766-776.
79. Smith, S., et al., *Structural variability in the human brain reflects fine-grained functional architecture at the population level.* Journal of Neuroscience, 2019. **39**(31): p. 6136-6149.
80. Quinn, A.J., et al., *Task-evoked dynamic network analysis through hidden markov modeling.* Frontiers in neuroscience, 2018. **12**: p. 603.
81. Herzfeld, D.J. and R. Shadmehr, *Motor variability is not noise, but grist for the learning mill.* nature neuroscience, 2014. **17**(2): p. 149-150.
82. Betzel, R.F., et al., *The community structure of functional brain networks exhibits scale-specific patterns of inter-and intra-subject variability.* Neuroimage, 2019. **202**: p. 115990.
83. Edelman, B.J., et al., *Noninvasive neuroimaging enhances continuous neural tracking for robotic device control.* Science robotics, 2019. **4**(31).
84. Faller, J., et al., *Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task.* Proceedings of the National Academy of Sciences, 2019. **116**(13): p. 6482-6490.
85. Saha, S. and M. Baumert, *Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review.* Frontiers in Computational Neuroscience, 2019. **13**: p. 87.
86. Luck, S.J., *An introduction to the event-related potential technique*. 2014: MIT press.
87. Ostry, D.J. and P.L. Gribble, *Sensory plasticity in human motor learning.* Trends in neurosciences, 2016. **39**(2): p. 114-123.
88. Radford, A., et al., *Improving language understanding by generative pre-training.* 2018.
89. Bai, O., et al., *Prediction of human voluntary movement before it occurs.* Clinical Neurophysiology, 2011. **122**(2): p. 364-372.
90. Craik, A., Y. He, and J.L. Contreras-Vidal, *Deep learning for electroencephalogram (EEG) classification tasks: a review.* Journal of neural engineering, 2019. **16**(3): p. 031001.
91. McFarland, D.J., W.A. Sarnacki, and J.R. Wolpaw, *Electroencephalographic (EEG) control of three-dimensional movement.* Journal of neural engineering, 2010. **7**(3): p. 036007.
92. Suwannarat, A., S. Pan-ngum, and P. Israsena, *Comparison of EEG measurement of upper limb movement in motor imagery training system.* Biomedical engineering online, 2018. **17**(1): p. 103.
93. Ang, K.K., et al., *A randomized controlled trial of EEG-based motor imagery brain-computer interface robotic rehabilitation for stroke.* Clinical EEG and neuroscience, 2015. **46**(4): p. 310-320.
94. Koo, B., et al., *A hybrid NIRS-EEG system for self-paced brain computer interface with online motor imagery.* Journal of neuroscience methods, 2015. **244**: p. 26-32.
95. Tam, W.-K., et al., *A minimal set of electrodes for motor imagery BCI to control an assistive device in chronic stroke subjects: a multi-session study.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2011. **19**(6): p. 617-627.

96. Seghier, M.L. and C.J. Price, *Interpreting and utilising intersubject variability in brain function.* Trends in Cognitive Sciences, 2018. **22**(6): p. 517-530.
97. Saha, S. and M. Baumert, *Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review.* Frontiers in computational neuroscience, 2020. **13**: p. 87.
98. Lee, D.-Y., et al. *Decoding movement imagination and execution from eeg signals using bci-transfer learning method based on relation network*. in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020. IEEE.
99. Wu, D., Y. Xu, and B.-L. Lu, *Transfer learning for EEG-based brain-computer interfaces: A review of progress made since 2016.* IEEE Transactions on Cognitive and Developmental Systems, 2020.
100. Zhang, K., et al., *Adaptive transfer learning for EEG motor imagery classification with deep Convolutional Neural Network.* Neural Networks, 2021. **136**: p. 1-10.
101. Fahimi, F., et al., *Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI.* Journal of neural engineering, 2019. **16**(2): p. 026007.