Post-hoc Labeling of Arbitrary M/EEG Recordings for Data-Efficient Evaluation of Neural Decoding Methods

Sebastián Castaño-Candamil[1]*, Andreas Meinel[1] and Michael Tangermann[1,2]*

[1] Brain State Decoding Lab, Department of Computer Science and BrainLinks-BrainTools Cluster of Excellence, University of Freiburg, Freiburg, Germany, [2] Autonomous Intelligent Systems, Department of Computer Science, University of Freiburg, Freiburg, Germany

Many cognitive, sensory and motor processes have correlates in oscillatory neural source activity, which is embedded as a subspace in the recorded brain signals. Decoding such processes from noisy magnetoencephalogram/electroencephalogram (M/EEG) signals usually requires data-driven analysis methods. The objective evaluation of such decoding algorithms on experimental raw signals, however, is a challenge: the amount of available M/EEG data typically is limited, labels can be unreliable, and raw signals often are contaminated with artifacts. To overcome some of these problems, simulation frameworks have been introduced which support the development of data-driven decoding algorithms and their benchmarking. For generating artificial brain signals, however, most of the existing frameworks make strong and partially unrealistic assumptions about brain activity. This limits the generalization of results observed in the simulation to real-world scenarios. In the present contribution, we show how to overcome several shortcomings of existing simulation frameworks. We propose a versatile alternative, which allows for an objective evaluation and benchmarking of novel decoding algorithms using real neural signals. It allows to generate comparatively large datasets with labels being deterministically recoverable from the arbitrary M/EEG recordings. A novel idea to generate these labels is central to this framework: we determine a subspace of the true M/EEG recordings and utilize it to derive novel labels. These labels contain realistic information about the oscillatory activity of some underlying neural sources. For two categories of subspace-defining methods, we showcase how such labels can be obtained—either by an exclusively data-driven approach (independent component analysis—ICA), or by a method exploiting additional anatomical constraints (minimum norm estimates—MNE). We term our framework *post-hoc labeling* of M/EEG recordings. To support the adoption of the framework by practitioners, we have exemplified its use by benchmarking three standard decoding methods—i.e., common spatial patterns (CSP), source power-comodulation (SPoC), and convolutional neural networks (ConvNets)—wrt. Varied dataset sizes, label noise, and label variability. Source code and data are made available to the reader for facilitating the application of our *post-hoc* labeling framework.

Keywords: neural decoding, M/EEG labeling, data-driven neural decoding, brain computer interfaces, band-power decoding

1. INTRODUCTION

Brain oscillatory phenomena measured with non-invasive imaging techniques, such as magneto- or electroencephalography (M/EEG), contain information about underlying neural processes. The possibility to describe these is a prerequisite to answer questions appearing in clinical contexts, as well as in basic neuroscience research. Examples are the monitoring of rehabilitation progress, the characterization of neurological and neuropsychiatric disorders (Herrmann and Demiralp, 2005), the investigation of memory processes (Klimesch et al., 2007), motor performance (Tangermann et al., 2015; Meinel et al., 2016), and visual perception (Marshall et al., 2018).

1.1. Benchmarking and Validation of Data-Driven Neural Decoding Algorithms

Since the rise of brain-computer interface (BCI) systems, great effort has been put into developing novel techniques for decoding neural sources from noisy M/EEG recordings using linear and nonlinear methods, both for classification and regression tasks (Lotte et al., 2018). For the development, validation, and benchmarking of such neural decoding algorithms, it is desirable to have multichannel datasets with large amounts of labeled data available. In the literature, two types of frameworks prevail. First, frameworks making use of real M/EEG recordings acquired during experimental sessions, and second, those using synthetically generated pseudo-M/EEG signals. Each comes with advantages and shortcomings, as explained below.

1.1.1. Real M/EEG Recordings

Using recordings of M/EEG data has the great advantage that their dynamics, the signal-to-noise ratio between oscillatory sources of interest and undesired background activity as well as any non-stationary behavior over time are provided naturally, such that they do not need to be set by the experimenter. Efforts have been made to provide benchmarking platforms, e.g., Jayaram and Barachant (2018) which contains real M/EEG data that has been recorded under specific experimental paradigms. However, the limitations of using such paradigm-specific M/EEG data remains.

1.1.1.1. Small Datasets

The amount of labeled real M/EEG data acquired in a single experimental session maximally lasts a couple of hours. This limited dataset size is rendered even smaller by subsequent data preprocessing steps, i.e., data segmentation, removal of inter-trial pauses and rejection of artifactual segments. Bigger datasets may be obtained by applying transfer learning techniques, with the aim of merging inter-subject and inter-session data (Krauledat et al., 2008). However, this comes with its own substantial challenges and is subject to active research (Jayaram et al., 2016; Lotte et al., 2018). Overall, the relative small dataset size is a clear drawback of using real M/EEG data for the benchmarking of algorithms.

1.1.1.2. Label Noise

In some experimental setups, M/EEG recordings are governed by a varying but known experimental parameter—such as the intensity of an external stimulus (Dähne et al., 2014b)—this parameter can be used as a target variable z, which serves as epoch-wise labels to support the supervised decoding of correlated oscillatory M/EEG activity. Unfortunately, the situation is more challenging, if an M/EEG correlate of open behavior or even an imagery task shall be decoded: a *reliable* behavioral surrogate is lacking, and the precise registration of open behavior may also be difficult, thus, the investigator may end up with a noisy estimate. This label noise can have many different origins: subjects may be unable to execute the task with a required timing, they may not follow the experimental instructions consistently, may change their mental strategies to solve a problem, or display varying levels of engagement over time. Compared to clean labels z, noisy label information is known to decrease the performance of decoding algorithms (Castaño-Candamil et al., 2015b). A number of decoding tasks, like the estimation of the motor tasks in imagery experiments (Höhne et al., 2014) in BCI, the prediction of hand motor performance (Meinel et al., 2016), or attention decoding (Martel et al., 2014), are considered very challenging, with label noise as a substantial part of the problem. As the experimenter typically neither knows the level of label noise contained in z nor can control it, behavioral experiments deliver suboptimal data for the benchmarking of decoding algorithms.

1.1.1.3. Task-Specific

Last but not least, the use of real M/EEG data for benchmarking comes with the drawback that switching between decoding approaches, e.g., classification and regression, may require to redesign M/EEG experiments and run them again to collect the necessary novel label types.

1.1.2. Synthetic Pseudo-M/EEG Signals

Some shortcomings of real data can be mitigated with synthetically generated M/EEG signals (Krol et al., 2018), which are utilized preferably in the fields of brain mapping and connectivity analysis (Castaño-Candamil et al., 2015a; Haufe and Ewald, 2016). Here, the assumption of a linear mapping from the neural source space to the M/EEG sensor space allows to simulate a neural target source, whose activity overlaps with measurement noise and task-irrelevant brain activity termed *background sources*. Special attention is dedicated to the modeling of sources, such that they match naturally occurring frequency spectra, e.g., reproducing a $1/f$-shaped frequency spectrum and a narrow-band oscillatory target source. However, since these simulations are based purely on synthetic data, they need to make strong assumptions about brain dynamics.

1.1.2.1. Assumptions About Neural Dynamics

The assumptions made are expressed by the choices of, e.g., the power ratio between target- and background sources, the noise level on the sensor space, and the time series of the sources. Synthetic datasets typically disregard more complex dynamics, which are present in real datasets and pose substantial challenges

for decoding methods. While being sufficient for proof-of-concept purposes (Dähne et al., 2014b; Lindgren et al., 2018), these purely synthetic datasets lack a sufficient level of realism to allow for generalizing simulated performance estimates to real-world scenarios.

1.1.2.2. Physiologically- and Functionally-Motivated Models

Simulated M/EEG time series are also used intensively in the field of computational neuroscience. Here, physiologically motivated linear and nonlinear stochastic systems are utilized to describe e.g., the dynamics of Alzheimer's disease, epilepsy, or sleeping disorders (Robinson et al., 2002; Kim et al., 2007) on the level of small networks and M/EEG. The complexity of such methods span from linear univariate, to highly detailed multivariate models motivated by complex functional and physiological constraints. Univariate linear models disregard the notion of spatiality inherent to M/EEG recordings and resemble simple oscillators. Multivariate models—at the other end of the complexity spectrum—account for highly specific networks and dynamics in the brain, but require control over a large number of parameters (Breakspear et al., 2010). To determine them based on data is difficult and may succeed only when very large data collections are accessible. In contrast to the physiologically motivated systems, purely data-driven approaches using recurrent neural networks methods (Forney et al., 2015) have been explored. These approaches are capable of generating a single channel of realistic, albeit artificial, M/EEG time-series but also disregard the spatial notion present in M/EEG recordings.

1.2. *Post-hoc* Labeling of Paradigm-Agnostic M/EEG Recordings

Motivated by the shortcomings of using real M/EEG recordings (few data and noisy labels) as well as of synthetically generated datasets (questionable assumptions about neural dynamics and noise), we propose a novel generation framework for labeled datasets. It is based on *post-hoc* labeling of pre-recorded real M/EEG signals, that generates novel labels using unsupervised subspace projection methods. As the original labels of the dataset are discarded, the framework is agnostic wrt. the original paradigm under which the M/EEG signals had been recorded, and to its original trial structure.

As a result, with our framework we aim at obtaining datasets with the following properties:

- Real neural dynamics, as contained in the M/EEG signals;
- highly efficient use of real M/EEG data (thus potentially yielding larger datasets), and
- labels deterministically recoverable from the available data, i.e., free of noise.

Our proposed framework is compatible with existing benchmarking frameworks, for instance the Mother of All BCI Benchmarks (MOABB) introduced in Jayaram and Barachant (2018).

2. METHODS

2.1. Generative Model of Brain Activity

Neural activity recorded by M/EEG can be represented by means of a linear forward model (Baillet et al., 2001; Grech et al., 2008):

$$X = A S + E \,, \tag{1}$$

where $X \in \mathcal{C} \subseteq \mathbb{R}^{N_c \times N_t}$ is a multivariate signal in the channel space $\mathcal{C}$ describing M/EEG data measured by N_c M/EEG channels at N_t discrete time samples, $S \in \mathcal{S} \subseteq \mathbb{R}^{N_s \times N_t}$ describes the time course of N_s neural sources in the source space $\mathcal{S}$ with covariance matrix $Q \in \mathbb{R}^{N_s \times N_s}$; and matrix $A \in \mathbb{R}^{N_c \times N_s}$ describes the linear projection $\mathcal{S} \to \mathcal{C}$ of the sources onto the sensor space, where the columns of A, $a \in \mathbb{R}^{N_c}$, are referred to as *spatial patterns*. Furthermore, the matrix E contains i.i.d. Gaussian noise with zero mean and a covariance matrix $Q_\epsilon \in \mathbb{R}^{N_c \times N_c}$.

Under this representation, it is widely accepted that surrogates of a wide range of cognitive processes can be decoded from the power of narrowband frequency oscillatory sources in S (Dähne et al., 2014b; Horschig et al., 2014). We will represent such a surrogate by the row vector $s_z^\mathsf{T} \in \mathbb{R}^{N_t}$ of S, whereas its envelope—the power of the source—will be denoted as $z \in \mathbb{R}^{N_t}$ and termed *target variable*. It delivers the labels and, consequently, represents the variable that is to be decoded for unseen data.

2.2. *Post-hoc* Labeling of Paradigm-Agnostic M/EEG Recordings

Our novel framework refrains from making (potentially problematic) assumptions about the dynamics of neural activity or about the signal-to-noise ratio between an oscillatory source of interest and background sources. The framework relies upon an unsupervised projection of an arbitrary M/EEG dataset X onto a source space $\mathcal{S}$ by means of a function $f : \mathcal{C} \to \mathcal{S}$.

Assuming we can find such a function which decomposes the M/EEG signals into *reasonable* sources (the next paragraphs will deal with this), we also propose that any source in $\mathcal{S}$ could be selected to serve as the target source s_z and that the oscillatory power of this source can be used to provide the labels z for the purpose of benchmarking arbitrary decoding methods.

2.2.1. Determining fs a Linear Projection Function

We propose two alternative strategies to choose function f: the first one uses an anatomically constrained source space $\mathcal{S}_a \subseteq \mathcal{S}$ while the second strategy defines the source space $\mathcal{S}_d \subseteq \mathcal{S}$ in a purely data-driven manner.

2.2.1.1. Anatomically Constrained Source Space $\mathcal{S}_a$

If an anatomically motivated head model A, potentially containing a very large number of sources, is available (Hallez et al., 2007), then f can be selected such that X is projected onto an anatomically constrained version of the source space, $\mathcal{S}_a$. To this end, a source reconstruction method may be used. Specifically, the maximum a-posteriori estimate of $S \in \mathcal{S}_a$ can be found as the minimizer of the following cost function (Grech et al., 2008; Castaño-Candamil et al., 2015a):

$$\underset{S}{\mathrm{argmin}} \{ ||X - AS||_Q^2 + \lambda \Theta (S) \} \,. \tag{2}$$

Here, $\|\cdot\|_{Q_\epsilon}$ is the matrix norm of the argument wrt. the covariance matrix Q_ϵ, and $\lambda \in \mathbb{R}^+$ is a regularization constant. The penalty term $\Theta(S) : S \mapsto \mathbb{R}^+$ can be utilized to formalize arbitrary constraints imposed upon the neural source activity. Many different algorithms, each with specific choices for $\Theta(S)$ and Q_ϵ, have been introduced (Grech et al., 2008), each of them representing different priors about the expected characteristics of sources. For the sake of simplicity and assuming stationarity wrt. Q and Q_ϵ, we have chosen $\Theta(S) = \|S\|_2^2$ and $Q_\epsilon = I_{N_c}$, where $I_{N_c} \in \mathbb{R}^{N_c \times N_c}$ is an identity matrix. This approach is commonly termed ℓ_2-norm regularization (Ng, 2004), also known as minimum norm estimate (MNE) (Pascual-Marqui, 1999; Grech et al., 2008). For this choice of Θ, it can be shown that the optimal solution for expression 2 is given conveniently by

$$S = A^\mathsf{T}(I_{N_c} + \lambda A A^\mathsf{T})^{-1} X \qquad (3)$$

where hyperparameter λ is determined, in our case, using the generalized crossvalidation procedure, as described in Grech et al. (2008).

Please note that the proposed *post-hoc* labeling framework is not limited to using MNE, and therefore, the assumption about stationary dynamics made by MNE is an external factor and is not intrinsically embedded in the proposed framework. If non-stationary dynamics in the underlying sources shall be taken into account, more complex mapping methods, as time-frequency mixed norm estimates (TF-MxNE) (Gramfort et al., 2013b) or spatio-temporal unifying tomography (STOUT) (Castaño-Candamil et al., 2015a) can be used to obtain S in $\mathcal{S}_a$.

2.2.1.2. Data-Driven Source Space $\mathcal{S}_d$
Choosing a data-driven approach, a set of underlying target sources can be estimated from X using standard unsupervised linear decomposition methods such as PCA, ICA, factor analysis, among others. In the following, we use the fastICA algorithm (Hyvärinen and Oja, 2000), which—among the different blind source separation methods—has been widely employed for the analysis of neural data (Makeig et al., 1996; Vigário et al., 2000; Delorme and Makeig, 2004). For this choice, the function f is defined as $S = f(X) = \Phi X$, where $\Phi \in \mathbb{R}^{N_s \times N_c}$ is a matrix spanning $\mathcal{S}_d$, a space of maximally independent components. Independence is achieved by maximizing non-Gaussianity of the sources $S \in \mathcal{S}_d$. Please note again, that the proposed dataset generation framework is not dependent on this specific choice of fastICA, and therefore, other blind source separation methods (including adaptive approaches, capable of dealing with non-stationary dynamics), can be used for estimating S in $\mathcal{S}_d$.

2.2.2. Extraction of Target Variable z
Once a set of sources has been determined by the two approaches mentioned above (or any analog thereof), a target source s_z is to be selected. The choice could be guided by a prior about the benchmarking problem (e.g., strong components only, or components that stem from a brain region known to be involved in a certain experimental task) or simply by random selection. Similarly, a arbitrary combination of sources could also be selected as target source. In this regard, our *post-hoc* labeling

framework offers absolute flexibility regarding the criteria used for obtaining the target source. After it has been selected, the labels z are computed in the following three-step procedure:

1. Since it is expected that the target source provides surrogate information about a cognitive process by means of its power in a narrow frequency band (Horschig et al., 2014), X is filtered with a bandpass filter to reflect this assumption and then projected onto $\mathcal{S}$, where a target source s_z is selected.
2. The band power envelope of the selected source is determined by computing, e.g., the magnitude of its Hilbert transform
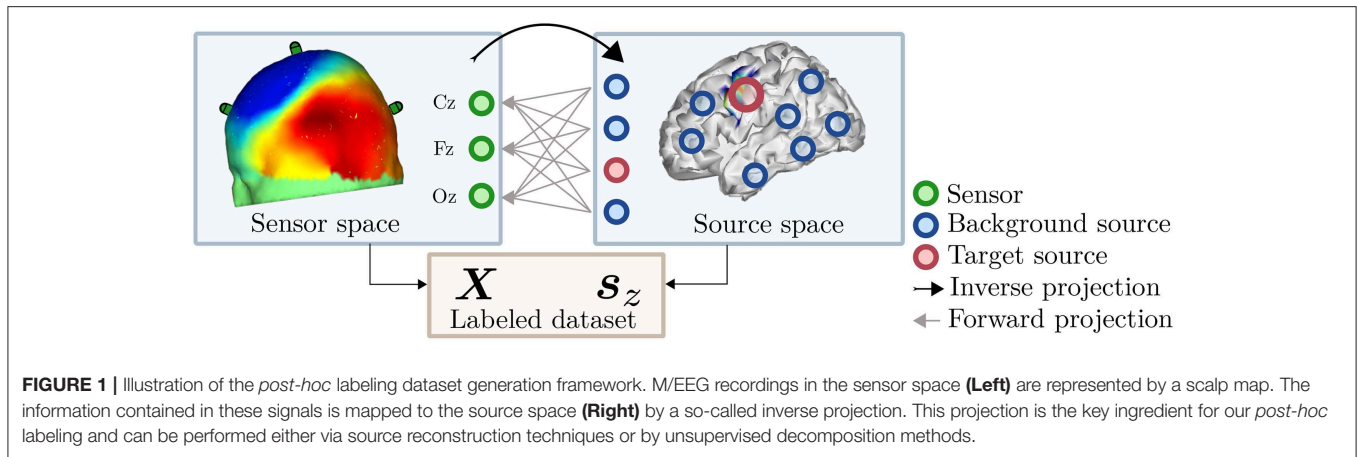
$$z = |\mathcal{H}\{s_z\}| \qquad (4)$$

3. If desired, the data and the labels can be segmented into epochs. It is important to remark that this step depends on the algorithm which shall be benchmarked later on.

The dataset resulting from these steps consists of real EEG recordings X and a noiseless continuous variable z containing the corresponding target labels. Using X and z, any *arbitrary* supervised decoding algorithm can be benchmarked and validated. Furthermore, the proposed formulation may be extended to obtain *discrete* labels y by assigning a class label depending on percentile memberships of individual labels z_i, thus extending the applicability of our proposed framework to classification tasks. **Figure 1** illustrates the general idea of the proposed *post-hoc* labeling framework for datasets generation.

3. EXPERIMENTAL SETUP

3.1. Decoding Methods
Among many different neural decoding methods found in the literature, linear subspace decomposition methods come with the advantage of computational simplicity and offer interpretability of the decoded information (Haufe et al., 2014b). Besides extracting label-informative oscillatory components, these approaches are often put to work for dimensionality reduction before applying more advanced processing methods. The pioneering work on joint covariance diagonalization presented by Fukunaga (2013) and reformulated by de Cheveigné and Parra (2014) serves as a generalized foundation for popular supervised linear subspace decomposition algorithms. One representative is the common spatial patterns (CSP) algorithm, suitable for the classification of oscillatory processes related to motor execution, motor imagery, and attempted motor execution tasks (Koles et al., 1990; Lemm et al., 2005). The relevance of CSP is not only indicated by its extensive use (Tangermann et al., 2012), but also by the plethora of derivatives that have been introduced after its original presentation, like finite impulse response CSP (FIR-CSP) (Higashi and Tanaka, 2013), sub-band CSP (SBCSP) (Novi et al., 2007), filter-bank CSP (FBCSP) (Ang et al., 2008), spectrally weighted-CSP (SPEC-CSP) (Tomioka et al., 2006), among others. While CSP is a supervised algorithm preferred for classification problems, the source power comodulation algorithm SPoC (Dähne et al., 2014b), together with its extensions canonical SPoC (cSPoC) (Dähne et al., 2014c) and multimodal SPoC (mSPoC) (Dähne

FIGURE 1 | Illustration of the *post-hoc* labeling dataset generation framework. M/EEG recordings in the sensor space **(Left)** are represented by a scalp map. The information contained in these signals is mapped to the source space **(Right)** by a so-called inverse projection. This projection is the key ingredient for our *post-hoc* labeling and can be performed either via source reconstruction techniques or by unsupervised decomposition methods.

et al., 2014a), lends itself to solve supervised linear regression tasks (Meinel et al., 2016). Unsupervised linear neural decoding methods are also extremely popular: After principal component analysis (PCA), the most widely applied class of algorithms is probably the family of independent component analysis (ICA) methods, which realize blind source separation (Makeig et al., 1996). Last but not least, in the context of unsupervised extraction of specific oscillatory components, the spatio-spectral decomposition (SSD) approach introduced by Haufe et al. (2014a) deserves to be mentioned.

In addition to these linear subspace methods, nonlinear decoding methods have been introduced. A recent example are convolutional neural networks (ConvNets) used for, e.g., the classification of motor tasks (Schirrmeister et al., 2017), visually evoked potentials (Lawhern et al., 2016), error-related negativity responses, movement-related cortical potentials, and sensory motor rhythms. Further decoding approaches which make use of various machine learning models have been described in the literature (see the review provided in Schirrmeister et al., 2017; Lotte et al., 2018).

We implemented three different decoding methods to exemplify the application of a our *post-hoc*-labeling framework. Specifically, we will report on two classification tasks (using CSP and ConvNets for a two- and three class classification problem, respectively) and SPoC for a regression task, covering a wide range of popular decoding algorithms. We will benchmark these methods wrt. several dataset-inherent parameters, i.e., dataset size, label noise level, and variance of the target label z. The performance of these decoding methods is evaluated using arbitrarily selected accuracy metrics for the sake of illustration, i.e., AUC, classification performance, or linear correlation coefficient.

3.1.1. Common Spatial Patterns (CSP)

The CSP algorithm is an established supervised method in the BCI community, used in classification tasks for constructing a set of N_c spatial filters $\boldsymbol{W}_{CSP} \in \mathbb{R}^{N_c \times N_c}$ that optimally discriminates epochs of two classes characterized by differing band-power features, where the labels are defined as $\boldsymbol{y} \in \{1, 2\}$, corresponding to a discretization of the continuous label z.

Assuming that M/EEG data $\boldsymbol{X}$ have been bandpass filtered to the frequency band of interest and segmented into a set of N epochs, where $\boldsymbol{X}(e)$ represent the e-th epoch of the M/EEG data, the CSP objective function is mathematically formalized as

$$\underset{\boldsymbol{W}_{CSP}}{\arg\max} \frac{\boldsymbol{W}_{CSP}^{\mathsf{T}} (\boldsymbol{C}_1 - \boldsymbol{C}_2) \, \boldsymbol{W}_{CSP}}{\boldsymbol{W}_{CSP}^{\mathsf{T}} \, \boldsymbol{C} \, \boldsymbol{W}_{CSP}}. \tag{5}$$

with the spatial covariance matrices of classes one and two defined as $\boldsymbol{C}_{1,2} = N_{1,2}^{-1} \sum_e^N \boldsymbol{X}_{1,2}(e) \boldsymbol{X}_{1,2}(e)^{\mathsf{T}}$ and with $\boldsymbol{C}$ being the pooled spatial covariance matrix. It can be shown that a solution to the CSP optimization problem can be found by solving the generalized eigenvalue problem

$$\boldsymbol{W}_{CSP}^{\mathsf{T}} (\boldsymbol{C}_1 - \boldsymbol{C}_2) = \boldsymbol{\Lambda} \, \boldsymbol{W}_{CSP}^{\mathsf{T}} \boldsymbol{C} \tag{6}$$

with $\boldsymbol{W}_{CSP}$ being a matrix containing (column-wise) the eigenvectors (i.e., spatial filters) which are related to the eigenvalues provided by the entries of the main diagonal of $\boldsymbol{\Lambda} \in \mathbb{R}^{N_f \times N_f}$.

For our tests, we reduced the full filter matrix to $\widehat{\boldsymbol{W}}_{CSP} \in \mathbb{R}^{N_c \times 2}$. Thus, it contains two eigenvectors only, one corresponding to the largest and one to the smallest eigenvalues in $\boldsymbol{\Lambda}$, representing each class, respectively. The selection of the number of CSP filters is an important hyperparameter for obtaining an optimal decoding performance. However, according to Blankertz et al. (2011), a good rule of thumb indicates that between 2 and 8 filters are likely to deliver a good performance.

Note that a spatial filters derived by CSP does not deliver an estimate of the target variable $\widehat{z}(e)$ yet. To derive estimates of the target variable, we thus trained a regularized LDA (rLDA) classifier on the power features (delivered by $\Theta_x(e) = \text{var}[\hat{\boldsymbol{s}}](e)$) of the spatially filtered data $\hat{\boldsymbol{s}}(e) = \widehat{\boldsymbol{W}}_{CSP}^{\mathsf{T}} \boldsymbol{X}(e)$ (Blankertz et al., 2011).

The results reported here were computed using the CSP implementation provided in the MNE toolbox by Gramfort et al. (2013a, 2014).

3.1.1.1. Decoding Accuracy

Depending on the application, an arbitrary metric can be used, which matches well with the given decoding method, e.g., classification accuracy, Type I/II errors, etc. To characterize CSP performance we chose the area ander the ROC curve (AUC).

3.1.2. Source Power Comodulation (SPoC)

Analogously to CSP, the multivariate neural decoding method SPoC (Dähne et al., 2014b) utilizes a supervised regression approach in order to estimate a spatial filter $w_{SPoC} \in \mathbb{R}^{N_c}$, onto which X will be linearly projected to extract the underlying *continuous* target source z.

Specifically, a SPoC spatial filter $w_{SPoC} \in \mathbb{R}^{N_c}$ is optimized such that the power of a projected epoch, $\Theta_x(e) = \text{var}[\hat{s}](e)$, maximally covaries with the target variable $z(e)$:

$$\underset{w}{\text{argmax}} \ \{\text{cov}\left[\Theta_x(e), z(e)\right]\} \ \forall \ e \qquad (7)$$

It can be shown that solving this optimization problem is equivalent to solving the generalized eigenvalue problem

$$C_z w_{SPoC} = \lambda C w_{SPoC} \qquad (8)$$

where $C = N^{-1} \sum_e^N X(e)X(e)^\top$ is the spatial covariance matrix of X and $C_z = N^{-1} \sum_e^N z(e)X(e)X(e)^\top$ is the epoch-wise z-weighted spatial covariance matrix.

Given a spatial filter w_{SPoC}, the target variable z can subsequently be estimated as $\hat{z}$ for each single epoch of unseen test data X_{te} via $\hat{z}(e) = \text{var}[w_{SPoC}^\top X_{te}(e)]$.

The results reported here were computed using the SPoC implementation provided in the MNE toolbox by Gramfort et al. (2013a, 2014).

3.1.2.1. Decoding Accuracy

In this specific scenario, the accuracy of the decoding provided by SPoC is assessed in terms of the linear correlation ρ between the estimated target labels $\hat{z}$ and the true labels z.

3.1.3. Convolutional Neural Network (ConvNet)

Finally, the third use-case for our *post-hoc* labeling framework shall be provided by a ConvNet as proposed and implemented by Schirrmeister et al. (2017). ConvNets provide an end-to-end decoding of raw EEG signals and thus may be a good method to chose when prior knowledge about relevant EEG features is missing. Specifically, we utilized a *shallow* ConvNet architecture. It focuses on both temporal and spatial convolutions and thus has the capacity to detect features in both time and spatial domains, similarly to features extracted by filters derived by CSP and SPoC. Unlike CSP and SPoC, however, the input representation of the EEG signals to the ConvNet does not assume any type of frequency pre-filtering. Instead, it consists of the N_c channels and N_t time points of the (epoched) raw EEG.

3.1.3.1. ConvNet Architecture

As shown in **Figure 2**, the temporal convolution step made use of a kernel size of 25 samples, containing 40 neural units. Subsequently, a layer with 40 units performed a spatial convolution step on all the channels. Finally, a log-power computation precedes a mean pooling stage and a fully connected layer with 3 units (softmax activation), one for each class.

3.1.3.2. Decoding Accuracy

We chose classification accuracy as evaluation metric for the shallow ConvNet. Nevertheless, as in CSP and SPoC, any given metric might have been utilized for this purpose, depending on the focus of the analysis.
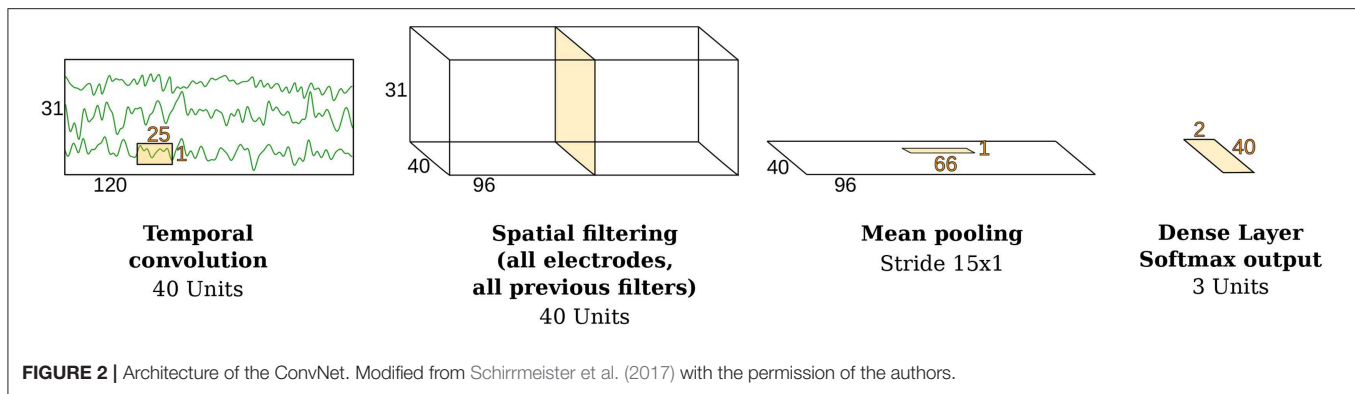
3.2. The EEG Dataset

3.2.1. Signal Acquisition

The EEG signals for our use-case were recorded from seven healthy subjects (three females) with a mean age of 28 years. Seventy three minutes of EEG data on average were recorded in a single session while subjects sat calmly in front of a computer screen and performed the sequential visual isometric pinch task (SVIPT) (Reis et al., 2009). Given the paradigm-agnostic character of the *post-hoc* labeling framework, details about the SVIPT paradigm remain outside the scope of this paper but can be consulted in Meinel et al. (2016). EEG signals were recorded from 31 passive Ag/AgCl electrodes (EasyCap GmbH, Germany) placed according to the extended 10-20 system. Impedances were kept below 20 kΩ. All channels were referenced against the nose at recording time and were re-referenced against the EEG common average during the *post-hoc* analysis. The EEG signals were registered by BrainAmp DC amplifiers (Brain Products GmbH, Germany) at a sampling rate of 1 kHz, with an analog lowpass filter of 250 Hz applied before digitization.

3.2.2. Pre-processing Only for *post-hoc* Labeling

All processing steps necessary to generate the *post-hoc* labeled datasets were performed in Matlab using the BBCI toolbox (Blankertz et al., 2016). The EEG signals were bandpass filtered between 0.2 and 48 Hz with a 5th order Butterworth filter and then sub-sampled to 120 Hz. Assuming the alpha band being in the focus of a benchmarking scenario, EEG data were further filtered with a 5th order Butterworth bandpass filter with cut-off frequencies at 8 and 12 Hz. This target frequency band can be modified according to different analysis goals, but for the sake of compactness in the use-case analysis, we have kept this parameter fixed. Finally, the label extraction procedure described in section 2.2.2 was applied, in order to obtain a labeled dataset comprised by X and z.

For the generation of the datasets containing anatomical constraints on the sources, the publicly available *New York Head* (Huang et al., 2016) was used. It describes a finite element model containing 2,000 sources located on the cortical surface. These sources were subsampled from a highly detailed model containing 74,382 sources, which had been computed from a non-linear average of 152 human brains. The *New York Head* takes scalp, skull, cerebro-spinal fluid, gray matter, and white matter into account. Sources were assumed to be perpendicularly oriented wrt. the cortical surface, however, our framework could also be used with models that allow for free source orientation.

FIGURE 2 | Architecture of the ConvNet. Modified from Schirrmeister et al. (2017) with the permission of the authors.

On the other hand, for the data-driven approach, a fixed number of 20 ICA components were extracted for each subject. Ideally, only components corresponding to actual neural sources should be selected for further analysis. For the identification of such neural components, the multiple artifact rejection algorithm (MARA) (Winkler et al., 2011) was applied, using a posterior probability threshold of 10^{-8} for components classified as having neural origin.

3.2.3. Pre-processing for Algorithm Benchmarking
The pre-processing pipeline we chose to apply on the newly labeled data was selected to match the requirements of the decoding methods presented as use-cases; consequently, it is independent of the *post-hoc* dataset generation framework.

For outlier detection, the continuous EEG data X were bandpass filtered between 0.7 and 25 Hz with a 5th order Butterworth filter. Segments of the continuous data with peak-to-peak amplitude exceeding 80 μV were marked as artifactual for later removal in the pre-processing pipeline. Only for CSP and SPoC, the original continuous data X, was filtered by a 5th order Butterworth filter to the band of 8–12 Hz. For ConvNets, the original raw data was used. Then, EEG data and the target source z were segmented in non-overlapping windows of 1 s duration. At this point, epochs marked earlier as artifactual were removed. For the remaining segments, the epoch-wise average power of z was extracted and used as the target variable to train the decoding algorithms. For CSP and ConvNet, epoch-wise discrete labels were generated. For the binary classification tasks they were determined by the top and bottom 50th percentile, whereas the 33th and 66th percentiles limits defined the three-class problem. At this point we want to point out, that the epoching does not necessarily need to obey the original time structure of the experimental paradigm, under which X was recorded.

3.3. Sweep Over Three Dataset-Inherent Parameters
The practitioner will probably strive for the best possible decoding performance. For doing so, he/she may have the choice between different decoding models or may try to improve upon existing methods. In a real application, the relative variance of the labels, dataset size, and label noise level are typically not or only weakly controllable, even though these dataset-inherent parameters may have a strong impact upon the decoding. In

our benchmarking scenario, however, absolute control over these or similar dataset-inherent parameters is granted for free, thus allowing to investigate, under which conditions a decoding model is applicable or which aspects of an existing method should be improved in order to optimize the decoding performance. Along the lines of the illustrative chart shown in **Figure 3**, we provide three exemplary use cases, where we tested the robustness of the three decoding methods wrt. the aforementioned parameters, namely: (1) relative variance of the labels, (2) dataset size, and (3) label noise level.

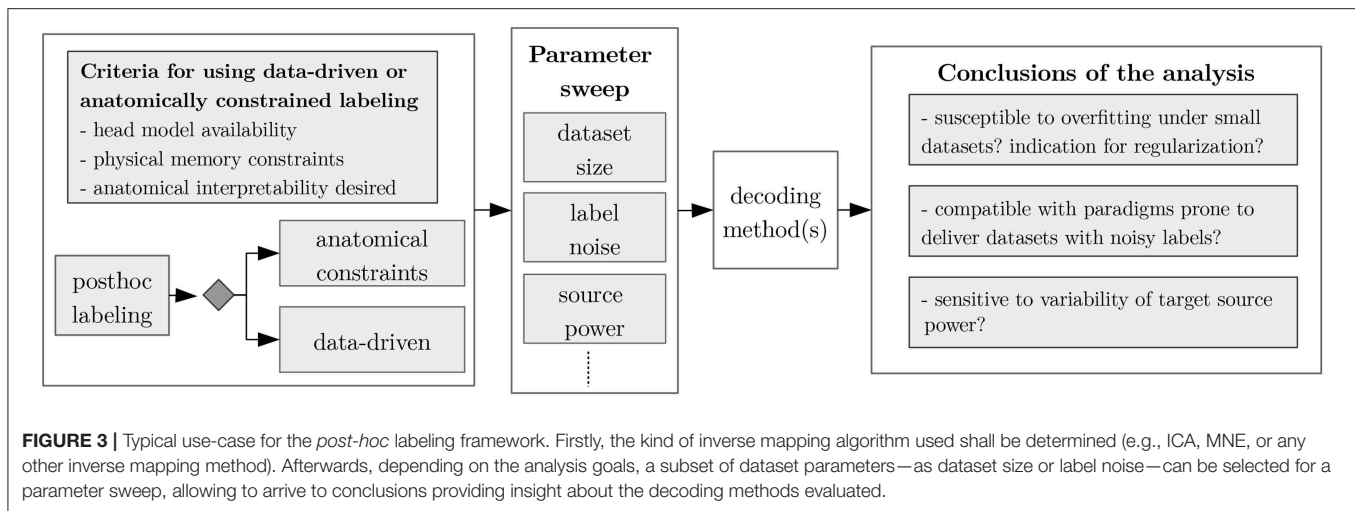### 3.3.1. Relative Variance of the Labels
First, machine learning methods favor datasets that contain high contrast in the labels. For example, Meinel et al. (2018) demonstrated that SPoC decoding performance is positively correlated with the variance of the target labels z over epochs e. In our current contribution, we analyzed the performance of the decoding methods for three type of target sources: sources with high, medium, and low power variability, each corresponding to the respective subject-wise *tertile membership* of the source power variability (z-variance). Note that when using ICA for generating $\mathcal{S}_d$, the scaling of the sources S is unknown. This does not represent any drawback for our framework; however, to allow for the analysis of performance with respect to z-variance, each of the sources in $S \in \mathcal{S}_d$ were normalized wrt. their ℓ_2-norm and scaled using the average covariance of each source with all the channels, i.e., $\tilde{s}_i = \left(1/N_c \sum_{N_c} s_i x_i^\top \right) s_i$; with $\tilde{s}_i$ being the i-th row in $\tilde{S} \in \mathcal{S}_d$, from which the z labels are finally extracted, according to section 2.2.2.

3.3.2. Dataset Size
Second, each decoding model's sensitivity wrt. the *number of training epochs* was evaluated by sweeping from 50 to 2,000 epochs, as larger datasets prevent overfitting and deliver more robust models.

3.3.3. Label Noise
Third, the influence of *label noise* was investigated, which was either imposed upon z (in the case of continuous labels in regression problems) or upon y (for classification problems). For both types of problems, the intensity of the noise was defined, respectively, as $\{\xi_n^{reg}, \xi_n^{class}\} \in \mathbb{R}$, $0 < \xi_n^{reg} < 1$, $0 < \xi_n^{class} < 1$, specifically:

FIGURE 3 | Typical use-case for the *post-hoc* labeling framework. Firstly, the kind of inverse mapping algorithm used shall be determined (e.g., ICA, MNE, or any other inverse mapping method). Afterwards, depending on the analysis goals, a subset of dataset parameters—as dataset size or label noise—can be selected for a parameter sweep, allowing to arrive to conclusions providing insight about the decoding methods evaluated.

Label noise for regression: The variable ξ_n^{reg} controls the correlation ρ_n between the original clean labels z and more challenging labels z_n, such that $\xi_n^{reg} = (1 - \rho_n)$ holds. Subsequently, noisy labels z_n are defined as

$$z_n = z + \frac{1 - (1 - \xi_n^{reg})^2}{(1 - \xi_n^{reg})^2} \text{var}(z) \, \eta \qquad (9)$$

where η is a normally distributed random variable.

Label noise for classification: The variable ξ_n^{class} for discrete labels is defined as

$$\xi_n^{class} = 2 \left(1 - P\left[y_n(e) = y(e)|y(e)\right]\right) \qquad (10)$$

that is, ξ_n^{class} controls the probability of assigning a given epoch e to a class different from the ground truth. For multiclass problems, if $y_n(e) \neq y(e)$, then $y_n(e)$ is assigned with equal probability to any of the remaining classes.

Sweeping over the three hyperparameters delivered more than 10,000 evaluations of different dataset configurations. The performance of each configuration is assessed by a chronological 5-fold cross-validation procedure. All results shown are based on the data of all the seven subjects. The results shown for SPoC and CSP were computed using the dataset generation framework based on data-driven source constraints, whereas for ConvNet, the version with physiologically constrained sources was used.

4. RESULTS OF USE-CASES

The datasets generated using the *post-hoc* labeling framework are characterized by **Figure 4**. Per subject, an average of 4,700 epochs could be obtained from only 73 min of average EEG recording time, while an average of 2,136 epochs were rejected as artifactual. Thus the proposed framework yields an acceptance rate of approximately 55 %. The accepted epochs could then be labeled in multiple ways, as each source can be utilized to define a label set. Per source, the resulting labels were analyzed for variance. We found, that the distribution of all label variances approximates a gamma distribution (see histograms in **Figure 4**),

both for the labels extracted via MNE and those extracted by the data-driven fastICA approach.

Next, we exemplify how the sensitivity of the three use-case algorithms SPoC, CSP, and ConvNet toward dataset-inherent parameters can be analyzed using the proposed framework. The top row of **Figure 5** shows the performance metrics obtained by a sweep over the parameter *label noise* while maintaining the full dataset size. The bottom row depicts the influence of the *dataset size* upon the performance, while no label noise was applied ($\xi_n^{class} = \xi_n^{reg} = 0$). For each subplot, results have been grouped into tertiles defined by z-variance. In addition, **Figure 6** provides a full bi-parametric analysis separately for the z-variance tertiles.

We observed, that label noise values of $\xi_n^{reg} > 0.2$ significantly reduced SPoC's performance score ρ. Interestingly, the variance of the labels z does not seem to play a systematic role for SPoC's decoding quality, if the training set is large. However, SPoC's performance suffers a drop for dataset sizes of ≈ 300 epochs or less, and this drop is most pronounced for the low and medium groups of z-variance. It can also be observed that high label noise cannot be compensated for by SPoC, even with large training datasets. However, for limited amount of noise, a larger dataset may be a sufficient countermeasure, which can be obtained by adapting the experimental design of a given paradigm in order to improve the data collection efficiency. The method CSP shows a similar behavior to SPoC. The critical thresholds of the investigated parameters, for which the performance dramatically drops, are around a label noise value of $\xi_n \approx 0.1$ and a dataset containing ≈ 400 samples. These results might be an indication for applying regularization techniques, specially if CSP or SPoC are to deal with small datasets. This conclusion agrees with state-of-the-art studies, e.g., by Meinel et al. (2018).

ConvNet models show a high performance variance over the full spectrum of both parameters. Its performance scores are very sensitivity wrt. the number of epochs and label noise: this data-hungry method requires approx. 1,000 epochs under no label noise to reach peak performance. Even small amounts of label noise influence ConvNet's performance, with label noise larger than $\xi_n^{class} \approx 0.1$ leading to a pronounced decline of the network's performance. It is necessary to remark that the 3-class
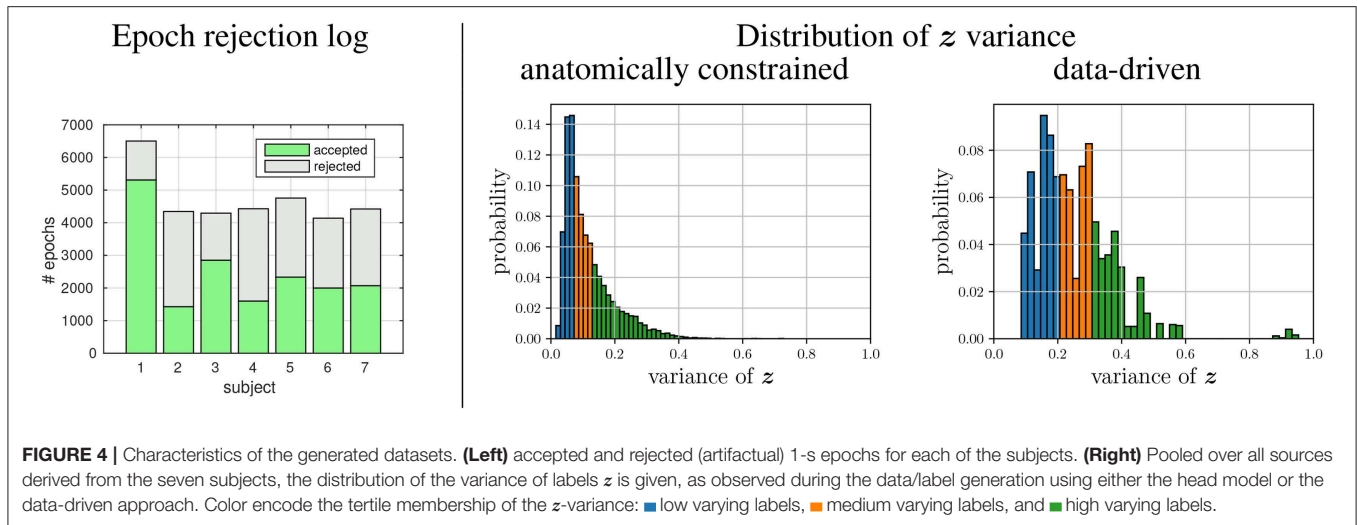
FIGURE 4 | Characteristics of the generated datasets. **(Left)** accepted and rejected (artifactual) 1-s epochs for each of the subjects. **(Right)** Pooled over all sources derived from the seven subjects, the distribution of the variance of labels z is given, as observed during the data/label generation using either the head model or the data-driven approach. Color encode the tertile membership of the z-variance: ■ low varying labels, ■ medium varying labels, and ■ high varying labels.
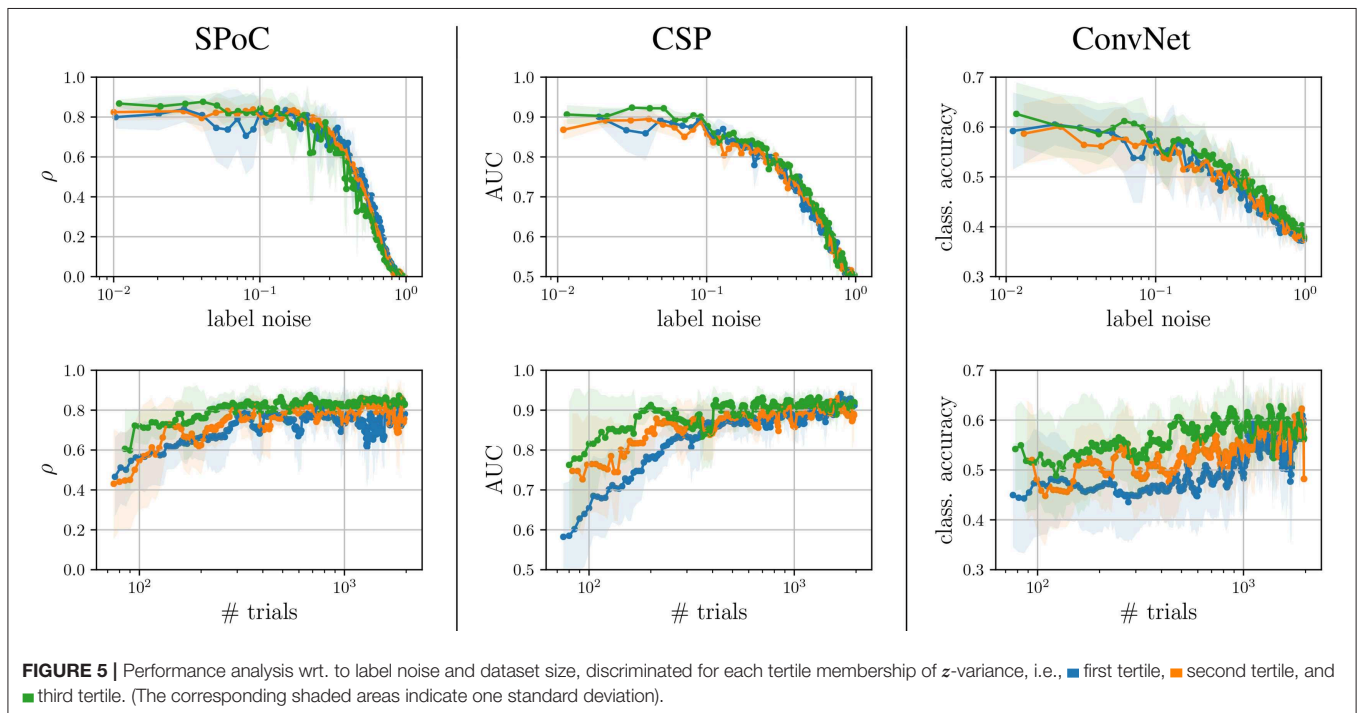

FIGURE 5 | Performance analysis wrt. to label noise and dataset size, discriminated for each tertile membership of z-variance, i.e., ■ first tertile, ■ second tertile, and ■ third tertile. (The corresponding shaded areas indicate one standard deviation).

classification task solved by ConvNet is the most complex one of the use-cases analyzed. This may partially explain the increased sensitivity of the method to the dataset size and label noise. However, another likely reason for such sensitivity is the large number of free parameters of the network to be tuned during its training, indicating that the complexity of the network should be reduced if the dataset size is not large enough.

5. DISCUSSION

For the development of decoding algorithms, an ideal testbench should be capable of providing large amounts of data, clean labels, and realistic neural dynamics. Unfortunately, state-of-the-art approaches lack one or several of these properties. To address this, we have introduced a labeled dataset generation framework. Its key idea is to implement a *post-hoc labeling* of (potentially very long) paradigm-agnostic pre-recorded M/EEG signals.

5.1. Advantages Over State-of-the-Art Frameworks

The *post-hoc* labeling framework allows to generate relatively large labeled datasets based on real neural signals and by doing so, it prescinds from making critical assumptions about neural dynamics. As a clear advantage of the proposed framework, the labels can be deterministically recoverable from
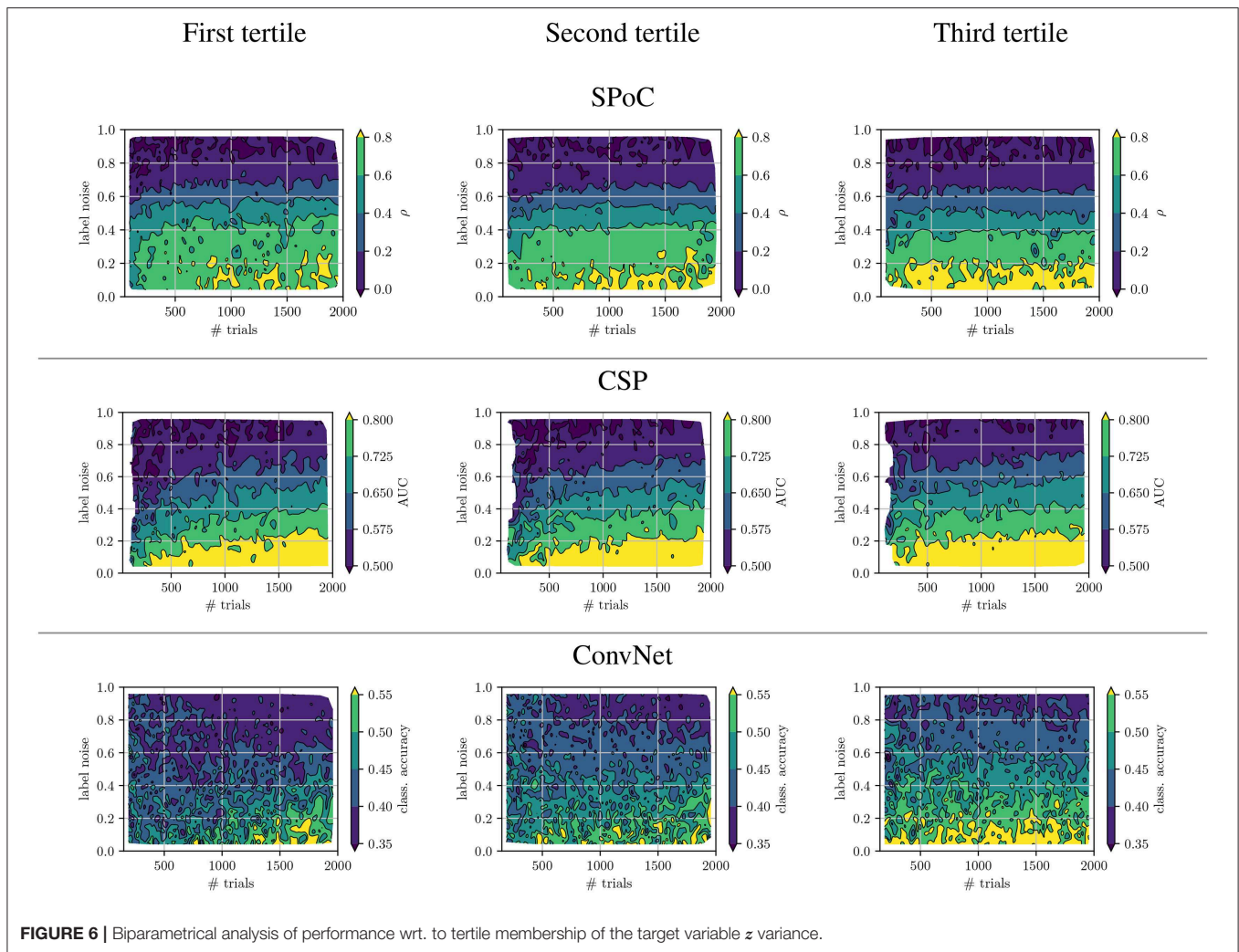
FIGURE 6 | Biparametrical analysis of performance wrt. to tertile membership of the target variable *z* variance.

the available data, thus, they are provided free of noise, from the perspective of the decoding methods. Furthermore, the *post-hoc* labeling of paradigm-agnostic M/EEG recordings offers greater efficiency in terms of data use. Compared to real datasets whose labels depend on the paradigm they were recorded under, our *post-hoc* labeling can also make use of recorded idle periods or preparatory intervals. For the provided EEG dataset, this led to an exploitation of effectively 55% of the overall M/EEG recording time for training and test data generation.

Our framework allows for absolute control over important parameters of the generated dataset and provides full knowledge about the statistical and—in case of the head model—anatomical properties of the target sources. It provides an ideal starting point for comparing competing decoding methods, as it yields insight into the data conditions under which methods stand out among the competing others.

Nevertheless, one relevant parameter unfortunately remains outside the control of our framework: the amount of sensor noise. This parameter is determined by the available real-M/EEG signals and can not be improved (only worsened) *post-hoc*.

Here, synthetic data generation approaches have a theoretical advantage, as they can control the level of sensor noise. In practice, however, it may not be straightforward to determine noise levels during synthetic data generation in order to match real experimental conditions.

5.2. Application in Development of Neural Decoding Methods

With the three use-cases presented as exemplary analyses enabled by *post-hoc* labeling, we intent to show how our framework can be used to investigate strengths and limitations of arbitrary decoding algorithms under different scenarios. For example, it could be easily observed, that the performance differences among labels with different variability is marginal, if the number of epochs surpasses a critical, method-specific threshold. However, for datasets below this threshold, the strongest varying labels showed a better performance than those belonging to the first and second tertile of label variability. Using the *post-hoc* labeling framework was an effective way to investigate these thresholds as well as the influence of label noise. However, not only the required (minimum)

TABLE 1 | Comparison of advantages and disadvantages of two state-of-the-art testbench scenarios against the proposed novel *post-hoc labeling* framework.

Dataset type	Dataset size	Label noise ctrl.	Sensor noise ctrl.	Real statistics and dynamics
Synthetic	Large ● (green)	Yes ● (green)	Yes ● (green)	No ● (red)
Real EEG	Small ● (red)	No ● (red)	No ● (red)	Yes ● (green)
Post-hoc labeling	Medium/large ◐ (yellow/green)	Yes ● (green)	No ● (red)	Yes ● (green)

amount of training data or feasible levels of label noise can be examined with the framework, also other specific constraints could easily be incorporated during the generation of the benchmark datasets. Examples are prior knowledge about central frequency, strength, or anatomical location of sources.

Nevertheless, frameworks using purely synthetically generated signals or real EEG recordings with paradigm specific labels have their place, next to the *post-hoc* labeling framework, in the development pipeline of neural decoding methods. Early development stages of decoding algorithms may benefit from strictly controlled simulation environments, as shown for example in the recent work presented by Krol et al. (2018). Furthermore, in late development stages and prior to deployment in real-world applications, validation on strictly real scenarios is necessary, for example, using the benchmark framework provided by Jayaram and Barachant (2018).

5.3. Considerations About Non-stationary Dynamics

By using MNE and ICA for inverse mapping, as shown in the exemplary use-cases, we assume a stationary mapping between M/EEG signals and neural sources, which may not always be true, specially for long recordings. Such assumption should be kept in the foreground, mainly in two scenarios: First, when claiming that the generated target sources exclusively correspond to a particular source with specific anatomical or physiological interpretation, since a stationary inverse mapping under non-stationary conditions would potentially deliver a time-varying mixture of underlying neural sources. Second, when benchmarking neural decoding methods designed to deal with non-stationary dynamics, since *post-hoc* labeling using stationarity assumptions may deliver a dataset where the strengths of any adaptive decoding method cannot be properly evaluated.

Using ICA and MNE in our use-cases was motivated by the observation, that the most popular decoding methods (for example, CSP, SPoC, or convNets) predominantly assume an underlying within-session stationary process. Challenging those popular decoding methods with labels derived by, e.g., STOUT or adaptive ICA would probably have an impact upon these decoding methods comparable to label noise.

Fortunately, the flexibility provided by the *post-hoc* labeling framework facilitate the use of inverse mapping methods, capable of dealing with underlying non-stationary processes. For example, state-of-the-art source reconstruction methods as TF-MxNE (Gramfort et al., 2013b) or STOUT (Castaño-Candamil et al., 2015a) are designed to extract neural sources from non-stationary M/EEG recordings, and could be employed in our *post-hoc* labeling framework, instead of MNE. Likewise, adaptive blind-source separation methods (usually deployed in online scenarios) can be used to perform a data-driven inverse mapping without assuming stationary dynamics (Hsu et al., 2015), as an alternative to the ICA procedure presented. Finally, under the assumption of piecewise stationarity, long recordings may be segmented into locally stationary windows—e.g., using statistical features of the spectrogram of the signals (Hory et al., 2002)—, and then the proposed *post-hoc* labeling framework can be carried out in the resulting (stationary) epochs.

To wrap up, **Table 1** summarizes the properties of our contribution (*post-hoc labeled data*) compared to that of other testbench approaches.

To facilitate the adoption of the *post-hoc* labeling framework as a tool for developing and testing decoding algorithms for oscillatory neural phenomena, both the source code and datasets utilized in the use-case scenarios have been made publicly available[1] (Castaño-Candamil et al., 2017a).

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and/or the supplementary files.

AUTHOR CONTRIBUTIONS

SC-C, AM, and MT: conceived the methods, collected dataset, and contributed analysis tools. SC-C and AM: method implementation. SC-C: performed the experiments. SC-C and MT: wrote the paper.

ACKNOWLEDGMENTS

---

[1]https://github.com/bsdlab/Post-HocLabeling

REFERENCES

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on* (Hong Kong), 2390–2397.

Baillet, S., Mosher, J. C., and Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Sig. Process. Mag.* 18, 14–30. doi: 10.1109/79.962275

Blankertz, B., Acqualagna, L., Dähne, S., Haufe, S., Schultze-Kraft, M., Sturm, I., et al. (2016). The Berlin brain-computer interface: progress beyond communication and control. *Front. Neurosci.* 10:530. doi: 10.3389/fnins.2016.00530

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of ERP components–a tutorial. *NeuroImage* 56, 814–825. doi: 10.1016/j.neuroimage.2010.06.048

Breakspear, M., Heitmann, S., and Daffertshofer, A. (2010). Generative models of cortical oscillations: neurobiological implications of the Kuramoto model. *Front. Hum. Neurosci.* 4:190. doi: 10.3389/fnhum.2010.00190

Castaño-Candamil, S., Höhne, J., Martínez-Vargas, J.-D., An, X.-W., Castellanos-Domínguez, G., and Haufe, S. (2015a). Solving the EEG inverse problem based on space–time–frequency structured sparsity constraints. *NeuroImage* 118, 598–612. doi: 10.1016/j.neuroimage.2015.05.052

Castaño-Candamil, S., Meinel, A., Dähne, S., and Tangermann, M. (2015b). Probing meaningfulness of oscillatory EEG components with bootstrapping, label noise and reduced training sets. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2015, 5159–5162. doi: 10.1109/EMBC.2015.7319553.

Castaño-Candamil, S., Meinel, A., and Tangermann, M. (2017a). *Post-hoc* labeling of arbitrary EEG recordings for data-efficient evaluation of neural decoding methods. Available online at: https://zenodo.org/record/1065107

Castaño-Candamil, S., Meinel, A., and Tangermann, M. (2017b). *Post-hoc* labeling of arbitrary EEG recordings for data-efficient evaluation of neural decoding methods. *arXiv[Preprint].arXiv:1711.08208.*

Dähne, S., Bießmann, F., Meinecke, F. C., Mehnert, J., Fazli, S., and Müller, K.-R. (2014a). "Multimodal integration of electrophysiological and hemodynamic signals," in *Brain-Computer Interface (BCI), 2014 International Winter Workshop on* (Jeongsun-kun: IEEE), 1–4.

Dähne, S., Meinecke, F. C., Haufe, S., Höhne, J., Tangermann, M., Müller, K.-R., et al. (2014b). SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. *NeuroImage* 86, 111–122. doi: 10.1016/j.neuroimage.2013.07.079

Dähne, S., Nikulin, V. V., Ramírez, D., Schreier, P. J., Müller, K.-R., and Haufe, S. (2014c). Finding brain oscillations with power dependencies in neuroimaging data. *NeuroImage* 96, 334–348. doi: 10.1016/j.neuroimage.2014.03.075

de Cheveigné, A., and Parra, L. C. (2014). Joint decorrelation, a versatile tool for multichannel data analysis. *NeuroImage* 98, 487–505. doi: 10.1016/j.neuroimage.2014.05.068

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. doi: 10.1016/j.jneumeth.2003.10.009

Forney, E. M., Anderson, C. W., Gavin, W. J., Davies, P. L., Roll, M. C., and Taylor, B. K. (2015). *Technical Report: Echo State Networks for Modeling and Classification of EEG Signals in Mental-Task Brain Computer Interfaces.* Technical Report MSU-CSE-06-2, Department of Computer Science, Colorado State University, Fort Collins.

Fukunaga, K. (2013). *Introduction to Statistical Pattern Recognition.* Academic Press.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013a). MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 7:267. doi: 10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage* 86, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M. S., and Kowalski, M. (2013b). Time-frequency mixed-norm estimates: sparse M/EEG imaging with non-stationary source activations. *NeuroImage* 70, 410–422. doi: 10.1016/j.neuroimage.2012.12.051

Grech, R., Cassar, T., Muscat, J., Camilleri, K. P., Fabri, S. G., Zervakis, M., et al. (2008). Review on solving the inverse problem in EEG source analysis. *J. Neuroeng. Rehabil.* 5:25. doi: 10.1186/1743-0003-5-25

Hallez, H., Vanrumste, B., Grech, R., Muscat, J., De Clercq, W., Vergult, A., et al. (2007). Review on solving the forward problem in EEG source analysis. *J. Neuroeng. Rehabil.* 4:46. doi: 10.1186/1743-0003-4-46

Haufe, S., Dähne, S., and Nikulin, V. V. (2014a). Dimensionality reduction for the analysis of brain oscillations. *NeuroImage* 101, 583–597. doi: 10.1016/j.neuroimage.2014.06.073

Haufe, S., and Ewald, A. (2016). A simulation framework for benchmarking EEG-based brain connectivity estimation methodologies. *Brain Topogr.* 32, 1–18. doi: 10.1007/s10548-016-0498-y

Haufe, S., Meinecke, F. C., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., et al. (2014b). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* 87, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

Herrmann, C., and Demiralp, T. (2005). Human EEG gamma oscillations in neuropsychiatric disorders. *Clin. Neurophysiol.* 116, 2719–2733. doi: 10.1016/j.clinph.2005.07.007

Higashi, H., and Tanaka, T. (2013). Simultaneous design of FIR filter banks and spatial patterns for EEG signal classification. *IEEE Trans. Biomed. Eng.* 60, 1100–1110. doi: 10.1109/TBME.2012.2215960

Höhne, J., Holz, E., Staiger-Sälzer, P., Müller, K.-R., Kübler, A., and Tangermann, M. (2014). Motor imagery for severely motor-impaired patients: evidence for brain-computer interfacing as superior control solution. *PLoS ONE* 9:e104854. doi: 10.1371/journal.pone.0104854

Horschig, J. M., Zumer, J. M., and Bahramisharif, A. (2014). Hypothesis-driven methods to augment human cognition by optimizing cortical oscillations. *Front. Syst. Neurosci.* 8:119. doi: 10.3389/fnsys.2014.00119

Hory, C., Martin, N., and Chehikian, A. (2002). Spectrogram segmentation by means of statistical features for non-stationary signal interpretation. *IEEE Trans. Sig. Process.* 50, 2915–2925. doi: 10.1109/TSP.2002.805489

Hsu, S.-H., Mullen, T. R., Jung, T.-P., and Cauwenberghs, G. (2015). Real-time adaptive EEG source separation using online recursive independent component analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 309–319. doi: 10.1109/TNSRE.2015.2508759

Huang, Y., Parra, L. C., and Haufe, S. (2016). The New York Head— A precise standardized volume conductor model for EEG source localization and tES targeting. *NeuroImage* 140, 150–162. doi: 10.1016/j.neuroimage.2015.12.019

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw.* 13, 411–430. doi: 10.1016/S0893-6080(00)00026-5

Jayaram, V., Alamgir, M., Altun, Y., Schölkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Comput. Intell. Mag.* 11, 20–31. doi: 10.1109/MCI.2015.2501545

Jayaram, V., and Barachant, A. (2018). MOABB: trustworthy algorithm benchmarking for BCIs. *arXiv-CoRR*, abs/1805.06427.

Kim, J., Shin, H., and Robinson, P. (2007). "Compact continuum brain model for human electroencephalogram," in *Microelectronics, MEMS, and Nanotechnology* (International Society for Optics and Photonics) (Canberra, ACT), 68020T–68020T.

Klimesch, W., Sauseng, P., and Hanslmayr, S. (2007). EEG alpha oscillations: the inhibition–timing hypothesis. *Brain Res. Rev.* 53, 63–88. doi: 10.1016/j.brainresrev.2006.06.003

Koles, Z. J., Lazar, M. S., and Zhou, S. Z. (1990). Spatial patterns underlying population differences in the background EEG. *Brain Topogr.* 2, 275–284. doi: 10.1007/BF01129656

Krauledat, M., Tangermann, M., Blankertz, B., and Müller, K.-R. (2008). Towards zero training for brain-computer interfacing. *PLoS ONE* 3:e2967. doi: 10.1371/journal.pone.0002967

Krol, L. R., Pawlitzki, J., Lotte, F., Gramann, K., and Zander, T. O. (2018). SEREEGA: simulating event-related EEG activity. *J. Neurosci. Methods* 309, 13–24. doi: 10.1016/j.jneumeth.2018.08.001

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2016). EEGNet: a compact convolutional network for EEG-based brain-computer interfaces. *arXiv[Preprint].arXiv:1611.08024.*

Lemm, S., Blankertz, B., Curio, G., and Müller, K.-R. (2005). Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Biomed. Eng.* 52, 1541–1548. doi: 10.1109/TBME.2005.851521

Lindgren, J. T., Merlini, A., Lécuyer, A., and Andriulli, F. P. (2018). simBCI-A framework for studying BCI methods by simulated EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 2096–2105. doi: 10.1109/TNSRE.2018.2873061

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *J. Neural Eng.* 15:031005. doi: 10.1088/1741-2552/aab2f2

Makeig, S., Bell, A. J., Jung, T.-P., and Sejnowski, T. J. (1996). "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems* (Denver, CO), 145–151.

Marshall, T. R., den Boer, S., Cools, R., Jensen, O., Fallon, S. J., and Zumer, J. M. (2018). Occipital alpha and gamma oscillations support complementary mechanisms for processing stimulus value associations. *J. Cogn. Neurosci.* 30, 119–129. doi: 10.1162/jocn_a_01185

Martel, A., Dähne, S., and Blankertz, B. (2014). EEG predictors of covert vigilant attention. *J. Neural Eng.* 11:035009. doi: 10.1088/1741-2560/11/3/035009

Meinel, A., Castaño-Candamil, S., Blankertz, B., Lotte, F., and Tangermann, M. (2018). Characterizing regularization techniques for spatial filter optimization in oscillatory eeg regression problems. *Neuroinformatics* 17, 1–17. doi: 10.1007/s12021-018-9396-7

Meinel, A., Castaño-Candamil, S., Reis, J., and Tangermann, M. (2016). Pre-trial EEG-based single-trial motor performance prediction to enhance neuroergonomics for a hand force task. *Front. Hum. Neurosci.* 10:170. doi: 10.3389/fnhum.2016.00170

Ng, A. Y. (2004). "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *Proceedings of the Twenty-First International Conference on Machine Learning* (Banff, AB: ACM), 78.

Novi, Q., Guan, C., Dat, T. H., and Xue, P. (2007). "Sub-band common spatial pattern (SBCSP) for brain-computer interface," in *Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference on* (Kohala Coast, HI: IEEE), 204–207.

Pascual-Marqui, R. D. (1999). Review of methods for solving the EEG inverse problem. *Int. J. Bioelectromagnet.* 1, 75–86.

Reis, J., Schambra, H. M., Cohen, L. G., Buch, E. R., Fritsch, B., Zarahn, E., et al. (2009). Noninvasive cortical stimulation enhances motor skill acquisition over multiple days through an effect on consolidation. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1590–1595. doi: 10.1073/pnas.0805413106

Robinson, P., Rennie, C., and Rowe, D. (2002). Dynamics of large-scale brain activity in normal arousal states and epileptic seizures. *Phys. Rev. E* 65:041924. doi: 10.1103/PhysRevE.65.041924

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., et al. (2012). Review of the BCI competition IV. *Front. Neurosci.* 6:55. doi: 10.3389/fnins.2012.00055

Tangermann, M., Reis, J., and Meinel, A. (2015). "Commonalities of motor performance metrics are revealed by predictive oscillatory EEG components," in *Proceedings of the 3rd International Congress on Neurotechnology, Electronics and Informatics (NEUROTECHNIX)* (Lisbon), 32–38.

Tomioka, R., Dornhege, G., Nolte, G., Blankertz, B., Aihara, K., and Müller, K.-R. (2006). *Spectrally Weighted Common Spatial Pattern Algorithm for Single Trial EEG Classification.* Mathematical Engineering Technical Report, METR–2006–40.

Vigário, R., Sarela, J., Jousmiki, V., Hamalainen, M., and Oja, E. (2000). Independent component approach to the analysis of EEG and MEG recordings. *IEEE Trans. Biomed. Eng.* 47, 589–593. doi: 10.1109/10.841330

Winkler, I., Haufe, S., and Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behav. Brain Funct.* 7:30. doi: 10.1186/1744-9081-7-30