PAPER

EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces

IOP Publishing Journal of Neural Engineering
J. Neural Eng. **15** (2018) 056013 (17pp) https://doi.org/10.1088/1741-2552/aace8c

EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces

Vernon J Lawhern[1,5], Amelia J Solon[1,2], Nicholas R Waytowich[1,3], Stephen M Gordon[1,2], Chou P Hung[1,4] and Brent J Lance[1]

[1] Human Research and Engineering Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, United States of America
[2] DCS Corporation, Alexandria, VA, United States of America
[3] Department of Biomedical Engineering, Columbia University, New York, NY, United States of America
[4] Department of Neuroscience, Georgetown University, Washington, DC, United States of America

E-mail: vernon.j.lawhern.civ@mail.mil

Received 9 May 2017, revised 20 June 2018
Accepted for publication 22 June 2018
Published 27 July 2018

Abstract

Objective. Brain–computer interfaces (BCI) enable direct communication with a computer, using neural activity as the control signal. This neural signal is generally chosen from a variety of well-studied electroencephalogram (EEG) signals. For a given BCI paradigm, feature extractors and classifiers are tailored to the distinct characteristics of its expected EEG control signal, limiting its application to that specific signal. Convolutional neural networks (CNNs), which have been used in computer vision and speech recognition to perform automatic feature extraction and classification, have successfully been applied to EEG-based BCIs; however, they have mainly been applied to single BCI paradigms and thus it remains unclear how these architectures generalize to other paradigms. Here, we ask if we can design a single CNN architecture to accurately classify EEG signals from different BCI paradigms, while simultaneously being as compact as possible. *Approach*. In this work we introduce EEGNet, a compact convolutional neural network for EEG-based BCIs. We introduce the use of depthwise and separable convolutions to construct an EEG-specific model which encapsulates well-known EEG feature extraction concepts for BCI. We compare EEGNet, both for within-subject and cross-subject classification, to current state-of-the-art approaches across four BCI paradigms: P300 visual-evoked potentials, error-related negativity responses (ERN), movement-related cortical potentials (MRCP), and sensory motor rhythms (SMR). *Main results*. We show that EEGNet generalizes across paradigms better than, and achieves comparably high performance to, the reference algorithms when only limited training data is available across all tested paradigms. In addition, we demonstrate three different approaches to visualize the contents of a trained EEGNet model to enable interpretation of the learned features. *Significance*. Our results suggest that EEGNet is robust enough to learn a wide variety of interpretable features over a range of BCI tasks. Our models can be found at: https://github.com/vlawhern/arl-eegmodels.

Keywords: EEG, convolutional neural network, brain–computer interface, deep learning, event-related potential, sensory motor rhythm

(Some figures may appear in colour only in the online journal)

[5] Author to whom any correspondence should be addressed.

1. Introduction

A brain–computer interface (BCI) enables direct communication with a machine via brain signals [1]. Traditionally, BCIs have been used for medical applications such as neural control of prosthetic artificial limbs [2]. However, recent research has opened up the possibility for novel BCIs focused on enhancing performance of healthy users, often with noninvasive approaches based on electroencephalography (EEG) [3–5]. Generally speaking, a BCI consists of five main processing stages [6]: a data collection stage, where neural data is recorded; a signal processing stage, where the recorded data is preprocessed and cleaned; a feature extraction stage, where meaningful information is extracted from the neural data; a classification stage, where a decision is interpreted from the data; and a feedback stage, where the result of that decision is provided to the user. While these stages are largely the same across BCI paradigms, each paradigm relies on manual specification of signal processing [7], feature extraction [8] and classification methods [9], a process which often requires significant subject-matter expertise and/or *a priori* knowledge about the expected EEG signal. It is also possible that, because the EEG signal preprocessing steps are often very specific to the EEG feature of interest (for example, band-pass filtering to a specific frequency range), that other potentially relevant EEG features could be excluded from analysis (for example, features outside of the band-pass frequency range). The need for robust feature extraction techniques will only continue to increase as BCI technologies evolve into new application domains [3–5, 10–12].

Deep learning has largely alleviated the need for manual feature extraction, achieving state-of-the-art performance in fields such as computer vision and speech recognition [13, 14]. Specifically, the use of deep convolutional neural networks (CNNs) has grown due in part to their success in many challenging image classification problems [15–19], surpassing methods relying on hand-crafted features (see [14] and [20] for recent reviews). Although the majority of BCI systems still rely on the use of handcrafted features, many recent works have explored the application of deep learning to EEG signals. For example, CNNs have been used for epilepsy prediction and monitoring [21–25], for auditory music retrieval [26, 27], for detection of visual-evoked responses [28–31] and for motor imagery classification [32, 33], while deep belief networks (DBNs) have been used for sleep stage detection [34], anomaly detection [35] and in motion-onset visual-evoked potential classification [36]. CNNs using time-frequency transforms of EEG data were used for mental workload classification [37] and for motor imagery classification [38–40]). Restricted Boltzman machines (RBMs) have been used for motor imagery [41]. An adaptive method based on stacked denoising autoencoders has been proposed for mental workload classification [42]). These studies focused primarily on classification in a single BCI task, often using task-specific knowledge in designing the network architecture. In addition, the amount of data used to train these networks varied significantly across studies, in part due to the difficulty in collecting data under different experimental designs. Thus, it remains unclear how these previous deep learning approaches would generalize both to other BCI tasks as well as to variable training data sizes.

In this work we introduce *EEGNet*, a compact CNN for classification and interpretation of EEG-based BCIs. We introduce the use of *depthwise* and *separable* convolutions, previously used in computer vision [43], to construct an EEG-specific network that encapsulates several well-known EEG feature extraction concepts, such as optimal spatial filtering and filter-bank construction, while simultaneously reducing the number of trainable parameters to fit when compared to existing approaches. We evaluate the generalizability of EEGNet on EEG datasets collected from four different BCI paradigms: P300 visual-evoked potential (P300), error-related negativity (ERN), movement-related cortical potential (MRCP) and the sensory motor rhythm (SMR), representing a spectrum of paradigms based on classification of event-related potentials (P300, ERN, MRCP) as well as classification of oscillatory components (SMR). In addition, each of these data collections contained varying amounts of data, allowing us to explore the efficacy of EEGNet on various training data sizes. Our results are as follows: We show that EEGNet achieves improved classification performance over an existing paradigm-agnostic EEG CNN model across nearly all tested paradigms when limited training data is available. In addition, we show that EEGNet effectively generalizes across all tested paradigms. We also show that EEGNet performs just as well as a more paradigm-specific EEG CNN model, but with two orders of magnitude fewer parameters to fit, representing a more efficient use of model parameters (an aspect that has been explored in previous deep learning literature, see [43, 44]). Finally, through the use of feature visualization and model ablation analysis, we show that neurophysiologically interpretable features can be extracted from the EEGNet model. This is important as CNNs, despite their ability for robust and automatic feature extraction, often produce hard to interpret features. For neuroscience practitioners, the ability to derive insights into CNN-derived neurophysiological phenomena may be just as important as achieving good classification performance, depending on the intended application. We validate our architecture's ability to extract neurophysiologically interpretable signals on several well-studied BCI paradigms.

The remainder of this manuscript is structured as follows. Section 2.1 gives a brief description of the four datasets used to validate our CNN model. Section 2.2 describes our EEGNet model as well as other BCI models (both CNN and non-CNN based models) used in our model comparison. Section 3 presents the results of both within-subject and cross-subject classification performance, as well as results of our feature explainability analysis. We discuss our findings in more detail in the discussion.

2. Materials and methods

2.1. Data description

BCIs are generally categorized into two types, depending on the EEG feature of interest [45]: event-related and oscillatory. *Event-related potential* (ERP) BCIs are designed to

Table 1. Summary of the data collections used in this study. Class imbalance, if present, is given as odds; i.e.: an odds of 2:1 means the class imbalance is 2/3 of the data for class 1 to 1/3 of the data for class 2. For the P300 and ERN datasets, the class imbalance is subject-dependent; therefore, the odds is given as the average class imbalance over all subjects.

Paradigm	Feature type	Bandpass filter (Hz)	# of Subjects	Trials per subject	# of Classes	Class imbalance?
P300	ERP	1–40	15	∼2000	2	Yes, ∼5.6:1
ERN	ERP	1–40	26	340	2	Yes, ∼3.4:1
MRCP	ERP/Oscillatory	0.1–40	13	∼1100	2	No
SMR	Oscillatory	4–40	9	288	4	No

detect a high amplitude and low frequency EEG response to a known, time-locked external stimulus. They are generally robust across subjects and contain well-stereotyped waveforms, enabling the time course of the ERP to be modeled through machine learning efficiently [46]. In contrast to ERP-based BCIs, which rely mainly on the detection of the ERP waveform from some external event or stimulus, *oscillatory* BCIs use the signal power of specific EEG frequency bands for external control and are generally asynchronous [47]. When oscillatory signals are time-locked to an external stimulus, they can be represented through event-related spectral perturbation (ERSP) analyses [48]. Oscillatory BCIs are more difficult to train, generally due to the lower signal-to-noise ratio (SNR) as well as greater variation across subjects [47]. A summary of the data used in this manuscript can be found in table 1

2.1.1. Dataset 1: P300 event-related potential (P300). The P300 event-related potential is a stereotyped neural response to novel visual stimuli [49]. It is commonly elicited with the visual oddball paradigm, where participants are shown repetitive 'non-target' visual stimuli that are interspersed with infrequent 'target' stimuli at a fixed presentation rate (for example, 1 Hz). Observed over the parietal cortex, the P300 waveform is a large positive deflection of electrical activity observed approximately 300 ms post stimulus onset, the strength of the observed deflection being inversely proportional to the frequency of the target stimuli. The P300 ERP is one of the strongest neural signatures observable by EEG, especially when targets are presented infrequently [49]. When the image presentation rate increases to 2 Hz or more, it is commonly referred to as rapid serial visual presentation (RSVP), which has been used to develop BCIs for large image database triage [50–52].

The EEG data used here have been previously described in [51]; a brief description is given below. 18 participants volunteered for an RSVP BCI study. Participants were shown images of natural scenery at 2 Hz rate, with images either containing a vehicle or person (target), or with no vehicle or person present (non-target). Participants were instructed to press a button with their dominant hand when a target image was shown. The target/non-target ratio was 20%/80%. Data from three participants were excluded from the analysis due to excessive artifacts and/or noise within the EEG data. Data from the remaining 15 participants (9 male and 14 right-handed) who ranged in age from 18 to 57 years (mean age 39.5 years) were further analyzed. EEG recordings were digitally sampled at 512 Hz from 64 scalp electrodes arranged in a 10–10 montage using a BioSemi active two system (Amsterdam, The Netherlands). Continuous EEG data were referenced offline to the average of the left and right earlobes, digitally bandpass filtered, using an FIR filter implemented in EEGLAB [53], to 1–40 Hz and downsampled to 128 Hz. EEG trials of target and non-target conditions were extracted at $[0, 1]$ s post stimulus onset, and used for a two-class classification.

2.1.2. Dataset 2: Feedback error-related negativity (ERN). Error-related negativity potentials are perturbations of the EEG following an erroneous or unusual event in the subject's environment or task. They can be observed in a variety of tasks, including time interval production paradigms [54] and in forced-choice paradigms [55, 56]. Here we focus on the feedback error-related negativity (ERN), which is an amplitude perturbation of the EEG following the perception of an erroneous feedback produced by a BCI. The feedback ERN is characterized as a negative error component approximately 350 ms, followed by a positive component approximately 500 ms, after visual feedback (see figure 7 of [57] for an illustration). The detection of the feedback ERN provides a mechanism to infer, and to possibly correct in real-time, the incorrect output of a BCI. This two-stage system has been proposed as a hybrid BCI in [58, 59] and has been shown to improve the performance of a P300 speller in online applications [60].

The EEG data used here comes from [57] and was used in the 'BCI Challenge' hosted by Kaggle (www.kaggle.com/c/inria-bci-challenge); a brief description is given below. 26 healthy participants (16 for training, 10 for testing) participated in a P300 speller task, a system which uses a random sequence of flashing letters, arranged in a 6×6 grid, to elicit the P300 response [61]. The goal of the challenge was to determine whether the feedback of the P300 speller was correct or incorrect. The EEG data were originally recorded at 600 Hz using 56 passive Ag/AgCl EEG sensors (VSM-CTF compatible system) following the extended 10–20 system for electrode placement. Prior to our analysis, the EEG data were band-pass filtered, using an FIR filter implemented in EEGLAB [53], to 1–40 Hz and down-sampled to 128 Hz. EEG trials of correct and incorrect feedback were extracted at $[0, 1.25]$ s post feedback presentation and used as features for a two-class classification.

2.1.3. Dataset 3: Movement-related cortical potential (MRCP). Some neural activities contain both ERP as well as an oscillatory components. One particular example of this is the movement-related cortical potential (MRCP), which can be elicited by voluntary movements of the hands and feet and is observable through EEG along the central and midline

electrodes, contralateral to the hand or foot movement [62–65]. The MRCP components can be seen before movement onset (a slow 0–5 Hz readiness potential [66, 67] and an early desynchronization in the 10–12 Hz frequency band), at movement onset (a slow motor potential [67, 68]), and after movement onset (a late synchronization of 20–30 Hz activity approximately 1 s after movement execution). The MRCP has been used previously to develop motor control BCIs for both healthy and physically disabled patients [69–71]

The EEG data used here have been previously described in [72]; a brief description is given below. In this study, 13 subjects performed self-paced finger movements using the left index, left middle, right index, or right middle fingers. The data was recorded using a 256 channel BioSemi Active II system at 1024 Hz. Due to extensive signal noise present in the data, the EEG data were first processed with the PREP pipeline [73]. The data were referenced to linked mastoids, bandpass filtered, using an FIR filter implemented in EEGLAB [53], between 0.1 Hz and 40 Hz, and then downsampled to 128 Hz. We further downsampled the channel space to the standard 64 channel BioSemi montage. The index and middle finger blocks for each hand were combined for binary classification of movements originating from the left or right hand. EEG trials of left and right hand finger movements were extracted at $[-.5, 1]$ s around finger movement onset and used for a two-class classification.

2.1.4. Dataset 4: Sensory motor rhythm (SMR). A common control signal for oscillatory-based BCI is the sensorimotor rhythm (SMR), wherein mu (8–12 Hz) and beta (18–26 Hz) bands desynchronize over the sensorimotor cortex contralateral to an actual or imagined movement. The SMR is very similar to the oscillatory component of the MRCP. Although SMR-based BCIs can facilitate nuanced, endogenous BCI control, they tend to be weak and highly variable across and within subjects, conventionally demanding user-training (neurofeedback) and long calibration times (20 min) in order to achieve reasonable performance [45].

The EEG data used here comes from BCI Competition IV Dataset 2A [74] (called the SMR dataset for the remainder of the manuscript). The data consists of four classes of imagined movements of left and right hands, feet and tongue recorded from nine subjects. The EEG data were originally recorded using 22 Ag/AgCl electrodes, sampled at 250 Hz and bandpass filtered between 0.5 and 100 Hz. We resampled the timeseries to 128 Hz, and follow the same EEG pre-processing procedure as described in [32], using software that was provided by the authors; a brief summary is given here. The data were causally filtered using a third-order Butterworth filter in the 4–40 Hz frequency band to minimize the influence of class-discriminative eye movements. The EEG signals were then standardized with an exponential moving average window with a decay factor of 0.999 (further details can be found in section A.7 of [32]).

For both the training and test sets we epoched the data at [0.5, 2.5] seconds post cue onset (the same window which was used in [40, 45]). Note that we make predictions for only this time range on the test set. We perform a four-class classification using accuracy as the summary measure.

2.2. Classification methods

2.2.1. EEGNet: compact CNN architecture. Here we introduce EEGNet, a compact CNN architecture for EEG-based BCIs that (1) can be applied across several different BCI paradigms, (2) can be trained with very limited data and (3) can produce neurophysiologically interpretable features. A visualization and full description of the EEGNet model can be found in figure 1 and table 2, respectively, for EEG trials, collected at 128 Hz sampling rate, having C channels and T time samples. We fit the model using the Adam optimizer, using default parameters as described in [75], minimizing the categorical cross-entropy loss function. We run 500 training iterations (epochs) and perform validation stopping, saving the model weights which produced the lowest validation set loss. All models were trained on an NVIDIA Quadro M6000 GPU, with CUDA 9 and cuDNN v7, in Tensorflow [76], using the Keras API [77]. We omit the use of bias units in all convolutional layers. Note that, while all convolutions are one-dimensional, we use two-dimensional convolution functions for ease of software implementation. Our software implementation can be found at https://github.com/vlawhern/arl-eegmodels.

- In block 1, we perform two convolutional steps in sequence. First, we fit F_1 2D convolutional filters of size $(1, 64)$, with the filter length chosen to be half the sampling rate of the data (here, 128 Hz), outputting F_1 feature maps containing the EEG signal at different band-pass frequencies. Setting the length of the temporal kernel at half the sampling rate allows for capturing frequency information at 2 Hz and above. We then use a *depthwise convolution* [43] of size $(C, 1)$ to learn a spatial filter. In CNN applications for computer vision the main benefit of a depthwise convolution is reducing the number of trainable parameters to fit, as these convolutions are not fully-connected to all previous feature maps (see figure 1 for an illustration). Importantly, when used in EEG-specific applications, this operation provides a direct way to learn spatial filters for each temporal filter, thus enabling the efficient extraction of frequency-specific spatial filters (see the middle column of figure 1). A depth parameter D controls the number of spatial filters to learn for each feature map ($D = 1$ is shown in figure 1 for illustration purposes). This two-step convolutional sequence is inspired in part by the filter-bank common spatial pattern (FBCSP) algorithm [78] and is similar in nature to another decomposition technique, bilinear discriminant component analysis [79]. We keep both convolutions linear as we found no significant gains in performance when using nonlinear activations. We apply batch normalization [80] along the feature map dimension before applying the exponential linear unit (ELU) nonlinearity [81]. To help regularize or model, we use the dropout technique [82]. We set the dropout probability to 0.5 for within-subject classification to help prevent over-fitting when training on small sample sizes, whereas we set the dropout probability to 0.25 in cross-subject classification, as the training set sizes are much larger

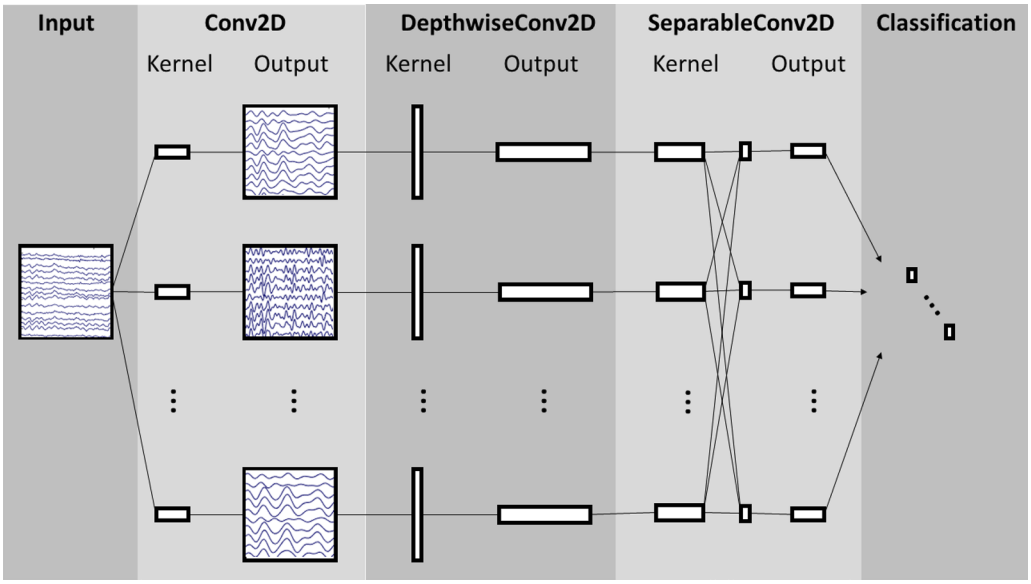

Figure 1. Overall visualization of the EEGNet architecture. Lines denote the convolutional kernel connectivity between inputs and outputs (called *feature maps*) . The network starts with a temporal convolution (second column) to learn frequency filters, then uses a depthwise convolution (middle column), connected to each feature map individually, to learn frequency-specific spatial filters. The separable convolution (fourth column) is a combination of a depthwise convolution, which learns a temporal summary for each feature map individually, followed by a pointwise convolution, which learns how to optimally mix the feature maps together. Full details about the network architecture can be found in table 2.

Table 2. EEGNet architecture, where C = number of channels, T = number of time points, F_1 = number of temporal filters, D = depth multiplier (number of spatial filters), F_2 = number of pointwise filters, and N = number of classes, respectively. For the Dropout layer, we use $p = 0.5$ for within-subject classification and $p = 0.25$ for cross-subject classification (see section 2.2.1 for more details)

Block	Layer	# filters	Size	# params	Output	Activation	Options
1	Input				(C, T)		
	Reshape				(1, C, T)		
	Conv2D	F_1	(1, 64)	$64 * F_1$	(F_1, C, T)	Linear	Mode = same
	BatchNorm			$2 * F_1$	(F_1, C, T)		
	DepthwiseConv2D	D $* F_1$	(C, 1)	$C * D * F_1$	(D $* F_1$, 1, T)	Linear	Mode = valid, depth = D, max norm = 1
	BatchNorm			$2 * D * F_1$	(D $* F_1$, 1, T)		
	Activation				(D $* F_1$, 1, T)	ELU	
	AveragePool2D		(1, 4)		(D $* F_1$, 1, T // 4)		
	Dropout$*$				(D $* F_1$, 1, T // 4)		$p = 0.25$ or $p = 0.5$
2	SeparableConv2D	F_2	(1, 16)	$16 * D * F_1 + F_2 * (D * F_1)$	(F_2, 1, T // 4)	Linear	Mode = same
	BatchNorm			$2 * F_2$	(F_2, 1, T // 4)		
	Activation				(F_2, 1, T // 4)	ELU	
	AveragePool2D		(1, 8)		(F_2, 1, T // 32)		
	Dropout$*$				(F_2, 1, T // 32)		$p = 0.25$ or $p = 0.5$
	Flatten				($F_2 *$ (T // 32))		
Classifier	Dense	N $* (F_2 *$ T // 32)			N	Softmax	Max norm = 0.25

(see section 2.3 for more details on our within- and cross-subject analyses). We apply an average pooling layer of size (1, 4) to reduce the sampling rate of the signal to 32 Hz. We also regularize each spatial filter by using a maximum norm constraint of 1 on its weights; $\|w\|^2 < 1$.

- In block 2, we use a *separable convolution*, which is a depthwise convolution (here, of size $(1, 16)$, representing 500 ms of EEG activity at 32 Hz) followed by F_2 $(1, 1)$ pointwise convolutions [43]. The main benefits of separable convolutions are (1) reducing the number of parameters to fit and (2) explicitly decoupling the relationship within and across feature maps by first learning a kernel summarizing each feature map individually, then optimally merging the outputs afterwards. When used for EEG-specific applica-

Table 3. Number of trainable parameters per model and per dataset for all CNN-based models. We see that the EEGNet models are up to two orders of magnitude smaller than both DeepConvNet and ShallowConvNet across all datasets. Note that we use a temporal kernel length of 32 samples for the SMR dataset as the data were high-passed at 4 Hz.

	Trial length (s)	DeepConvNet	ShallowConvNet	EEGNet-4,2	EEGNet-8,2
P300	1	174 127	104 002	**1066**	2258
ERN	1.25	169 927	91 602	**1082**	2290
MRCP	1.5	175 727	104 722	**1098**	2322
SMR∗	2	152 219	40 644	**796**	1716

tions this operation separates learning how to summarize individual feature maps in time (the depthwise convolution) with how to optimally combine the feature maps (the pointwise convolution). This operation is also particularly useful for EEG signals as different feature maps may represent data at different time-scales of information. In our case we first learn a 500 ms 'summary' of each feature map, then combine the outputs afterwards. An average pooling layer of size $(1, 8)$ is used for dimension reduction.

- In the classification block, the features are passed directly to a softmax classification with N units, N being the number of classes in the data. We omit the use of a dense layer for feature aggregation prior to the softmax classification layer to reduce the number of free parameters in the model, inspired by the work in [83].

We investigate several different configurations of the EEGNet architecture by varying the number of filters, F_1, and the number of spatial filters per temporal filter, D to learn. We set $F_2 = D * F_1$ (the number of temporal filters along with their associated spatial filters from block 1) for the duration of the manuscript, although in principle F_2 can take any value; $F_2 < D * F_1$ denotes a compressed representation, learning fewer feature maps than inputs, whereas $F_2 > D * F_1$ denotes an overcomplete representation, learning more feature maps than inputs. We use the notation EEGNet-F_1,D to denote the number of temporal and spatial filters to learn; i.e.: EEGNet-4,2 denotes learning four temporal filters and two spatial filters per temporal filter.

2.2.2. Comparison with existing CNN approaches. We compare the performance of EEGNet against the DeepConvNet and ShallowConvNet models proposed by [32]; full table descriptions of both models can be found in the appendix. We implemented these models in Tensorflow and Keras, following the descriptions found in the paper. As their architectures were originally designed for 250 Hz EEG signals (as opposed to 128 Hz signals used here) we divided the lengths of temporal kernels and pooling layers in their architectures by two to correspond approximately to the sampling rate used in our models. We train these models in the same way we train the EEGNet model (see section 2.2.1).

The DeepConvNet architecture consists of five convolutional layers with a softmax layer for classification (see figure 1 of [32]). The ShallowConvNet architecture consists of two convolutional layers (temporal, then spatial), a squaring nonlinearity ($f(x) = x^2$), an average pooling layer and a log nonlinearity ($f(x) = \log(x)$). We would like to emphasize that the ShallowConvNet architecture was designed specifically for oscillatory signal classification (by extracting features related to log band-power); thus, it may not work well on ERP-based classification tasks. However, the DeepConvNet architecture was designed to be a general-purpose architecture that is not restricted to specific feature types [32], and thus it serves as a more valid comparison to EEGNet. Table 3 shows the number of trainable parameters per model across all CNN models.

2.2.3. Comparison with traditional approaches. We also compare the performance of EEGNet to that of the best performing traditional approach for each individual paradigm. For all ERP-based data analyses (P300, ERN, MRCP) the traditional approach is the approach which won the Kaggle BCI competition (code and documentation at http://github.com/alexandrebarachant/bci-challenge-ner-2015), which uses a combination of xDAWN spatial filtering [84], Riemannian geometry [85, 86], channel subset selection and L_1 feature regularization (referred to as xDAWN + RG for the remainder of the manuscript). Here we provide a summary of the approach, which is done in five steps:

1. Train two set of 5 xDAWN spatial filters, one set for each class of a binary classification task, using the ERP template concatenation method as described in [86, 87].
2. Perform EEG electrode selection through backward elimination [88] to keep only the most relevant 35 channels.
3. Project the covariance matrices onto the tangent space using the log-Euclidean metric [85, 89].
4. Perform feature normalization using an L_1 ratio of 0.5, signifying an equal weight for L_1 and L_2 penalties. An L_1 penalty encourages the sum of the absolute values of the parameters to be small, whereas an L_2 penalty encourages the sum of the squares of the parameters to be small (a theoretical overview of these penalties can be found in [90]).
5. Perform classification using an elastic net regression.

We use the same xDAWN+RG model parameters across all comparisons (P300, ERN, MRCP) with the exception of the initial number of EEG channels to use, which was set to 56 for ERN and 64 for P300 and MRCP. While the original solution used an ensemble of bagged classifiers, for this analysis we only compared a single model with this approach to a single EEGNet model on identical training and test sets, as we expect any gains from ensemble learning to benefit both approaches equally. The original solution also used a set of 'meta features' that were specific to that data collection. As the goal of this work is to investigate a general-purpose CNN model for EEG-based BCIs, we omitted the use of these features as they are specific to that particular data collection.

For oscillatory-based classification of SMR, the traditional approach is our own implementation of the one-versus-rest (OVR) filter-bank common spatial pattern (FBCSP) algorithm as described in [78]. Here we provide a brief summary of our approach:

1. Bandpass filter the EEG signal into nine non-overlapping filter banks in 4 Hz steps, starting at 4 Hz: 4–8 Hz, 8–12 Hz,..., 36–40 Hz.
2. As the classification problem is multi-class, we use OVR classification, which requires that we train a classifier for all pairs of OVR combinations, which there are four here (class 1 versus all others, class 2 versus all others, etc). We train two CSP filter pairs (four filters total) for each filter bank on the training data using the auto-covariance shrinkage method by [91]. This will give a total of 36 features (nine filter banks $\times$ four CSP filters) for each trial and each OVR combination.
3. Train an elastic-net logistic regression classifier [92] for each OVR combination. We set the elastic net penalty $\alpha = 0.95$.
4. Find the optimal λ value for the elastic-net logistic regression that maximizes the validation set accuracy by evaluating the trained classifiers on a held-out validation set. The multi-class label for each trial is the classifier that produces the highest probability among the 4 OVR classifiers.
5. Apply the trained classifiers to the test set, using the λ values obtained in step 4.

Note that this approach differs slightly from the original technique as proposed in [78], where they use a naive Bayes Parzen window classifier. We opted to use an elastic net logistic regression for ease of implementation, and the fact that it has been used in existing software implementations of FBCSP (for example, in BCILAB [93]).

2.3. Data analysis

Classification results are reported for two sets of analyses: within-subject and cross-subject. Within-subject classification uses a portion of the subjects data to train a model specifically for that subject, although cross-subject classification uses the data from other subjects to train a subject-agnostic model. While within-subject models tend to perform better than cross-subject models on a variety of tasks, there is ongoing research investigating techniques to minimize (or possibly eliminate) the need for subject-specific information to train robust systems [45, 52].

For within-subject, we use four-fold blockwise cross-validation, where two of the four blocks are chosen to be the training set, one block as the validation set, and the final block as testing. We perform statistical testing using a repeated-measures analysis of variance (ANOVA), modeling classification results (AUC for P300/MRCP/ERN and classification accuracy for SMR) as the response variable with subject number and classifier type as factors. For cross-subject analysis in P300 and MRCP we choose, at random, four subjects for the validation set, one subject for the test set, and all remaining subjects for the training set (see table 1 for number of subjects per dataset). This process was repeated 30 times, producing 30 different folds. We follow the same procedure for the ERN dataset, except we use the ten test subjects from the original Kaggle competition as the test set for each fold. We perform statistical testing using a one-way analysis of variance, using classifier type as the factor. For the SMR dataset, we partitioned the data as follows: For each subject, select the training data from five other subjects at random to be the training set and the training data from the remaining three subjects to be the validation set. The test set remains the same as the original test set for the competition. Note that this enforces a fully cross-subject classification analysis as we never use the test subjects' training data. This process is repeated ten times for each subject, creating 90 different folds. The mean and standard error of classification performance were calculated over the 90 folds. We perform statistical testing for this analysis using the same testing procedure as the within-subject analysis.

When training both the within-subject and cross-subject models, we apply a class-weight to the loss function whenever the data is imbalanced (unequal number of trials for each class). The class-weight we apply is the inverse of the proportion in the training data, with the majority class set to 1. For example, in the P300 dataset, there is a 5.6:1 odds between non-targets and targets (table 1) . In this case the class-weight for non-targets was set to 1, while the class-weight for targets was set to 6 (when the odds are a fraction, we take the next highest integer). This procedure was applied to the P300 and ERN datasets only, as these were the only datasets where significant class imbalance was present.

Note that for the SMR analysis, we set the temporal kernel length to be 32 samples long (as opposed to 64 samples long as given in table 2) since the data were high-passed at 4 Hz.

2.4. EEGNet feature explainability

The development of methods for enabling feature explainability from deep neural networks has become an active research area over the past few years, and has been proposed as an essential component of a robust model validation procedure, to ensure that the classification performance is being driven by relevant features as opposed to noise or artifacts in the data [16, 94–100]. We present three different approaches for understanding the features derived by EEGNet:

1. **Summarizing averaged outputs of hidden unit activations:** This approach focuses on summarizing the activations of hidden units at layers specified by the user. In this work we choose to summarize the hidden unit activations representing the data after the depthwise convolution (the spatial filter operation in EEGNet). Because the spatial filters are tied directly to a particular temporal filter, they provide additional insights into the spatial localization of narrow-band frequency activity. Here we summarize the spatially-filtered data by calculating the difference in averaged time-frequency representations between classes, using Morlet wavelets [101].
2. **Visualizing the convolutional kernel weights:** This approach focuses on directly visualizing and interpreting the convolutional kernel weights from the model. Generally speaking, interpreting the convolutional kernel

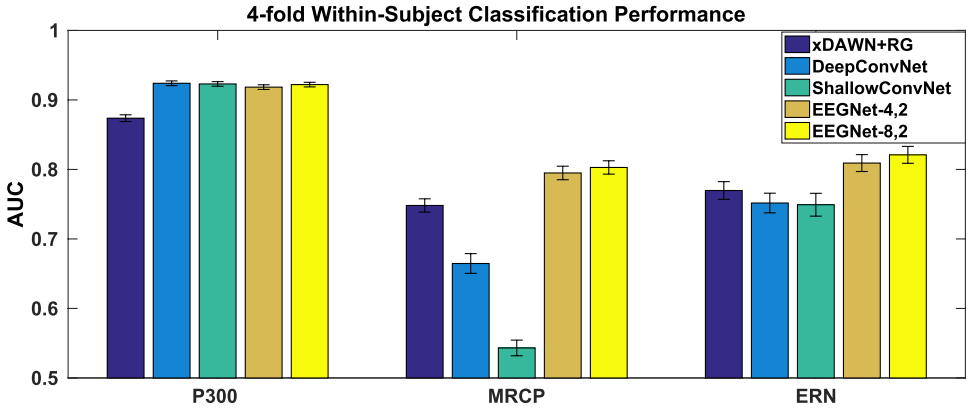

Figure 2. 4-fold within-subject classification performance for the P300, ERN and MRCP datasets for each model, averaged over all folds and all subjects. Error bars denote two standard errors of the mean. We see that, while there is minimal difference between all the CNN models for the P300 dataset, there are significant differences in the MRCP dataset, with both EEGNet models outperforming all other models. For the ERN dataset we also see both EEGNet models performing better than all others ($p < 0.05$).

weights is very difficult due to the cross-filter-map connectivity between any two layers. However, because EEGNet limits the connectivity of the convolutional layers (using depthwise and separable convolutions), it is possible to interpret the temporal convolution as narrow-band frequency filters and the depthwise convolution as frequency-specific spatial filters.

3. **Calculating single-trial feature *relevance* on the classification decision:** This approach focuses on calculating, on a single-trial basis, the *relevance* of individual features on the resulting classification decision. Positive values of relevance denote evidence supporting the outcome, while negative values of relevance denote evidence against the outcome. In our analysis we used DeepLIFT with the Rescale rule [98], as implemented in [99], to calculate single-trial EEG feature relevance. DeepLIFT is a gradient-based relevance attribution method that calculates relevance values per feature relative to a 'reference' input (here, an input of zeros, as is suggested in [98]), and is a technique similar to layerwise relevance propagation (LRP) which has been used previously for EEG analysis [33] (a summary of gradient-based relevance attribution methods can be found in [99]). This analysis can be used to elucidate feature relevance from high-confidence versus low-confidence predictions, and can be used to confirm that the relevant features learned are interpretable, as opposed to noise or artifact features.

3. Results

3.1. Within-subject classification

We compare the performance of both the CNN-based reference algorithms (DeepConvNet and ShallowConvNet) and the traditional approach (xDAWN+RG for P300/MRCP/ERN and FBCSP for SMR) with EEGNet-4, 2 and EEGNet-8, 2. Within-subject four-fold cross-validation results across all algorithms for P300, MRCP and ERN datasets are shown in figure 2. We observed, across all paradigms, that there was no statistically significant difference between EEGNet-4, 2 and EEGNet-8,2 ($p > 0.05$), indicating that the increase in model complexity did not statistically improve classification performance. For the P300 dataset, all CNN-based models significantly outperform xDAWN+RG ($p < 0.05$) while not performing significantly different amongst themselves. For the ERN dataset, EEGNet-8, 2 outperforms DeepConvNet, ShallowConvNet and xDAWN+RG ($p < 0.05$), while EEGNet-4, 2 outperforms DeepConvNet and ShallowConvNet ($p < 0.05$). The biggest difference observed among all the approaches is in the MRCP dataset, where both EEGNet models statistically outperform all others by a significant margin (DeepConvNet, ShallowConvNet and xDAWN+RG, $p < 0.05$ for each comparison).

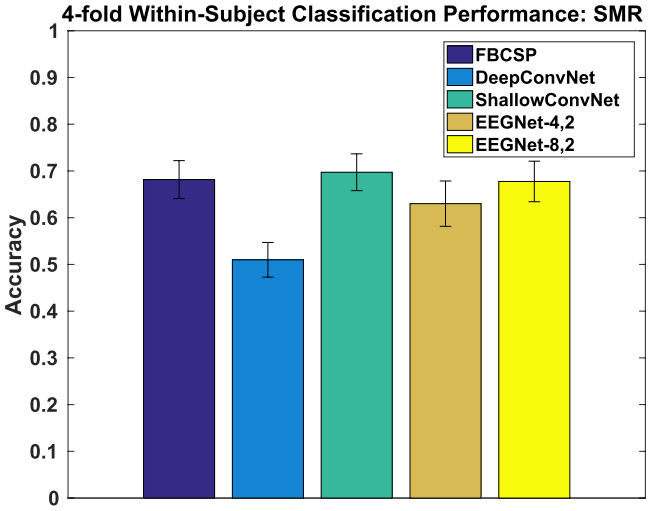

Figure 3. 4-fold within-subject classification performance for the SMR dataset for each model, averaged over all folds and all subjects. Error bars denote two standard errors of the mean. Here we see DeepConvNet statistically performed worse than all other models ($p < 0.05$). ShallowConvNet and EEGNet-8,2 performed similarly to that of FBCSP.

Four-fold cross-validation results for the SMR dataset are shown in figure 3. Here we see the performances of ShallowConvNet and FBCSP are very similar, replicating previous results as reported in [32], while DeepConvNet performance is significantly lower. We also see that EEGNet-8, 2 performance is similar to FBCSP as well.

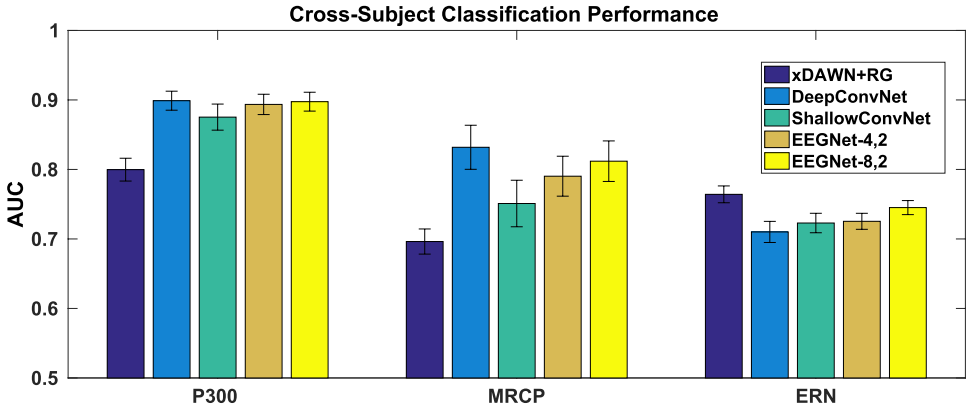

Figure 4. Cross-Subject classification performance for the P300, ERN and MRCP datasets for each model, averaged for 30 folds. Error bars denote two standard errors of the mean. For the P300 and MRCP datasets there is minimal difference between the DeepConvNet and the EEGNet models, with both models outperforming ShallowConvNet. For the ERN dataset the reference algorithm (xDAWN + RG) significantly outperforms all other models.

3.2. Cross-subject classification

Cross-subject classification results across all algorithms for P300, MRCP and ERN datasets are shown in figure 4. Similar to the within-subject analysis, we observed no statistical difference between EEGNet-4, 2 and EEGNet-8, 2 across all datasets ($p > 0.05$). For the P300 dataset, all CNN-based models significantly outperform xDAWN + RG ($p < 0.05$) while not performing significantly different amongst themselves. For the MRCP dataset EEGNet-8, 2 and DeepConvNet significantly outperform ShallowConvNet ($p < 0.05$). We also see that both DeepConvNet and ShallowConvNet performance is better when compared to its within-subject performance for the MRCP dataset. For the ERN dataset, xDAWN+RG outperforms all CNN models ($p < 0.05$). Cross-subject classification results for the SMR dataset are shown in figure 5, where we found no significant difference in performance across all CNN-based models ($p > 0.05$).

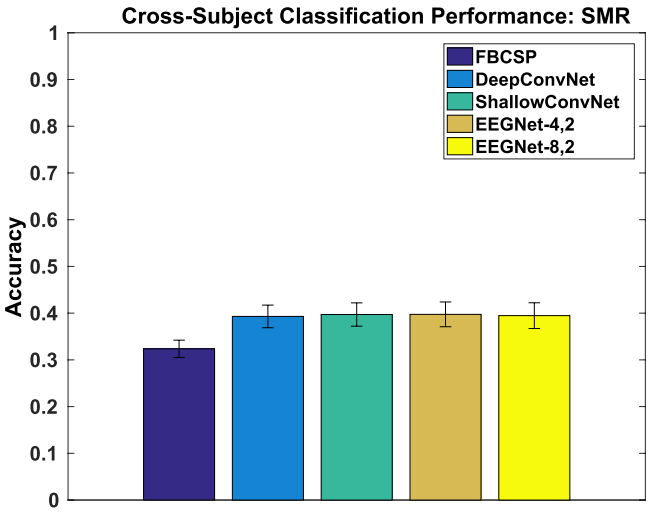

Figure 5. Cross-Subject classification performance for the SMR for each model, averaged over all folds and all subjects. Error bars denote two standard errors of the mean. We see that all CNN-based models perform similarly, while slightly outperforming FBCSP.

3.3. EEGNet feature characterization

We illustrate three different approaches to characterize the features learned by EEGNet: (1) Summarizing averaged outputs of hidden unit activations, (2) visualizing convolutional kernel weights, and (3) calculating single-trial feature relevances on classification decision. We illustrate approach 1 on the P300 dataset for a cross-subject trained EEGNet-4, 1 model. We chose to analyze the filters from the P300 dataset due to the fact that multiple neurophysiological events occur simultaneously: participants were told to press a button with their dominant hand whenever a target image appeared on the screen. Because of this, target trials contain both the P300 event-related potential as well as the alpha/beta desynchronizations in contralateral motor cortex due to button presses. Here we were interested in whether or not the EEGNet architecture was capable of separating out these confounding events. We were also interested in quantifying the classification performance of the architecture whenever specific filters were removed from the model.

Figure 6 shows the spatial topographies of the four filters along with an average wavelet time-frequency difference, calculated using Morlet wavelets [101], between all target trials and all non-target trials. Here we see four distinct filters appear. The time-frequency analysis of filter 1 shows an increase in low-frequency power approximately 500ms after image presentation, followed by desynchronizations in alpha frequency. As nearly all subjects in the P300 dataset are right-handed, we also see significant activity along the left motor cortex. Time-frequency analysis of filter 2 appears to show a significant theta-beta relationship; while increases in theta activity have been previously noted in the P300 literature in response to targets [102], a relationship between theta and beta has not previously been noted. The time-frequency difference for filter 4 appears to correspond with the P300, with an increase low-frequency power approximately 350 ms after image presentation.

We also conducted a feature ablation study, where we iteratively removed a set of filters (by replacing the filters with zeros) and re-applied the model to predict trials in the test set. We do this for all combinations of the four filters. Classification results for this ablation study are shown in table 4. We see that test set performance is minimally impacted by the removal of any single filter, with the largest

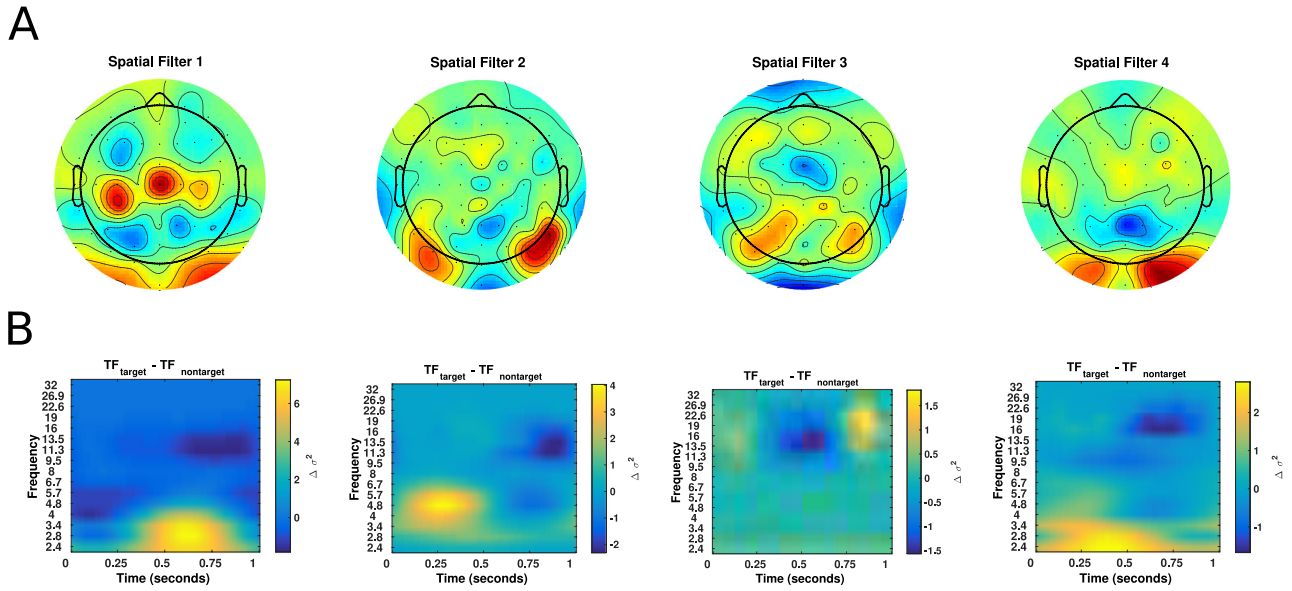

Figure 6. Visualization of the features derived from an EEGNet-4, 1 model configuration for one particular cross-subject fold in the P300 dataset. (A) Spatial topoplots for each spatial filter. (B) The mean wavelet time-frequency difference between target and non-target trials for each individual filter.

decrease occurring when removing filter 4. As expected, when removing pairs of filters the decrease in performance is more pronounced, with the largest decrease observed when removing filters 3 and 4. Removing filters 2 and 3 results in practically no change in classification performance when compared to the full model, suggesting that the most important features in this task are being captured by filters 1 and 4. This finding is further reinforced when looking at classification performance when three filters are removed; a model that contains only filter 4 (0.8637 AUC) performs fairly well when compared to models that contain only filter 2 (0.7108 AUC) or filter 1 (0.7970 AUC).

Figure 7 shows the filters learned for the EEGNet-8,2 model for a within-subject classification of subject 3 for the SMR dataset. Each column of this figure denotes the learned temporal kernel (top row) with its two associated spatial filters (bottom two rows). Note that we are learning temporal filters of length 32 samples, which correspond to 0.25 s in time; hence, we estimate the frequency for each temporal filter as four times the number of observed cycles. Here we see that EEGNet-8,2 learns both slow-frequency activity at approximately 12 Hz (filters 1, 2, 6 and 8, which show three cycles in a 0.25 s window) and high-frequency activity at approximately 32 Hz (filter 3, which show eight cycles). Figure 8 compares the spatial filters associated with 8–12 Hz frequency band learned by EEGNet-8,2 with the spatial filters learned by FBCSP in the 8–12 Hz filter-bank for each of the four OVR combinations. For ease of description we will use the notation X-Y to denote the row-column filter. Here we see many of the filters are strongly positively correlated across models (i.e.: the 1–1 filter of EEGNet-8,2 with the 3–1 filter for FBCSP ($\rho = 0.93$) and the 2–1 filter of EEGNet-8,2 with the 3–4 filter of FBCSP ($\rho = 0.83$)), while some are strongly negatively correlated (the 3–1 filter of EEGNet-8,2 with the 1–1 filter of FBCSP ($\rho = -0.93$)), indicating a similar filter up to a sign ambiguity.

Table 4. Performance of a cross-subject trained EEGNet-4, 1 model when removing certain filters from the model, then using the model to predict the test set for one randomly chosen fold of the P300 dataset. AUC values in bold denote the best performing model when removing one, two or three filters at a time. As the number of filters removed increases, we see decreases in classification performance, although the magnitude of the decrease depends on which filters are removed.

Filters removed	Test set AUC
(1)	0.8866
(2)	**0.9076**
(3)	0.8910
(4)	0.8747
(1, 2)	0.8875
(1, 3)	0.8593
(1, 4)	0.8325
(2, 3)	**0.8923**
(2, 4)	0.8721
(3, 4)	0.8206
(1, 2, 3)	**0.8637**
(1, 2, 4)	0.8202
(1, 3, 4)	0.7108
(2, 3, 4)	0.7970
None	**0.9054**

Figure 9 shows the single-trial feature relevances for EEGNet-8,2, calculated using DeepLIFT, for three three different test trials for one cross-subject fold of the MRCP dataset. Here we see that the high-confidence predictions (figures 9(A) and (B), for left and right finger movement, respectively) both correctly show the contralateral motor cortex relevance as expected, whereas for a low-confidence prediction (figure 9(C)), the feature relevance is more broadly distributed, both in time and in space on the scalp.

Figure 10 shows an additional example of using DeepLIFT to analyze feature relevance for a cross-subject trained EEGNet-4,2 model for one test subject of the ERN

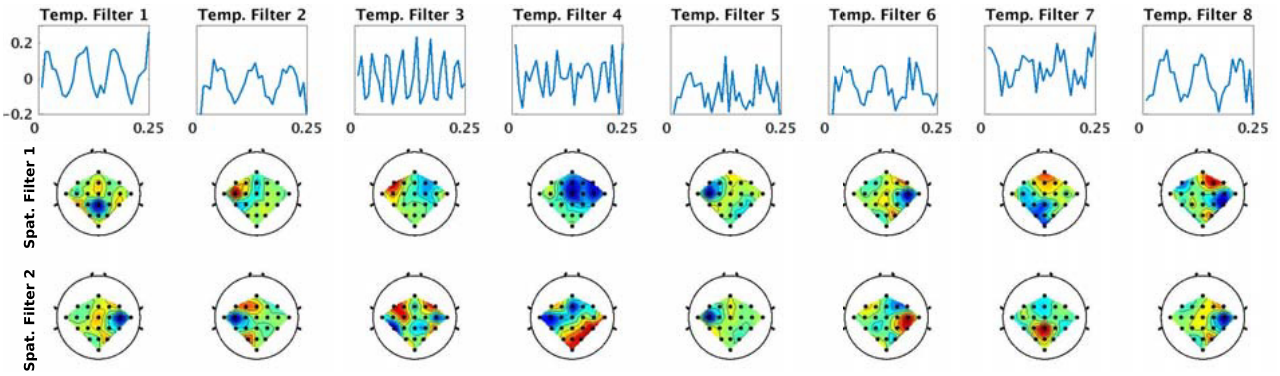

Figure 7. Visualization of the features derived from a within-subject trained EEGNet-8, 2 model for Subject 3 of the SMR dataset. Each of the 8 columns shows the learned temporal kernel for a 0.25 s window (top) with its two associated spatial filters (bottom two). We see that, while many of the temporal filters are isolating slower-wave activity, the network identifies a higher-frequency filter at approximately 32 Hz (temp. filter 3, which shows eight cycles in a 0.25 s window).

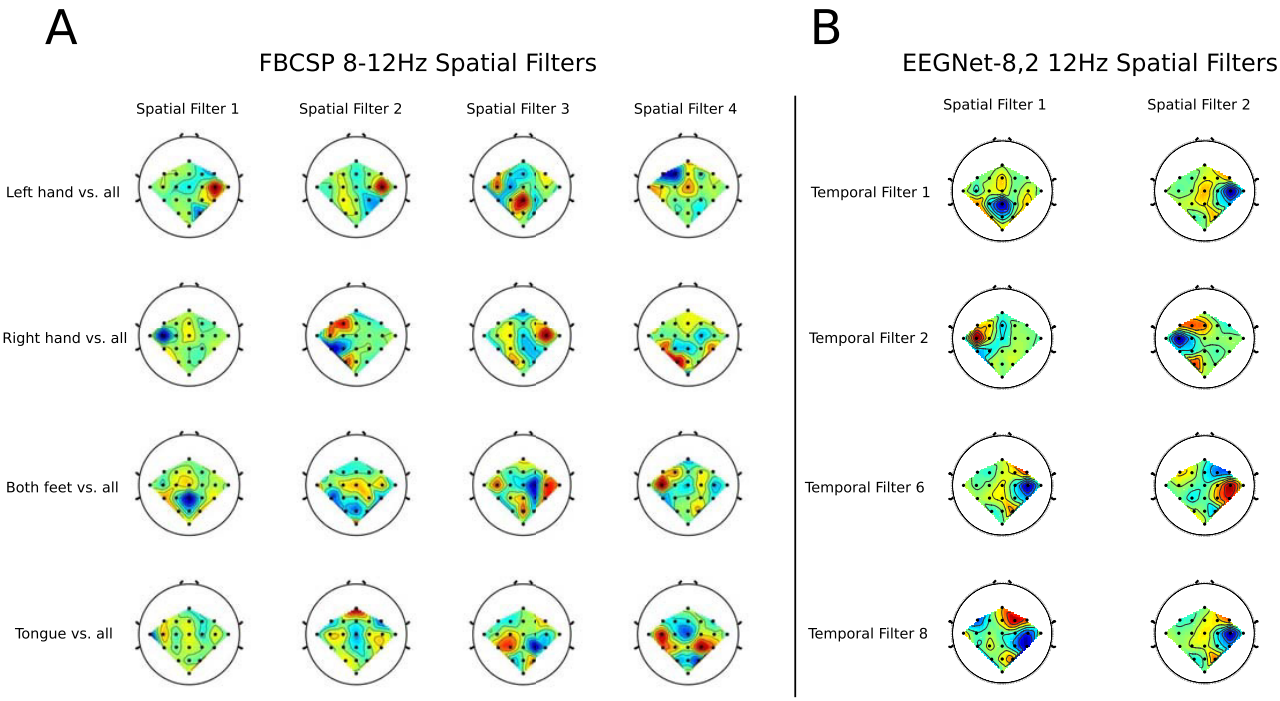

Figure 8. Comparison of the four spatial filters learned by FBCSP in the 8–12 Hz filter bank for each OVR class combination (A) with the spatial filters learned by EEGNet-8, 2 (B) for four temporal filters that capture 12 Hz frequency activity for subject 3 of the SMR dataset (temporal filters 1, 2, 6 and 8, see figure 7). We see that, for this subject, similar filters appear across both FBCSP and EEGNet-8, 2.

dataset. Margaux *et al* [57] previously noted that the average ERP for correct feedback trials has an earlier peak positive potential, corresponding to approximately 325 ms, whereas the positive peak potential for incorrect trials occurs slightly later, approximately 475 ms. Here we see the same temporal difference in the timing of the peak positive potential for incorrect feedback trials (vertical line in top row of figure 10) and correct feedback trials (vertical line in bottom row of figure 10). We also see the DeepLIFT feature relevances align very closely to that of the peak positive potential for both classes, suggesting that the network has focused on the peak positive potential as the relevant feature for ERN classification. This finding supports results previously reported in [57], where they showed a strong positive correlation between the amplitude of the peak positive potential and the accuracy of error detection.

4. Discussion

In this work we proposed *EEGNet*, a compact convolutional neural network for EEG-based BCIs that can generalize across different BCI paradigms in the presence of limited data and can produce interpretable features. We evaluated EEGNet against the state-of-the-art approach for both ERP and oscillatory-based BCIs across four EEG datasets: P300 visual-evoked potentials, error-related negativity (ERN), movement-related cortical potentials (MRCP) and sensory motor rhythms (SMR). To the best of our knowledge, this represents the first work that has validated the use of a single network architecture across multiple BCI datasets, each with their own feature characteristics and data set sizes. Our work introduced the use of depthwise and separable convolutions [43] for EEG signal classification, and showed that they can be used to construct

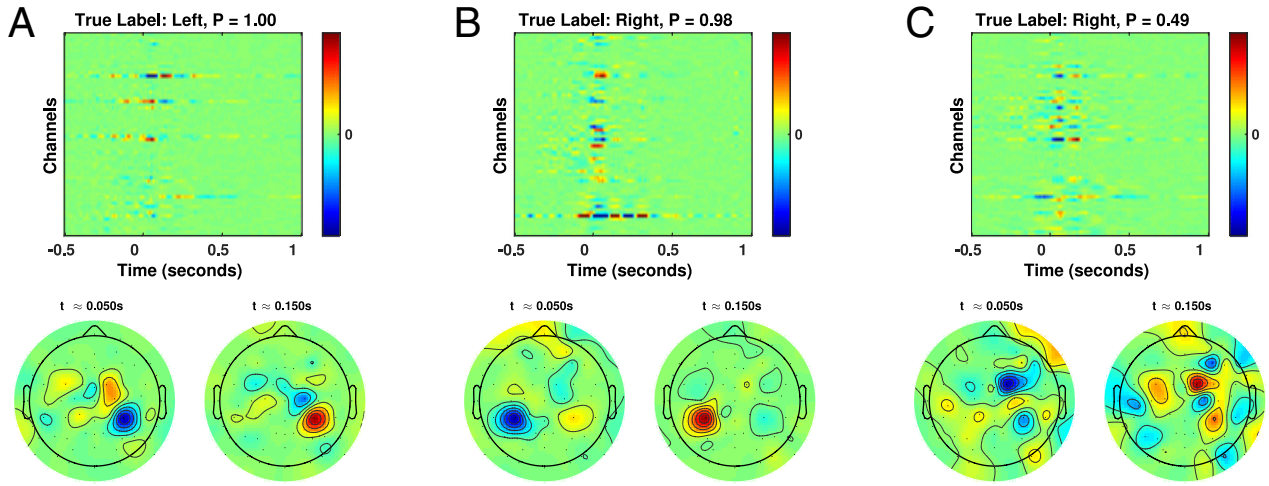

Figure 9. (Top row) Single-trial EEG feature relevance for a cross-subject trained EEGNet-8,2 model, using DeepLIFT, for three different test trials of the MRCP dataset: (A) a high-confidence, correct prediction of left finger movement, (B) a high-confidence, correct prediction of right finger movement and (C) a low-confidence, incorrect prediction of left finger movement. Titles include the true class label and the predicted probability of that label. (Bottom row) Spatial topoplots of the relevances at two time points: approximately 50 ms and 150 ms after button press. As expected, the high-confidence trials show the correct relevances corresponding to contralateral motor cortex for left (A) and right (B) button presses, respectively. For the low-confidence trial we see the relevances are more mixed and broadly distributed, without a clear spatial localization to motor cortices.

an EEG-specific model which encapsulates well-known EEG feature extraction concepts. Finally, through the use of feature visualization and ablation analysis, we show that neurophysiologically interpretable features can be extracted from the EEGNet model. This last finding is particularly important, as it is a critical component to understanding the validity and robustness of CNN model architectures not just for EEG [32, 33], but for CNN architectures in general [16, 95, 100].

The learning capacity of CNNs comes in part from their ability to automatically extract intricate feature representations from raw data. However, since the features are not hand-designed by human engineers, understanding the meaning of those features poses a significant challenge in producing interpretable models [96]. This is especially true when CNNs are used for the analysis of EEG data where features from neural signals are often non-stationary and corrupted by noise artifacts [103, 104]. In this study, we illustrated three different approaches for visualizing the features learned by EEGNet: (1) analyzing spatial filter outputs, averaged over trials, on the P300 dataset, (2) visualizing the convolutional kernel weights on the SMR dataset and comparing them to the weights learned by FBCSP, and (3) performing single-trial relevance analysis on the MRCP and SMR datasets. For the ERN dataset we compared single-trial feature relevances to averaged ERPs and showed that relevant features coincided with the peak of the positive potential for correct and incorrect feedback trials, which has been shown in previous literature to be positively correlated to classifier performance [57]. In addition, we conducted a feature ablation study to understand the impact of a classification decision on the presence or absence of a particular feature on the P300 dataset. In each of these analyses, we showed that EEGNet was capable of extracting interpretable features that generally corresponded to known neurophysiological phenomena.

Generally speaking, the classification performance of DeepConvNet and EEGNet were similar across all cross-subject analyses, whereas DeepConvNet performance was lower across nearly all within-subject analyses (with the exception of P300). One possible explanation for this discrepancy is the amount of training data used to train the model; in cross-subject analyses the training set sizes were about 10–15 times larger than that of within-subject analyses. This suggests that DeepConvNet is more data-intensive compared to EEGNet, an unsurprising result given that the model size of DeepConvNet is two orders of magnitude larger than EEGNet (see table 3). We believe this intuition is consistent with the findings originally reported by the developers of DeepConvNet [32], where they state that a training data augmentation strategy was needed to obtain good classification performance on the SMR dataset. In contrast to their work, we show that EEGNet performed well across all tested datasets without the need for data augmentation, making the model simpler to use in practice.

In general we found that, both in within- and cross-subject analyses, that ShallowConvNet tended to perform worse on the ERP BCI datasets than on the oscillatory BCI dataset (SMR), while the opposite behavior was observed with DeepConvNet. We believe this is due to the fact that the ShallowConvNet architecture was designed specifically to extract log band-power features; in situations where the dominant feature is signal amplitude (as is the case in many ERP BCIs), ShallowConvNet performance tended to suffer. The opposite situation occurred with DeepConvNet; as its architecture was not designed to extract frequency features, its performance was lower in situations where frequency power is the dominant feature. In contrast, we found that EEGNet performed just as well as ShallowConvNet in SMR classification and just as well as DeepConvNet in ERP classification (and outperforming in the case of within-subject MRCP, ERN and SMR classifications), suggesting that EEGNet is robust enough to learn a wide variety of features over a range of BCI tasks.

The severe underperformance of ShallowConvNet on within-subject MRCP classification was unexpected, given the similarity in neural responses between the MRCP and SMR,

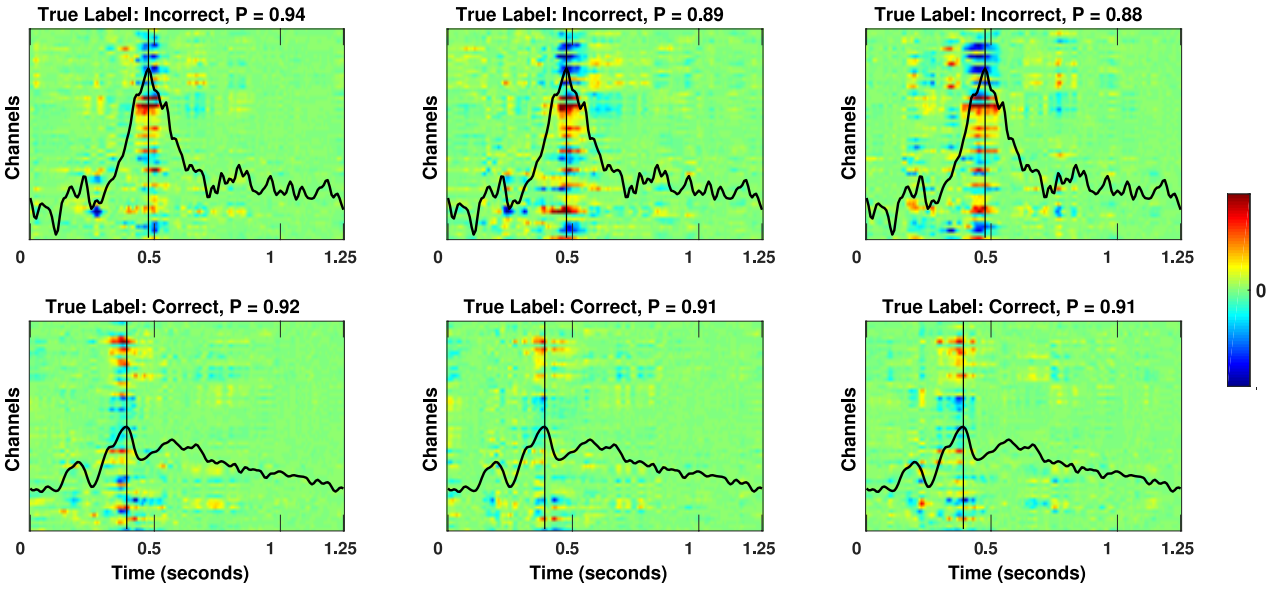

Figure 10. Single-trial EEG feature relevance for a cross-subject trained EEGNet-4,2 model, using DeepLIFT, for the one test subject of the ERN dataset. (top row) Feature relevances for three correctly predicted trials of incorrect feedback, along with its predicted probability *P*. (bottom row) Same as the top row but for three correctly predicted trials of correct feedback. The black line denotes the average ERP, calculated at channel Cz, for incorrect feedback trials (top row) and for correct feedback trials (bottom row). The thin vertical line denotes the positive peak of the average ERP waveform. Here we see feature relevances coincide strongly with the positive peak of the average ERP waveform for each trial. We also see the positive peak occurring slightly earlier for correct feedback trials versus incorrect feedback trials, consistent with the results in [57].

and the fact that ShallowConvNet performed well on SMR. This discrepancy in performance is not due to the amount of training data used, as within-subject MRCP classification has approximately 700 training trials, evenly split among left and right finger movements, whereas the SMR dataset has only 192 training trials, evenly split among four classes. In addition, we did not observe large deviations in ShallowConvNet performance on the other datasets (P300 and ERN). In fact, ShallowConvNet performed fairly well on within-subject ERN classification, even though this dataset is the smallest among all datasets used in this study (only having 170 training trials total). Determining the underlying source of this phenomena will be explored in future research.

Deep learning models for EEG generally employ one of three input styles, depending on their targeted application: (1) the EEG signal of all available channels, (2) a transformed EEG signal (generally a time-frequency decomposition) of all available channels [37] or (3) a transformed EEG signal of a subset of channels [38]. Models that fall in (2) generally see a significant increase in data dimensionality, thus requiring either more data or more model regularization (or both) to learn an effective feature representation. This introduces more hyperparameters that must be learned, increasing the potential variability in model performance due to hyperparameter misspecification. Models that fall in (3) generally require *a priori* knowledge about the channels to select. For example, the model proposed in [38] uses the time-frequency decomposition of channels Cz, C3 and C4 as the inputs for a motor imagery classification task. This channel selection is intentional, given the fact that neural responses to motor actions (the sensory motor rhythm) are observed strongest at those channels and are easily observed through a time-frequency analysis. Also, by only working with three channels, the authors reduce the significant increase in dimensionality of the data. While this approach works well if the feature of interest is known beforehand, this approach is not guaranteed to work well in other applications where the features are not observed at those channels, limiting the overall utility of this approach. We believe models that fall in (1), such as EEGNet and others [28, 30, 31], offer the best tradeoff between input dimensionality and the flexibility to discover relevant features by providing all available channels. This is especially important as BCI technologies evolve into novel application spaces, as the features needed for these future BCIs may not be known beforehand [3–5, 10–12].

Acknowledgments

This project was sponsored by the US Army Research Laboratory under ARL-H70-HR52, ARL-74A-HRCYB and through the Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix

Table A1. DeepConvNet architecture, where C = number of channels, T = number of time points and N = number of classes, respectively.

Layer	# filters	Size	# params	Activation	Options
Input		(C, T)			
Reshape		(1, C, T)			
Conv2D	25	(1, 5)	150	Linear	Mode = valid, max norm = 2
Conv2D	25	(C, 1)	25 * 25 * C + 25	Linear	Mode = valid, max norm = 2
BatchNorm			2 * 25		epsilon = 1×10^{-05}, momentum = 0.1
Activation				ELU	
MaxPool2D		(1, 2)			
Dropout					$p = 0.5$
Conv2D	50	(1, 5)	25 * 50 * C + 50	Linear	Mode = valid, max norm = 2
BatchNorm			2 * 50		epsilon = 1×10^{-05}, momentum = 0.1
Activation				ELU	
MaxPool2D		(1, 2)			
Dropout					$p = 0.5$
Conv2D	100	(1, 5)	50 * 100 * C + 100	Linear	Mode = valid, max norm = 2
BatchNorm			2 * 100		epsilon = 1×10^{-05}, momentum = 0.1
Activation				ELU	
MaxPool2D			(1, 2)		
Dropout					$p = 0.5$
Conv2D	200	(1, 5)	100 * 200 * C + 200	Linear	Mode = valid, max norm = 2
BatchNorm			2 * 200		epsilon = 1×10^{-05}, momentum = 0.1
Activation				ELU	
MaxPool2D		(1, 2)			
Dropout					$p = 0.5$
Flatten					
Dense	N			Softmax	Max norm = 0.5

Table A2. ShallowConvNet architecture, where C = number of channels, T = number of time points and N = number of classes, respectively. Here, the 'square' and 'log' activation functions are given as $f(x) = x^2$ and $f(x) = \log(x)$, respectively. Note that we clip the log function such that the minimum input value is a very small number ($\epsilon = 10 \times 10^{-7}$) for numerical stability.

Layer	# filters	Size	# params	Activation	Options
Input		(C, T)			
Reshape		(1, C, T)			
Conv2D	40	(1, 13)	560	Linear	Mode = same, max norm = 2
Conv2D	40	(C, 1)	40 * 40 * C	Linear	Mode = valid, max norm = 2
BatchNorm			2 * 40		epsilon = 1×10^{-05}, momentum = 0.1
Activation				Square	
AveragePool2D		(1, 35), stride (1, 7)			
Activation				Log	
Flatten					
Dropout					$p = 0.5$
Dense	N			Softmax	Max norm = 0.5

A.1. DeepConvNet and ShallowConvNet architectures

The DeepConvNet and ShallowConvNet architectures are given in tables A1 and A2, respectively. The DeepConvNet was designed to be a general-purpose architecture that is not restricted to specific feature types, whereas ShallowConvNet is designed specifically for oscillatory signal classification.

ORCID iDs

Vernon J Lawhern https://orcid.org/0000-0002-3921-8723

References

[1] Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91

[2] Schwartz A B, Cui X T, Weber D and Moran D W 2006 Brain-controlled interfaces: movement restoration with neural prosthetics *Neuron* **52** 205–20

[3] van Erp J, Lotte F and Tangermann M 2012 Brain–computer interfaces: beyond medical applications *Computer* **45** 26–34

[4] Saproo S, Faller J, Shih V, Sajda P, Waytowich N R, Bohannon A, Lawhern V J, Lance B J and Jangraw D 2016 Cortically coupled computing: a new paradigm for synergistic human–machine interaction *Computer* **49** 60–8

[5] Lance B J, Kerick S E, Ries A J, Oie K S and McDowell K 2012 Brain–computer interface technologies in the coming decades *Proc. IEEE* **100** 1585–99

[6] Nicolas-Alonso L F and Gomez-Gil J 2012 Brain–computer interfaces, a review *Sensors* **12** 1211

[7] Bashashati A, Fatourechi M, Ward R K and Birch G E 2007 A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals *J. Neural Eng.* **4** R32

[8] McFarland D J, Anderson C W, Muller K R, Schlogl A and Krusienski D J 2006 Bci meeting 2005-workshop on BCI signal processing: feature extraction and translation *IEEE Trans. Neural Syst. Rehabil. Eng.* **14** 135–8

[9] Lotte F, Congedo M, Lécuyer A, Lamarche F and Arnaldi B 2007 A review of classification algorithms for EEG-based brain–computer interfaces *J. Neural Eng.* **4** R1

[10] Zander T O and Kothe C 2011 Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general *J. Neural Eng.* **8** 025005

[11] Blankertz B *et al* 2010 The Berlin brain–computer interface: non-medical uses of BCI technology *Frontiers Neurosci.* **4** 198

[12] Gordon S M, Jaswa M, Solon A J and Lawhern V J 2017 Real world BCI: cross-domain learning and practical applications *Proc. 2017 ACM Workshop on an Application-Oriented Approach to BCI out of the Laboratory* (New York, NY, USA: ACM) pp 25–8

[13] Hinton G *et al* 2012 Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups *IEEE Signal Process. Mag.* **29** 82–97

[14] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44

[15] Krizhevsky A and Sutskever I 2012 Hinton imagenet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* ed F Pereira *et al* pp 1097–105

[16] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition *CoRR* (arXiv:1409.1556)

[17] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S E, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2014 Going deeper with convolutions *CoRR* (arXiv:1409.4842)

[18] He K, Zhang X, Ren S and Sun J 2015 Deep residual learning for image recognition *CoRR* (arXiv:1512.03385)

[19] Huang G, Liu Z, Weinberger K Q and van der Maaten L 2016 Densely connected convolutional networks *CoRR* (arXiv:1608.06993)

[20] Schmidhuber J 2014 Deep learning in neural networks: an overview (arXiv:1404.7828)

[21] Antoniades A, Spyrou L, Took C C and Sanei S 2016 Deep learning for epileptic intracranial EEG data *IEEE 26th Int. Workshop on Machine Learning for Signal Processing* pp 1–6

[22] Liang J, Lu R, Zhang C and Wang F 2016 Predicting seizures from electroencephalography recordings: a knowledge transfer strategy *IEEE Int. Conf. on Healthcare Informatics* pp 184–91

[23] Page A, Shea C and Mohsenin T 2016 Wearable seizure detection using convolutional neural networks with transfer learning *IEEE Int. Symp. on Circuits and Systems* pp 1086–9

[24] Mirowski P, Madhavan D, LeCun Y and Kuzniecky R 2009 Classification of patterns of EEG synchronization for seizure prediction *Clin. Neurophysiol.* **120** 1927–40

[25] Thodoroff P, Pineau J and Lim A 2016 Learning robust features using deep learning for automatic seizure detection *CoRR* (arXiv:1608.00220)

[26] Stober S, Cameron D J and Grahn J A 2014 Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings *Advances in Neural Information Processing Systems 27* ed Z Ghahramani pp 1449–57

[27] Stober S, Sternin A, Owen A M and Grahn J A 2015 Deep feature learning for EEG recordings *CoRR* (arXiv:1511.04306)

[28] Cecotti H and Graser A 2011 Convolutional neural networks for p300 detection with application to brain–computer interfaces *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 433–45

[29] Manor R and Geva A 2015 Convolutional neural network for multi-category rapid serial visual presentation BCI *Frontiers Comput. Neurosci.* **9** 146

[30] Shamwell J, Lee H, Kwon H, Marathe A R, Lawhern V and Nothwang W 2016 Single-trial EEG rsvp classification using convolutional neural networks *Proc. SPIE* **9836** 983622

[31] Cecotti H, Eckstein M P and Giesbrecht B 2014 Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering *IEEE Trans. Neural Netw. Learn. Syst.* **25** 2030–42

[32] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420

[33] Sturm I, Lapuschkin S, Samek W and Müller K-R 2016 Interpretable deep neural networks for single-trial EEG classification *J. Neurosci. Methods* **274** 141–5

[34] Längkvist M, Karlsson L and Loutfi A 2012 Sleep stage classification using unsupervised feature learning *Adv. Artif. Neu. Sys.* **2012** 5

[35] Wulsin D F, Gupta J R, Mani R, Blanco J A and Litt B 2011 Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement *J. Neural Eng.* **8** 036015

[36] Ma T, Li H, Yang H, Lv X, Li P, Liu T, Yao D and Xu P 2017 The extraction of motion-onset vep BCI features based on deep learning and compressed sensing *J. Neurosci. Methods* **275** 80–92

[37] Bashivan P, Rish I, Yeasin M and Codella N 2015 Learning representations from EEG with deep recurrent-convolutional neural networks *CoRR* (arXiv:1511.06448)

[38] Tabar Y R and Halici U 2017 A novel deep learning approach for classification of EEG motor imagery signals *J. Neural Eng.* **14** 016003

[39] An X, Kuang D, Guo X, Zhao Y and He L 2014 A deep learning method for classification of EEG data based on motor imagery *Intelligent Computing in Bioinformatics (ICIC 2014) (Lect. Not. Comput. Sci.* vol 8590) ed D S Huang *et al* (Cham: Springer) pp 203–10

[40] Sakhavi S, Guan C and Yan S 2015 Parallel convolutional-linear neural network for motor imagery classification *23rd European Signal Processing Conf.* pp 2736–40
[41] Lu N, Li T, Ren X and Miao H 2017 A deep learning scheme for motor imagery classification based on restricted Boltzmann machines *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 566–76
[42] Yin Z and Zhang J 2017 Cross-session classification of mental workload levels using eeg and an adaptive deep learning model *Biomed. Signal Process. Control* **33** 30–47
[43] Chollet F 2016 Xception: deep learning with depthwise separable convolutions *CoRR* (arXiv:1610.02357)
[44] Yang Z, Moczulski M, Denil M, Freitas N D, Smola A, Song L and Wang Z 2015 Deep fried convnets *IEEE Int. Conf. on Computer Vision* pp 1476–83
[45] Lotte F 2015 Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces *Proc. IEEE* **103** 871–90
[46] Fazel-Rezai R, Allison B Z, Guger C, Sellers E W, Kleih S C and Kübler A 2012 P300 brain–computer interface: current challenges and emerging trends *Frontiers Neuroeng.* **5** 14
[47] Pfurtscheller G and Neuper C 2001 Motor imagery and direct brain–computer communication *Proc. IEEE* **89** 1123–34
[48] Makeig S 1993 Auditory event-related dynamics of the {EEG} spectrum and effects of exposure to tones *Electroencephalogr. Clin. Neurophysiol.* **86** 283–93
[49] Polich J 2007 Updating p300: an integrative theory of P3a and P3b *Clin. Neurophysiol.* **118** 2128–48
[50] Sajda P, Pohlmeyer E, Wang J, Parra L C, Christoforou C, Dmochowski J, Hanna B, Bahlmann C, Singh M K and Chang S F 2010 In a blink of an eye and a switch of a transistor: cortically coupled computer vision *Proc. IEEE* **98** 462–78
[51] Marathe A R, Lawhern V J, Wu D, Slayback D and Lance B J 2016 Improved neural signal classification in a rapid serial visual presentation task using active learning *IEEE Trans. Neural Syst. Rehabil. Eng.* **24** 333–43
[52] Waytowich N, Lawhern V, Bohannon A, Ball K and Lance B 2016 Spectral transfer learning using information geometry for a user-independent brain-computer interface *Frontiers Neurosci.* **10** 430
[53] Delorme A and Makeig S 2004 Eeglab: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21
[54] Miltner W H R, Braun C H and Coles M G H 1997 Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a generic neural system for error detection *J. Cogn. Neurosci.* **9** 788–98
[55] Gehring W J, Goss B, Coles M G H, Meyer D E and Donchin E 1993 A neural system for error detection and compensation *Psychol. Sci.* **4** 385–90
[56] Falkenstein M, Hohnsbein J, Hoormann J and Blanke L 1991 Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks *Electroencephalogr. Clin. Neurophysiol.* **78** 447–55
[57] Margaux P, Emmanuel M, Sébastien D, Olivier B and Jérémie M 2012 Objective and subjective evaluation of online error correction during p300-based spelling *Adv. Hum. Comput. Interact.* **2012** 4
[58] Zander T O, Kothe C, Welke S and Roetting M 2009 Utilizing secondary input from passive brain–computer interfaces for enhancing human–machine interaction *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience (FAC 2009) (Lect. Not. Comput. Sci.* vol 5638) ed D D Schmorrow *et al* (Berlin: Springer) pp 759–71
[59] Millán J D R *et al* 2010 Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges *Frontiers Neurosci.* **4** 161
[60] Spüler M, Bensch M, Kleih S, Rosenstiel W, Bogdan M and Kübler A 2012 Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a p300-BCI *Clin. Neurophysiol.* **123** 1328–37
[61] Krusienski D J, Sellers E W, McFarland D J, Vaughan T M and Wolpaw J R 2008 Toward enhanced p300 speller performance *J. Neurosci. Methods* **167** 15–21
[62] Toro C, Deuschl G, Thatcher R, Sato S, Kufta C and Hallett M 1994 Event-related desynchronization and movement-related cortical potentials on the ECoG and EEG *Electroencephalogr. Clin. Neurophysiol. Evoked Potentials Sect.* **93** 380–9
[63] Pfurtscheller G and Aranibar A 1977 Event-related cortical desynchronization detected by power measurements of scalp EEG *Electroencephalogr. Clin. Neurophysiol.* **42** 817–26
[64] Pfurtscheller G and da Silva F L 1999 Event-related eeg/meg synchronization and desynchronization: basic principles *Clin. Neurophysiol.* **110** 1842–57
[65] Liao K, Xiao R, Gonzalez J and Ding L 2014 Decoding individual finger movements from one hand using human EEG signals *PLoS ONE* **9** 1–12
[66] Barrett G, Shibasaki H and Neshige R 1986 Cortical potentials preceding voluntary movement: evidence for three periods of preparation in man *Electroencephalogr. Clin. Neurophysiol.* **63** 327–39
[67] Yilmaz O, Birbaumer N and Ramos-Murguialday A 2015 Movement related slow cortical potentials in severely paralyzed chronic stroke patients *Frontiers Hum. Neurosci.* **8** 1033
[68] Deecke L, Scheid P and Kornhuber H H 1969 Distribution of readiness potential, pre-motion positivity, and motor potential of the human cerebral cortex preceding voluntary finger movements *Exp. Brain Res.* **7** 158–68
[69] Leuthardt E C, Schalk G, Moran D and Ojemann J G 2006 The emerging world of motor neuroprosthetics: a neurosurgical perspective *Neurosurgery* **59** 1–14
[70] Yom-Tov E and Inbar G F 2003 Detection of movement-related potentials from the electro-encephalogram for possible use in a brain–computer interface *Med. Biol. Eng. Comput.* **41** 85–93
[71] Karimi F, Kofman J, Mrachacz-Kersting N, Farina D and Jiang N 2017 Detection of movement related cortical potentials from EEG using constrained ICA for brain–computer interface applications *Frontiers Neurosci.* **11** 356
[72] Gordon S, Lawhern V, Passaro A and McDowell K 2015 Informed decomposition of electroencephalographic data *J. Neurosci. Methods* **256** 41–55
[73] Bigdely-Shamlo N, Mullen T, Kothe C, Su K M and Robbins K A 2015 The prep pipeline: standardized preprocessing for large-scale EEG analysis *Frontiers Neuroinformatics* **9** 16
[74] Tangermann M *et al* 2012 Review of the BCI competition iv *Frontiers Neurosci.* **6** 55
[75] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)
[76] Abadi M *et al* 2016 Tensorflow: a system for large-scale machine learning *Proc. 12th USENIX Conf. on Operating Systems Design and Implementation* (Berkeley, CA, USA: USENIX Association) pp 265–83
[77] Chollet F 2015 Keras https://github.com/fchollet/keras
[78] Ang K K, Chin Z Y, Wang C, Guan C and Zhang H 2012 Filter bank common spatial pattern algorithm on BCI competition iv datasets 2a and 2b *Frontiers Neurosci.* **6** 39

[79] Dyrholm M, Christoforou C and Parra L C 2007 Bilinear discriminant component analysis *J. Mach. Learn. Res.* **8** 1097–111
[80] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift (arXiv:1502.03167)
[81] Clevert D, Unterthiner T and Hochreiter S 2015 Fast and accurate deep network learning by exponential linear units (elus) *CoRR* (arXiv:1511.07289)
[82] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
[83] Springenberg J T, Dosovitskiy A, Brox T and Riedmiller M A 2014 Striving for simplicity: the all convolutional net (arXiv:1412.6806)
[84] Rivet B, Souloumiac A, Attina V and Gibert G 2009 xDAWN algorithm to enhance evoked potentials: application to brain–computer interface *IEEE Trans. Biomed. Eng.* **56** 2035–43
[85] Barachant A, Bonnet S, Congedo M and Jutten C 2012 Multiclass brain–computer interface classification by Riemannian geometry *IEEE Trans. Biomed. Eng.* **59** 920–8
[86] Barachant A and Congedo M 2014 A plug & play P300 BCI using information geometry (arXiv:1409.0107 [cs, stat])
[87] Congedo M, Barachant A and Andreev A 2013 A new generation of brain–computer interface based on riemannian geometry *CoRR* (arXiv:1310.8115)
[88] Barachant A and Bonnet S 2011 Channel selection procedure using riemannian distance for bci applications *5th Int. IEEE/EMBS Conf. on Neural Engineering* pp 348–51
[89] Barachant A, Bonnet S, Congedo M and Jutten C 2013 Classification of covariance matrices using a Riemannian-based kernel for BCI applications *Neurocomputing* **112** 172–8
[90] Ng A Y 2004 Feature selection, l1 versus l2 regularization, and rotational invariance *Proc. 21st Int. Conf. on Machine Learning* (New York, NY: ACM) p 78
[91] Ledoit O and Wolf M 2004 A well-conditioned estimator for large-dimensional covariance matrices *J. Multivariate Anal.* **88** 365–411
[92] Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *J. R. Stat. Soc.* B **67** 301–20
[93] Kothe C A and Makeig S 2013 Bcilab: a platform for brain–computer interface development *J. Neural Eng.* **10** 056014
[94] Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K and MÃžller K-R 2010 How to explain individual classification decisions *J. Mach. Learn. Res.* **11** 1803–31
[95] Zeiler M D and Fergus R 2014 Visualizing and understanding convolutional networks *Computer Vision—ECCV* ed D Fleet *et al* (Cham: Springer) pp 818–33
[96] Nguyen A M, Yosinski J and Clune J 2014 Deep neural networks are easily fooled: high confidence predictions for unrecognizable images *CoRR* (arXiv:1412.1897)
[97] Ribeiro M T, Singh S and Guestrin C 2016 'Why should I trust you?': explaining the predictions of any classifier In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (New York, NY: ACM) pp 1135–44
[98] Shrikumar A, Greenside P and Kundaje A 2017 Learning important features through propagating activation differences *CoRR* (arXiv:1704.02685)
[99] Ancona M, Ceolini E, Öztireli C and Gross M 2018 Towards better understanding of gradient-based attribution methods for deep neural networks *Int. Conf. on Learning Representations*
[100] Montavon G, Samek W and Müller K-R 2018 Methods for interpreting and understanding deep neural networks *Digit. Signal Process.* **73** 1–15
[101] Torrence C and Compo G P 1998 A practical guide to wavelet analysis *Bull. Am. Meteorol. Soc.* **79** 61–78
[102] Mazaheri A and Picton T W 2005 EEG spectral dynamics during discrimination of auditory and visual targets *Cogn. Brain Res.* **24** 81–96
[103] Johnson G, Waytowich N and Krusienski D J 2011 The challenges of using scalp-EEG input signals for continuous device control *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems. Int. Conf. on Foundations of Augmented Cognition* ed D D Schmorrow and C M Fidopiastis (Berlin: Springer) pp 525–7
[104] Lawhern V, Hairston W D, McDowell K, Westerfield M and Robbins K 2012 Detection and classification of subject-generated artifacts in EEG signals using autoregressive models *J. Neurosci. Methods* **208** 181–9