Computer Methods and Programs in Biomedicine 271 (2025) 109008

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Unveiling multi-domain signatures of EEG oscillations using a fully-interpretable convolutional neural network

Davide Borra *, Elisa Magosso

Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI), University of Bologna, Cesena, Forlì-Cesena, Italy

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.16156955, https://www.bbci.de/competition/iv/, https://gigadb.org/dataset/100542

Keywords:
Neural oscillations
EEG analysis
Interpretable convolutional neural networks
Motor imagery

ABSTRACT

Background and Objective: Neural oscillations are widely investigated to characterize brain functions. However, their analysis in the frequency, spatial, and temporal domains from electroencephalographic (EEG) signals (e.g., via event-related spectral perturbations) is affected by choices that limit the quality, reproducibility, and reliability of results. For example, different pre-processing and processing steps strongly affect the results, and the steps are often implemented with manual/semi-automatic algorithms. Moreover, due to the high dimensionality of the involved measures, generally a few frequency intervals/brain regions/time intervals of interest are selected exploiting a priori knowledge, and then analyzed. Therefore, it is desired an end-to-end approach that automatically learns the optimal strategy to process minimally pre-processed EEG to highlight the most relevant signatures of brain oscillations in the frequency, spatial, and temporal domains with minimal a priori assumptions.

Methods: In this study, we design a novel framework for characterizing EEG oscillations in the frequency, spatial, and temporal domains, based on the features learned by a fully-interpretable convolutional neural network. The network learns a bank of bandpass filters to be applied to minimally pre-processed EEG. Then, frequency-specific spatial and temporal filtering allow the learning of the most salient spatial and time samples, separately for each frequency component. Finally, the framework processes the learned interpretable features to reveal meaningful EEG signatures.

Results: We test our approach by applying it to real data recorded during motor imagery tasks. Our neural network-empowered approach reveals the modulations of brain oscillations known to occur during motor imagery, and match results obtained with classic analyses. Specifically, the alpha band (8–13 Hz) was the most important, together with the electrodes covering motor areas and the time samples closer to the cue indicating the action to imagine.

Conclusions: The proposed framework enables the characterization of brain oscillations in an automatic, optimal and end-to-end way, and could be conveniently exploited for boosting our comprehension of brain functions in healthy participants and in patients, tracking their neuropathological alterations.

1. Introduction

Neural oscillations, reflecting the synchronized rhythmic fluctuations of local neuronal ensembles, have long been investigated to help characterize cognitive functions in both healthy individuals and neurological disorders [1–4]. In particular, electroencephalographic (EEG) signals are widely analyzed to quantify neural oscillations in a non-invasive and cost-effective way, by exploiting time–frequency representations (e.g., power representations based on Wavelet analysis) separately for each spatial location (i.e., EEG channel), thus providing their characterization in the frequency, spatial, and temporal domains. Due to the complex, noisy and high-dimensional nature of EEG signals, the quantification of EEG oscillations is affected by many factors. First, the EEG needs to be heavily pre-processed to attenuate noise components before utilizing it, for example by filtering data in a specific frequency range of interest (i.e., retaining only a specific EEG band), removing noisy EEG channels, removing artifacts (e.g., eye blinks) via independent component analysis [5,6]. The specific pre-processing pipeline strongly affects the quality of the results [7]. Indeed, task-relevant neural activity may be masked out, for example by selecting a priori a specific EEG band for the processing, or by removing independent components containing mixed activity (related to both brain and noise components). Moreover, EEG studies often rely on manual/semi-automatic and non-standardized pre-processing pipelines [8]; these aspects limit the reproducibility and reliability of the results. Besides

* Corresponding author.
E-mail address: davide.borra2@unibo.it (D. Borra).

https://doi.org/10.1016/j.cmpb.2025.109008
Received 9 May 2025; Received in revised form 19 July 2025; Accepted 31 July 2025
Available online 8 August 2025
0169-2607/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

the pre-processing steps, also the processing methods adopted for extracting and characterizing the EEG oscillations affect the results. For example, time–frequency power computation can be performed using either fast Fourier or Wavelet transforms. In turn, Wavelet transform can also have different variants, depending on the mother wavelet used (e.g., Morlet, Daubechies, Meyer) [9,10]. In addition to variants related to different spectral methods, other variants may affect the computation of power of EEG oscillations related to a specific task. Indeed, as task-related power changes are often subtle and difficult to observe, they are computed relative to a baseline (baseline correction) [10]. However, there is no consensus on how to perform baseline correction, and different approaches have been proposed and are used (e.g., subtractive or divisive approach, at the level of single-trial or after trial averaging [10]). Of course, also the choice of the baseline (e.g., its length) influences the results. In addition to these aspects limiting the analysis of neural oscillations, an important challenge regards the large number of measures involved, that is, multiple spatial locations, time samples, and frequencies characterizing the time–frequency representation. Typically, one way to solve this is to exploit the existing literature to define a priori different regions (i.e., scalp sites)/time intervals/frequency intervals of interest, and restrict the analysis only on them [9]. Of course, all the previous aspects require strong a priori choices that impact on the final results. This poses the challenge of fairly comparing the results across different studies, and, critically, might result in discarding potentially relevant information (e.g., neural modifications unknown so far). Furthermore, adopting a priori choices may not be feasible when analyzing patients' signals, as their neural activity may be characterized by unpredictable features and unknown modifications, also depending on the severity of neuropathology.

Even though part of these aspects has been addressed in the past years, for example by proposing automatic and standardized parts of the pre-processing pipeline [11,12] or by defining guidelines [13,14], the previous limitations still affect the analysis of neural oscillations, which results in a time-consuming trial-and-error procedure, often guided by the expectation of results about the investigated cognitive functions. Thus, it is desired an end-to-end approach that automatically learns the optimal strategy to process raw/minimally pre-processed EEG to highlight, for a target cognitive task, the most relevant neural signatures in the frequency, spatial, and temporal domains with no/minimal a priori selections.

Convolutional neural networks (CNNs) [15] are widely used for processing single-trial EEG, by mapping it to the corresponding brain state (e.g., right-hand motor imagery) characterizing the cognitive task (e.g., motor imagery) [16,17]. These algorithms have been mainly applied to the EEG with the aim of decoding, for example to realize accurate decoders to be used, in prospective, in brain–computer interfaces. At variance with traditional machine learning, in which a separation between feature extraction, selection, and classification occurs, CNNs address neural decoding in an end-to-end fashion, linking raw/lightly pre-processed EEG time series to the relevant variables to be decoded (e.g., brain states of interest). CNNs outperformed traditional machine learning, widely across diverse decoding problems [18], e.g., decoding of motor imagery [19,20], P300 [20–22], and steady-state visual evoked potential [23–25]. From their first application to P300 decoding with a compact architecture in Cecotti and Graser [26], CNNs improved over the past years. For example, Schirrmeister et al. [19] realized an architecture that mimics the processing operated by a traditional machine learning approach (filter bank common spatial pattern), facilitating the learning of spectrum power-related EEG features. This specialized architecture resulted beneficial for decoding problems exploiting sensorimotor oscillations (e.g., motor imagery). Then, general-purpose architectures have been proposed and applied. These typically involve convolutions operating individually in specific domains, generally processing time series separately in the temporal domain (filtering the time series), and in the spatial domain (learning the optimal combinations of channels). Optionally, they might further process signals in time with deeper temporal convolutions. Among the proposed architectures, EEGNet [20] and its variants [21,22,24,27–29] represent the most used ones [18], providing a good trade-off between model size (i.e., number of trainable parameters), training time, and decoding performance, also reaching state-of-the-art performance in various international EEG decoding competitions [21,27]. Finally, improvements regarded the adoption of multi-resolution temporal feature learning [22,28,30–32], by means of parallel convolutions operating at different time scales. In addition to these advancements, attention-based transformer models were introduced in CNNs to design convolutional transformers [33]. These networks have the advantage of focusing not only on local features (learned by convolutional filters within local receptive fields), but also on long-term relationships.

Importantly, CNNs could also have a key role in EEG analysis and not only for EEG decoding. Indeed, even from raw/minimally pre-processed EEG, these networks are able to automatically learn the most important neural features to link the input EEG to output brain states, separately in different domains (e.g., temporal domain). By leveraging this automatic feature learning, it might be possible to analyze the EEG in a data-driven way. However, these models are scarcely interpretable out-of-the-box. Solutions were proposed in the literature for increasing the interpretability of the features learned by the network, by coupling models with explanation techniques (post-hoc model explanation) [34–36], and by incorporating interpretability in a given domain directly into the network structure (ad-hoc interpretable modeling), realizing interpretable CNNs [37–39]. Post-hoc model explanations, by relying on external algorithms to increase interpretability, pose the challenge of not only defining the network structure but also of defining the explanation technique and its settings for properly analyzing EEG [34–36]. Moreover, post-hoc model explanations are generally applied to non-interpretable networks (i.e., networks with layers not designed to learn directly interpretable features). Hence, they typically reveal the most salient samples of the network input, since the network input actually represents the sole interpretable element of the network; thus, they analyze the most relevant characteristics only in the domains represented in the input (e.g., spatial and temporal domains, in case the network processes multi-channel EEG time series as input). For these reasons, adopting a inherently interpretable model would help more in reducing the settings influencing the EEG analysis, and by properly defining the interpretable components of the network, would help analyzing EEG oscillations in all domains characterizing it (i.e., frequency, spatial, and temporal domains). Regarding this point, Borra et al. [37,40] designed a network component (temporal convolutional layer) that forces the learning of a bank of ideal bandpass filters (i.e., based on the composition of sinc functions), by training the cutoff frequencies of filters (parameters directly interpretable) instead of all the filter coefficients. Zhao et al. [38] proposed a similar approach, in which a bank of bandpass filters based on real Morlet wavelets was used, learning the central frequency and bandwidth of the wavelets. More recently, Ludwig et al. [39] generalized these two approaches, by considering a generalized Gaussian function for modeling the frequency response of the bank of bandpass filters; the frequency response can be tuned to describe a response closer to a sinc-based bandpass filter or to a Morlet-based one, in addition to be tuned in the central frequency and bandwidth. Remarkably, all these approaches provide an increased interpretability of the frequency-domain features that are used by the network to predict the output brain state when processing the input EEG. Additionally, Borra et al. [37] also promoted the increase of interpretability in the spatial-domain features, by designing network components such that frequency-specific spatial features are learned, enabling the understanding of the EEG electrode sites most contributing to the discrimination, frequency by frequency.

Despite these remarkable steps, the state-of-the-art interpretable networks [37–40] enable the interpretation only of a single domain relevant for EEG processing – the frequency domain – with exception of a network architecture (Sinc-EEGNet [37]) that enables also the

interpretability of the frequency-related spatial features, thus characterizing the frequency and spatial domains. Importantly, increasing the interpretability also in the temporal domain would enable the identification of the most salient time samples of EEG signals during cognitive tasks. This would provide valuable insights about the dynamics of the neural processing from neural signals time-locked with respect to an event of interest (e.g., a stimulus onset). Moreover, the interpretability of the network features only covers a small portion (less than 50%) of the total features learned by the network; thus, the state-of-the-art interpretable networks still remain predominantly non-interpretable, overall. Finally, the previous studies [37–40] attempted to interpret the EEG features from the interpretable networks only for validation purposes and not for analysis purposes, aiming at validating the network feature learning, i.e., for checking that the network decision was based on neurophysiologically meaningful correlates of the predicted variables (e.g., motor imagery). On the contrary, the prior studies did not address the definition of an EEG analysis tool based on the interpretable feature learning operated by the CNN, which could be useful to analyze neural oscillations in an automatic, data-driven, and end-to-end way.

In this study, our goal is to move a step forward, towards the definition and use of CNN-based pipelines not merely aimed at *decoding* EEG signals but mainly aimed at *analyzing* EEG signals. Importantly, such approaches can be relevant to overcome the limitations of traditional EEG analysis, that often rely on highly pre-processed signals and are often biased by a priori methodological choices, driven by the expectation of results about the investigated cognitive functions. To this aim, here we propose the design of an innovative framework for automatically analyzing the EEG in a multiple domains – i.e., frequency, spatial, and temporal domains – by leveraging on the knowledge automatically learned by a fully-interpretable CNN, from raw/minimally pre-processed signals. Therefore, the primary contribution of our study regards bridging CNNs – traditionally exploited for EEG decoding – to EEG analysis, promoting an *automatic*, *optimal* (i.e., a data-driven procedure that is optimally tailored to the specific cognitive task under investigation), and *end-to-end* analysis of EEG signals (i.e., highlighting neural correlates of brain states directly from minimally pre-processed neural time series). Specifically, we design a compact and fully-interpretable CNN that learns to link the input EEG to the output brain states. Thanks to its architecture, the network enables a straightforward interpretation of the learned features in the frequency, spatial and temporal domains, thus revealing the most salient EEG features characterizing the decoded cognitive task, in multiple domains. In particular, the network first processes minimally pre-processed single-trial EEG by learning the optimal bank of bandpass filters to filter it. Then, frequency-specific spatial filtering and frequency-specific temporal filtering are performed, by learning the most salient EEG channels and time samples, separately for each bandpass-filtered EEG signal. Finally, based on these high-level features, the network predicts the output brain states. To test the proposed framework as an EEG analysis tool, we conducted experiments on real data collected during a motor imagery task. We used this representative task since the EEG signatures of motor imagery are well-established and documented. We showcase how the proposed network, realized fully-interpretable in its architecture, can be exploited to analyze the EEG in multiple domains, revealing the neural signatures of a cognitive task under investigation (motor imagery). Finally, as a secondary contribution, we provide a comparative analysis of interpretable CNNs for neural decoding, which is lacking in the literature, by comparing the main interpretable CNNs both in terms of decoding performance and interpretability.

2. Methods

In the following, the proposed EEG analysis framework is presented. First, we describe our fully-interpretable neural network for automatically learning the most salient frequency-, spatial-, and temporal-domain EEG features of the investigated cognitive task (e.g., motor imagery). Then, we illustrate how to process these EEG features to provide useful insights about the multi-domain signatures of EEG oscillations. Finally, the experiments conducted while applying our framework to real data are described.

2.1. Proposed EEG analysis framework

2.1.1. Learning multi-domain EEG features using a fully-interpretable neural network

Let us denote with $X \in \mathbb{R}^{C\times T}$ a minimally pre-processed single-trial EEG recorded from a specific participant (C: number of EEG channels, T: number of time samples) and with l the corresponding brain state, i.e., $l \in L = \left\{l_0, \ldots, l_{N_c-1}\right\}$ in case of N_c possible brain states (i.e., classes). A neural network can be used to find the optimal relationship that maps the input neural activity to the output brain state, solving a neural decoding problem. Here we use a fully-interpretable and fully-convolutional neural network, to enable the interpretation of multi-domain EEG features related to the decoded brain states, and promote the design of an analysis framework based on the network feature learning (addressed in Section 2.1.2).

Our neural network design is composed of three interpretable and trainable layers. Through these layers, the network progressively learns, from the input EEG (X), the most useful *frequency-domain features*, *spatial-domain features* and finally *time-domain features* for discriminating between the output brain states. Remarkably, thanks to the adoption of the proposed architecture, the EEG features learned by the network are forced to live within these domains and to be easily interpretable once the network training has ended. The neural network is presented in the following, separately for each type of feature learning, and it is illustrated in Fig. 1 and Table 1.

i. Frequency-domain feature learning. A temporal convolutional layer is applied to the input EEG (X), learning a bank of bandpass filters. To do so, each temporal convolutional kernel – $k[n], 0 \le n \le M-1$ – is modeled by forcing its values to describe a bandpass filter in the frequency domain. To this aim, the analytical expression for defining the frequency response of the bandpass filter has to be known (e.g., the expression of an ideal bandpass filter) and used to set the kernel values. Consequently, instead of exposing all the values of the kernel ($k[n], \forall n$) to the training process, only a small set of parameters θ_f are trained, containing the parameters that uniquely define the frequency response of the filter (e.g., the cutoff frequencies). By denoting with $K\left(f;\theta_f\right)$ the frequency response of the designed bandpass filter (parametrized in θ_f), its impulse response is obtained by computing the inverse Fourier transform $k\left(t;\theta_f\right) = \mathcal{F}^{-1}\left\{K\left(f;\theta_f\right)\right\}$. Finally, the kernel values $k\left[n;\theta_f\right]$, representing the coefficients of a finite-impulse–response filter, are obtained by windowing the impulse response using a Hamming window with $M = 63$ points.
Three different options were considered in this study to model the analytical expression of the bandpass filter frequency response, which are briefly introduced in the following and also illustrated in Fig. 2. Hence, three different versions of our network (and thus, of our analysis framework) were designed, each one differing as to the functional form used to express the frequency response of the bandpass filters.

   – Sinc-based functional form. This option forces the learning of ideal band-pass filters [37,40] (i.e., with an impulse response defined as the composition of sinc functions). The frequency response of such kernels is given by the composition of two heavy-side step functions (see also Fig. 2a):

$$K\left(f;f_1,f_2\right) = rect\left(\frac{f}{2f_2}\right) - rect\left(\frac{f}{2f_1}\right), \quad (1)$$

   where f_1 and f_2 denote the low and high cutoff frequencies of an ideal band-pass filter, respectively, representing the

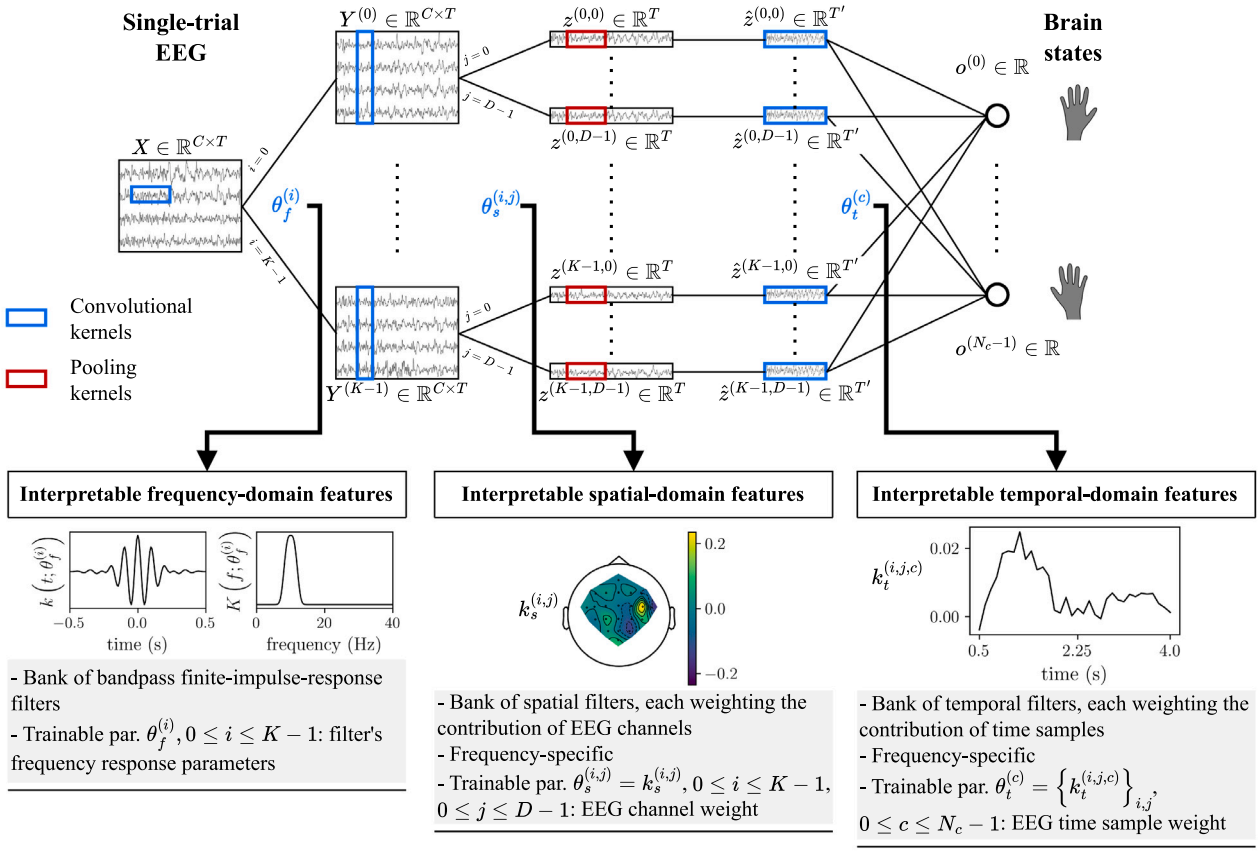

Fig. 1. Fully-interpretable convolutional neural network for decoding the EEG: high-level scheme. The network processes the input EEG progressively in the frequency, spatial, and temporal domains, enabling an easy understanding of the most salient EEG features for separating the brain states under analysis. A representative visualization of the EEG features learned by the network in each domain is provided on the bottom. See Section 2.1.1 for further details and the used symbology.

Table 1
Network architecture: layers, main hyperparameters, intermediate output shapes, and number of trainable parameters. If not specified differently, unitary stride (S) and no padding (P) are applied. The intermediate output shapes and number of parameters are expressed in terms of the network hyperparameters and of the input dimension (C: number of EEG channels, T: number of time samples). See also Section 2.1.1 for further details about the symbols. In the last column, we also provide within brackets the number of trainable parameters as resulting on average from our experiments, see Section 2.2. Here, a range of values is reported, depending on the functional form used to design the frequency response of bandpass filters.

Domain	Layer	Main hyperparameters	Output shape	No. of trainable parameters
	Input	–	$(1, C, T)$	0
Frequency	Frequency-Conv2D	$K = 32,\ F = (1, 63)$	(K, C, T)	$2K - 3K$ (64-96)
	BatchNorm2D	–	(K, C, T)	$2K$ (64)
Spatial	Space-Conv2D	$D = 2,\ F = (C, 1)$	$(KD, 1, T)$	KDC (1744)
	BatchNorm2D	–	$(KD, 1, T)$	$2KD$ (128)
	ELU	–	$(KD, 1, T)$	0
Temporal	AvgPool2D	$F = (1, 63),\ S = (1, 13)$	$(KD, 1, T')$	0
	Dropout	$p = 0.5$	$(KD, 1, T')$	0
	Time-Conv2D	$N_c \in \{2, 4\}$, $F = (1, T')$	$(N_c,)$	$KDT'N_c$ (6531)
	Softmax	–	$(N_c,)$	0
				8531–8563

frequency response parameters. Note that, similarly to the following approaches, the frequency response can be characterized in terms of the central frequency f_c and full-width half-maximum bandwidth h instead of the cutoff frequencies, where $f_c = (f_1 + f_2)/2$ and $h = f_2 - f_1$. The set of parameters $\theta_f = \{f_1, f_2\}$ (or equivalently $\theta_f = \{f_c, h\}$) constitutes the interpretable spectral parameters of the sinc-based frequency response.

– Morlet wavelet (MW)-based functional form. This option forces the learning of real Morlet wavelets [38]. The frequency response of such kernels is given by the composition of two Gaussian functions (see also Fig. 2b):

$$K(f; f_c, \xi) = \frac{1}{2}\left[e^{-\xi\pi^2(f-f_c)^2} + e^{-\xi\pi^2(f+f_c)^2}\right], \quad (2)$$

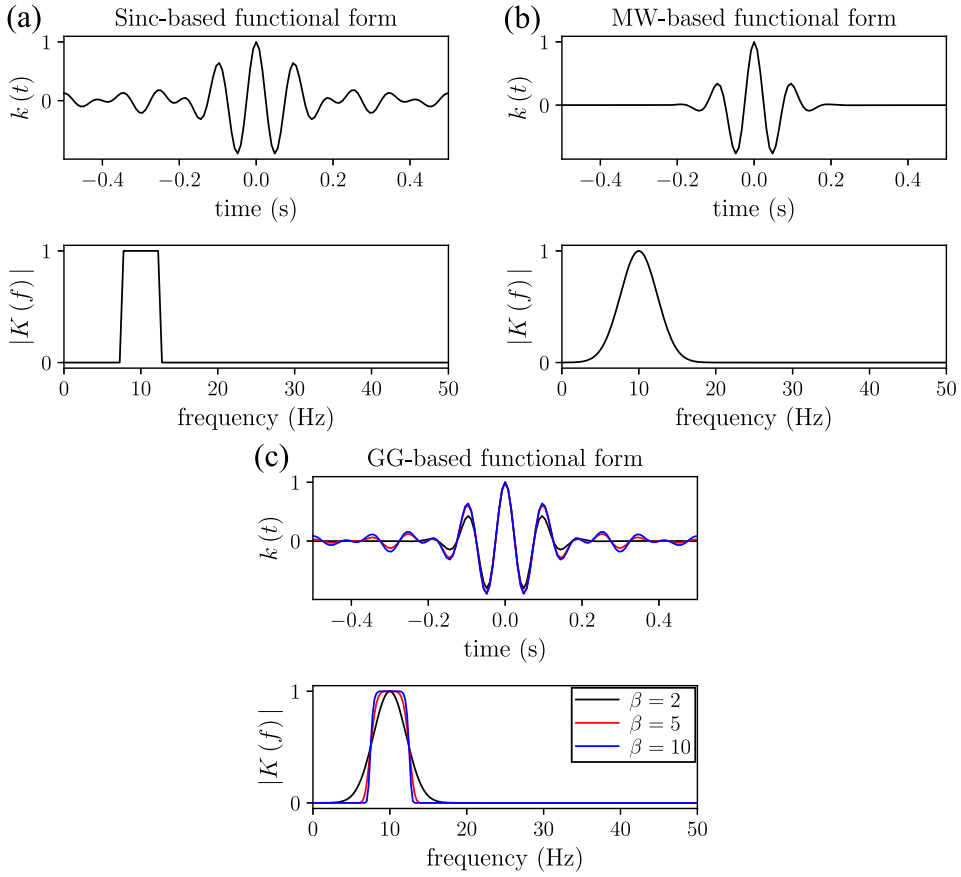

Fig. 2. Functional forms used for modeling the analytical expression of the frequency response of the bandpass filters. Here, a bandpass filter centered at f_c = 10 Hz with a full-width half-maximum bandwidth h = 5 Hz is designed, separately by exploiting sinc-based, Morlet wavelet (MW)-based, and generalized Gaussian (GG)-based functions (panels a–c). Both impulse responses and frequency responses of the filters are normalized with respect to their maximum value.

where f_c is the central frequency and ξ is the time decay of the wavelet (affecting the variance), representing the frequency response parameters. Similar to the previous case, the frequency response could be characterized in terms of the full-width half-maximum bandwidth h instead of the time decay parameter ξ [41], together with f_c. The set of parameters $\theta_f = \{f_c, \xi\}$ (or equivalently, $\theta_f = \{f_c, h\}$) constitutes the interpretable spectral parameters of the MW-based frequency response.

– Generalized Gaussian (GG)-based functional form. Here, kernels are forced to describe a generalized Gaussian function [39]. The frequency response of the kernels is given by (see also Fig. 2c):

$$\begin{cases} K\left(f; f_c, h, \beta\right) = e^{-\left(|f-f_c|/\alpha\right)^{\beta}} \\ \alpha = \frac{h}{2ln(2)^{1/\beta}}, \end{cases} \tag{3}$$

where f_c is the central frequency, h is the full-width half-maximum bandwidth and β is a shape parameter, representing the frequency response parameters. The set of parameters $\theta_f = \{f_c, h, \beta\}$ constitutes the interpretable spectral parameters of the GG-based frequency response.

For $\beta = 2$, $K\left(f; f_c, h, \beta\right)$ corresponds to the normal distribution (i.e., a complex Morlet wavelet in the time domain) [41]. For $\beta \to \infty$, $K\left(f; f_c, h, \beta\right)$ tends to an ideal sinc-based bandpass filter, centered at f_c with bandwidth h. Indeed:

$$\lim_{\beta\to\infty} K\left(f; f_c, h, \beta\right) = \lim_{\beta\to\infty} e^{-ln(2)\left(|f-f_c|/\frac{h}{2}\right)^{\beta}} = \begin{cases} 1, |f-f_c| < \frac{h}{2} \\ \frac{1}{2}, |f-f_c| = \frac{h}{2} \\ 0, |f-f_c| > \frac{h}{2}. \end{cases} \tag{4}$$

Therefore, with this functional form, the filter shifts from a Morlet-based representation to a sinc-based representation, by increasing the parameter β.

These three different types of filters, due to their characteristics in the frequency and time domains, intrinsically have different advantages and disadvantages. The sinc-based formulation realizes ideal band-pass filters with sharp transition from stopband to passband, thus providing high frequency selectivity, but possibly introducing ringing and oscillations in the time domain particular in presence of abrupt edges in the signals. Conversely, the MW formulation being characterized by a Gaussian-shape in the frequency domain and thus without sharp transitions, minimizes ringing and artifactual oscillations in the time domain, but at cost of reduced frequency selectivity which makes the filter less ideal. The GG formulation has the clear advantage of higher flexibility since the trainable shape parameter β allows the filter to move from a Gaussian shape to a rectangular shape during the course of training. Of course, these differences could traduce in different decoding performance across the three network variants, when applied to a specific EEG decoding problem, since one implementation could be more appropriate than the other

to extract useful discriminative features, also depending on the nature and patterns of the input EEG signals.

The coefficients of the finite-impulse–response filter $k\left[n;\theta_f\right]$ depend on the parameters contained in θ_f. These parameters are learned during network training and represent the *interpretable frequency-domain EEG features*. The coefficients $k\left[n;\theta_f\right]$ are then used as values of the convolutional kernel (consisting of $M = 63$ values), and applied to each EEG time series $x[n]$, one per channel, obtaining the bandpass filtered time series $y[n]$:

$$y[n] = (x * k)\left[n;\theta_f\right] = \sum_{m=0}^{M-1} x[n-m]\, k\left[m;\theta_f\right]. \quad (5)$$

Here, zero-padding is applied to produce an output time series with the same dimension of the input. Finally, by stacking each filtered time series, the filtered multi-variate time series is obtained: $Y = \left[y_0[n], \ldots, y_{C-1}[n]\right] \in \mathbb{R}^{C\times T}$. For brevity, the previous considerations were made considering only one bandpass filter; however, the network learns a bank of $K = 32$ bandpass filters. Therefore, the trainable parameters are $\theta_f^{(i)}, 0 \le i \le K-1$. Thus, this layer learns how to optimally filter the input time series, and produces as output a collection of K versions of the input ($Y^{(i)} \in \mathbb{R}^{C\times T}$), each filtered adopting a different bandpass filter given by $\theta_f^{(i)}$.

ii. Spatial-domain feature learning. A depthwise spatial convolutional layer is applied to the filtered multi-variate time series $Y^{(i)}$, learning a set of $D = 2$ spatial convolutional kernels separately for each bandpass filtered version of the input (i.e., $\forall i$, thus, 64 kernels in total) [37]. In other words, this layer learns the optimal combination of EEG channels for predicting the output brain states, separately for each bandpass-filtered signal. Therefore, the trainable parameters are the convolutional kernels $k_s^{(i,j)}$, $0 \le i \le K-1$, $0 \le j \le D-1$, used to weight each EEG channel, representing the *interpretable spatial-domain EEG features* $\theta_s^{(i,j)} = k_s^{(i,j)} \in \mathbb{R}^C$. Thus, the convolution utilizes kernels with C values – that is, with a number of values equal to the number of EEG channels to combine – and produces as output the time series $z^{(i,j)}[n]$, $0 \le i \le K-1$, $0 \le j \le D-1$ ($z^{(i,j)} \in \mathbb{R}^T$) by recombining in space the bandpass filtered version of the input EEG:

$$z^{(i,j)}[n] = \sum_{m=0}^{C-1} Y^{(i)}[m,n]\, k_s^{(i,j)}[m], \quad (6)$$

where i and j represent the indices of the applied bandpass filters ($0 \le i \le K-1$) and spatial filters ($0 \le j \le D-1$), respectively.

iii. Time-domain feature learning. The time series contained in $z^{(i,j)}[n]$ are pooled in time for reducing the computational cost, by employing an average pooling layer with a pooling size and stride of approximately 500 ms and 100 ms, respectively (corresponding to 63 and 13 time samples with the adopted sampling rate, see Section 2.2.1). This results in the pooled time series $\hat{z}^{(i,j)}[n]$ ($\hat{z}^{(i,j)} \in \mathbb{R}^{T'}$, $T' < T$). Then, a temporal convolutional layer is applied; it learns a set of N_c kernels (one per output brain state), for optimally recombining the temporal information contained in each time series of $\hat{z}^{(i,j)}$. In this study, we applied the network to both binary and multi-class classification problems, with 2 or 4 output classes (i.e., $N_c \in \{2,4\}$), see also Section 2.2.1.

Therefore, the convolutional kernels $k_t^{(c)}$, $0 \le c \le N_c - 1$, used to weight each time sample of the processed time series, are trained during network training and represent the *interpretable time-domain EEG features* $\theta_t^{(c)} = k_t^{(c)} = \left\{k_t^{(i,j,c)}\right\}_{i,j}, k_t^{(i,j,c)} \in \mathbb{R}^{T'}$. The convolution utilizes kernels covering all time samples (T') and all intermediate features (KD), producing as output a single scalar value $o^{(c)} \in \mathbb{R}$ for each output brain state ($\forall c$), according to:

$$o^{(c)} = \sum_{i=0}^{K-1}\sum_{j=0}^{D-1}\sum_{m=0}^{T'-1} \hat{z}^{(i,j)}[m]\, k_t^{(i,j,c)}[m]. \quad (7)$$

Importantly, this layer finalizes the classification problem, mapping $\hat{z}^{(i,j)}$ to the output brain states.

Based on the previous description, the interpretable features are directly accessible in the parameter set $\theta = \left\{\theta_f^{(i)}, \theta_s^{(i,j)}, \theta_t^{(c)}\right\}, \forall i,j,c$. In addition to the processing steps presented above, batch normalization [42] is included after the first temporal convolution and after the spatial depthwise convolution. Dropout [43] (dropout rate $p = 0.5$) is applied immediately before the last convolutional layer. The network neurons are activated via an exponential-linear function after the spatial depthwise convolution, and via a softmax function after the output convolutional layer. Thus, the network provides as output the conditional probability $p(l|X)$. Finally, the main hyper-parameters of the network – such as the number of samples of the windowed impulse response (M), number of bandpass filters (K), depthwise multiplier (D), pooling size and stride, and the dropout rate – are set as in previous studies exploiting deep neural networks (either interpretable and non-interpretable) with EEG signals [19,20,37,40].

2.1.2. Processing of the learned features: computation of multi-domain EEG signatures characterizing the cognitive task

Neural networks were trained separately for each participant to discriminate different brain states characterizing a cognitive task from minimally pre-processed single-trial EEG signals (see later Section 2.2 for further details about the conducted experiments). To highlight the signatures of EEG oscillations characterizing the cognitive task, the multi-domain EEG features contained in θ of each trained model were processed as follows.

i. Frequency-domain feature processing. Frequency-domain features were processed to highlight the frequency components that enable a better separation between the output brain states. From the frequency-domain features ($\theta_f^{(i)}, \forall i$), we derived the probability that a specific frequency component was processed by the network. To do so, the frequency axis was discretized ($F = 500$ points) within $\left[f_{min}, f_{max}\right]$ – denoting with f_{min} and f_{max} the frequency limits of the pre-processed time series – and the probability that each frequency component was included in the passband of a filter was computed, by identifying the passband from $f_1 = f_c - h/2$ to $f_2 = f_c + h/2$. This results in a *frequency-level probability* ($\in \mathbb{R}^F$), summarizing the frequency-domain processing of the model.

ii. Spatial-domain feature processing. Spatial-domain features were processed to highlight the contribution of different EEG channels to discriminate between the output brain states, separately for each frequency component. The absolute value of spatial features ($\theta_s^{(i,j)}, \forall i,j$) was computed, as we aspired at quantifying the strength of the contribution of different spatial sites for the discrimination. The absolute value of convolutional kernels has been used in prior works [26,37,44] to quantify the 'discriminatory power' of features, e.g., the higher the kernel sample the higher the contribution of that sample to discriminate between the output brain states. As explained in Section 2.1.1, each bandpass filter is linked to a set of D spatial kernels. We derived the *frequency-level spatial discriminatory power*, quantifying the discriminatory power in space separately for each frequency component. To do so, the frequency axis was discretized as in point i., and for each frequency component, we averaged together the spatial kernels linked to bandpass filters containing that frequency component in their passband (i.e., within $\left[f_1 = f_c - h/2, f_2 = f_c + h/2\right]$). This way, we obtained a representation $\in \mathbb{R}^{F\times C}$ that quantifies how much each EEG channel, for each frequency component, contributes to the discrimination. Finally, to summarize the information across frequency components, we also derived the *spatial discriminatory power*, by averaging the frequency-level spatial discriminatory power across all the frequency bins, obtaining a representation $\in \mathbb{R}^C$.

iii. Temporal-domain feature processing. Temporal-domain features were processed to highlight the contribution of different time samples to discriminate between the output brain states, separately for each frequency component. It is worth remarking that each time-domain feature ($\theta_t^{(c)}, \forall c$) weights the time samples separately for each frequency-specific spatially filtered time series (i.e., for each frequency interval processed by the bandpass filters). The absolute value of $\theta_t^{(c)}, \forall c$ was computed, as we were interested in analyzing the strength of the contribution of different time samples for the discrimination, and then was averaged across the different temporal-domain features (i.e., temporal kernels), summarizing the information across brain states. We derived a *frequency-level temporal discriminatory power*, quantifying the discriminatory power in time separately for each frequency component. To do so, the frequency axis was discretized as in point i., and for each frequency component, we averaged together the temporal discriminatory power linked to bandpass filters containing that frequency component in their passband (i.e., within $\left[f_1 = f_c - h/2, f_2 = f_c + h/2\right]$). This way, we obtained a representation $\in \mathbb{R}^{F \times T'}$ that quantifies how much each time sample, for each frequency component, contributes to the discrimination. Finally, similar to point ii., to summarize the information across frequency components, we also derived the *temporal discriminatory power*, by averaging the frequency-level temporal discriminatory power across all the frequency bins, obtaining a representation $\in \mathbb{R}^{T'}$.

Overall, this procedure enables our framework to highlight the most salient frequency-, spatial-, temporal-domain signatures of the EEG signals that better allow the prediction of the output brain states (e.g., left-hand vs. right-hand motor imagery) characterizing the cognitive task under investigation (e.g., motor imagery), that is, the signatures that are modulated during the cognitive task.

2.2. Experiments

2.2.1. EEG data and pre-processing

We conducted our experiments on three public motor imagery datasets widely adopted in the literature: BNCI2014-001 [45], BNCI2014-004 [46], Lee2019-MI [47]. Overall, 72 participants and 22500 EEG trials were considered. A description of these public datasets is provided in the following. Please refer to the original publications for further details about data recording and acquisition.

i. BNCI2014-001 [45]. This dataset consists of 22-channel EEG recorded from 9 healthy participants across 2 recording sessions. Signals were recorded from Fz, FC3, FC1, FCz, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP3, CP1, CPz, CP2, CP4, P1, Pz, P2, POz ($C = 22$). The task consisted in the imagination of feet, left-hand, right-hand and tongue movements for 4 s. Before the task, a 2s-length rest interval was included, in which the participant fixated a fixation cross displayed on the screen. For each participant and each session, 288 trials were recorded, balanced between classes. EEG signals were sampled at 250 Hz.
ii. BNCI2014-004 [46]. Here, 3-channel EEG was collected from 9 healthy participants in 5 recording sessions. Signals were recorded from C3, Cz, C4 ($C = 3$). The task consisted in the imagination of left-hand and right-hand movements for 4.5 s. Before the task, a 3s-length rest interval was included. For each participant and each session, approximately 145 trials were recorded (on average across participants and sessions), balanced between classes. EEG signals were sampled at 250 Hz.
iii. Lee2019-MI [47]. In this dataset, 62-channel EEG was recorded from 54 healthy participants across 2 recording sessions. Electrodes were placed according to 10-10 international system ($C = 62$). The task consisted in the imagination of left-hand and right-hand grasping for 4 s. Before the task, a 3s-length rest interval was included, in which the participant fixated a fixation cross displayed on the screen. For each participant and each session, 100 trials were recorded, balanced between classes. EEG signals were sampled at 1000 Hz.

Signals of all datasets were minimally pre-processed by: (i) downsampling to 125 Hz, to reduce the computational cost; (ii) band-pass filtering between $f_{min} = 4$ and $f_{max} = 40$ Hz (4th order zero-phase infinite-impulse–response Butterworth filter), to select the frequency range covering from theta (4–8 Hz) to low-gamma (30–40 Hz) EEG bands; (iii) epoching between 0 and 4 s with respect to the motor imagery onset ($T = 500$ time samples). After the pre-processing, each EEG trial was represented as a 2D matrix with shape ($C \in \{3, 22, 62\}, T = 500$). The single-trial EEG represented the network input, while its corresponding motor imagery state represented the network output.

2.2.2. Multi-domain characterization of motor imagery with the proposed framework

Neural networks were randomly initialized and trained to classify between distinct motor imagery states ($N_c = 4$ classes for BNCI2014-001 dataset, while $N_c = 2$ classes for the other two datasets). The categorical cross-entropy was used as loss function. Trainable parameters were optimized via mini-batch stochastic gradient descent using adaptive moment estimation [48] (up to 500 epochs, learning rate of 1e-3, mini-batch size of 32). Early stopping was performed, stopping the training once the validation loss did not decrease anymore. During training, the central frequency of bandpass filters f_c was constrained to be in the frequency range $\left[f_{min}, f_{max}\right]$ (i.e., given by the EEG pre-processing), the full-width half-maximum bandwidth h was constrained to be within the range $[2, 20]$ Hz, and the shape parameter β (used in the GG-based frequency response only) was constrained to be within the interval $[2, 10]$, as suggested by previous studies [37,39].

Participant-specific decoders were trained, utilizing a leave-one-session-out cross-validation. That is, for each participant, one recording session was reserved as the test set, and the remaining sessions represented the training and validation sets (split with a 80%–20% ratio). This is equivalent to a session-level cross-validation scheme, where the number of cross-validation folds is equal to the number of total sessions. The adopted data split resulted into 232, 464, 80 training trials, 56, 116, 20 validation trials, and 288, 145, 100 test trials respectively for the BNCI2014-001, BNCI2014-004, Lee2019-MI datasets, on average across participants and cross-validation folds. We have chosen to rely on participant-specific decoders as the proposed approach, based on the network learned features, is intended to be used as a tool for revealing participant-specific neural signatures. This is of extreme relevance as this approach, in prospective, may be also extended to patients' signals, highlighting patient-specific neural signatures. Moreover, previous studies [28,44] have shown that participant-specific decoders outperform cross-participant decoders; therefore, the features learned by participant-specific decoders are more likely to be related to neurophysiological/neuropathological EEG characteristics.

For each trained model, i.e., for each participant and each cross-validation fold, we applied the processing procedure to the learned interpretable features as described in Section 2.1.2. Thus, we derived the frequency-level probability, the spatial discriminatory power representations and the temporal discriminatory power representations for each model. Then, these were averaged across cross-validation folds, deriving variables that characterize the motor imagery neural modulations in multiple domains, in a participant-specific way. The results provided by the analysis performed with our framework were assessed (see Section 4.1) in light of the modulations known in literature occurring during motor imagery, and also compared with the modulations we obtained on the same datasets using a classic analysis approach based on event-related spectral perturbations (see Appendix A).

2.2.3. Decoding performance and comparison with state-of-the-art approaches

To quantify the quality of the multi-domain EEG feature learning operated by our fully-interpretable CNN, we computed the decoding performance. First, we computed the accuracy for classifying among all the available classes characterizing each dataset (i.e., 4-class classification for BNCI2014-001 and 2-class classification for BNCI2014-004 and Lee2019-MI). Additionally, as the left-hand and right-hand motor imagery conditions were common across all datasets and as it is common in the literature to address the binary classification of left-hand vs. right-hand motor imagery [18,49], we also estimated the accuracy for the 2-class classification of left-hand vs. right-hand motor imagery conditions. In this case, for the BNCI2014-001 dataset (originally with 4 classes) we limited to consider the output neurons' prediction of left-hand and right-hand motor imagery only. The accuracy was computed on the test set for each trained network. Then, it was averaged across cross-validation folds (i.e., across sessions) separately for each participant, resulting in an average performance value for each participant. The accuracy scored by our network – in all its three variants (sinc-based, MW-based, and GG-based) – was compared with the chance level, separately for each decoding problem (i.e., dataset BNCI2014-001 4-class decoding, BNCI2014-001 2-class decoding, BNCI2014-004 2-class decoding, Lee2019-MI 2-class decoding). Here, to compare the performance with the chance level (0.25 or 0.5, respectively for 4-class decoding and 2-class decoding), non-parametric pairwise Wilcoxon signed-rank tests were applied [50], corrected for multiple tests using the Benjamini–Hochberg procedure [51] (3 in total). Finally, we compared the accuracy scored by the three variants of the proposed network by using the non-parametric statistical test developed by M. Friedman [52] (significance level of 0.05). As no significance was found ($p \geq 0.05$) when comparing the three variants of our network, post-hoc pairwise comparisons were not performed.

Moreover, to also provide a comparative analysis with state-of-the-art approaches – even though a benchmark analysis about neural decoding is not the main aim of this study – our proposed fully-interpretable neural network was compared with other 5 state-of-the-art interpretable networks and other 6 state-of-the-art non-interpretable networks. The state-of-the-art interpretable models were: Sinc-EEGNet [37] (sinc-based architecture, inspired from the widely used design of EEGNet [20]), WaSFCNN [38] (MW-based architecture), and finally MagEEGminer, CorrEEGminer, PLVEEGminer [38] (GG-based architectures). As concerning the interpreted domains, the state-of-the-art designs mainly enable the interpretation of frequency-domain EEG features, with the exception of the first design (Sinc-EEGNet) that enables the interpretation of frequency-specific spatial features too. On the other hand, importantly, our network design (in all the three variants) enables the analysis of the EEG features in all frequency, spatial and temporal domains (see Section 2.1.1). The state-of-the-art non-interpretable models were: EEGNet [20], ShallowConvNet [19], DeepConvNet [19], EEGITNet [53], EEGInception [30], and EEGConformer [33]. These are CNNs commonly adopted for neural decoding, that also won international EEG decoding competitions (EEGNet) [21, 27] and that also employ recent methodologies devoted at improving neural decoding, e.g., multi-scale temporal feature learning via parallel temporal convolutions (EEGInception) and attention-based mechanisms to realize convolutional transformers (EEGConformer). For each decoding problem, we compared the accuracy of each variant of the proposed fully-interpretable CNN with the accuracy of state-of-the-art networks, by first detecting differences across multiple networks using the non-parametric statistical test developed by M. Friedman [52], separately for state-of-the-art interpretable and non-interpretable networks (significance level of 0.05). Then, as significance ($p < 0.05$) was found, for each decoding problem, post-hoc pairwise comparisons were performed between the accuracy of each variant of our neural network and each state-of-the-art network, adopting the non-parametric Wilcoxon signed-rank test [50] corrected for multiple tests (5 and 6, in total, respectively in case of the state-of-the-art interpretable and non-interpretable networks) with the Benjamini–Hochberg procedure [51]. In addition to the network performance, for each neural network we computed the number of trainable parameters (quantifying the model size), and the percentage of the interpretable trainable parameters (quantifying the level of interpretability of the model). The latter was computed as the ratio between the number of interpretable trainable parameters and the total number of trainable parameters (expressed in percentage). We also computed the training time and inference time, quantifying the computational time during training and testing, respectively. These are expressed as the time to process a mini-batch of examples (32 EEG epochs) on GPU (NVIDIA Titan V), respectively during training and inference. Therefore, we performed a comprehensive comparative analysis with the state-of-the-art in terms of network performance (accuracy), size (number of parameters), interpretability (interpreted domains and percentage of interpretable parameters), and computational time (training and inference times).

Finally, we performed channel pruning experiments, to compare our interpretable decoders when using the same dataset but selecting different small subsets of electrodes among a relative large number of available electrodes. This may provide valuable insights for selecting simpler EEG layouts to use in practical applications, and for reducing the model size (e.g., for prospective applications that exploit devices with limited capabilities). To this aim we considered the dataset Lee2019-MI with 62 channels available and we selected 6 subsets of electrodes, with different degrees of coverage over the sensorimotor cortex. For each subset, we trained our GG-based fully interpretable network, and we computed the number of trainable parameters and the decoding accuracy, with the same procedure adopted in the previous experiments. Then, for each subset of electrodes, the obtained decoding accuracy was compared with the decoding accuracy obtained when using all 62 available channels (reference condition), using a non-parametric pairwise Wilcoxon signed-rank test [50] corrected for multiple comparisons (6 in total) with the Benjamini–Hochberg procedure [51].

3. Results

3.1. Multi-domain characterization of motor imagery

In all decoding problems, our fully-interpretable CNN (in each of the three variants) performed well above the chance level (see Section 3.2), indicating the learning of meaningful features for motor imagery classification. Here, since our main aim is to design a CNN-based framework for EEG analysis, we first report the results about the learned features and the multi-domain EEG signatures of motor-imagery derived from the learned features. In particular, since no significant difference in decoding performance (see Section 3.2) was found among the three variants of our fully-interpretable CNN (sinc- vs. MW- vs. GG-based networks), for brevity, here we report only the EEG analysis results obtained using the GG-based network. The results obtained with the other two variants of our network can be found in Appendix B. We show here the results obtained with the GG-based network as this can be considered a more general case; indeed, the bandpass filtering based on generalized Gaussian functions can describe frequency responses resembling both the sinc-based and MW-based frequency responses, depending on the shape parameter β (see Section 2.1.1). To better illustrate this aspect, Fig. 3 shows the distribution of β (Fig. 3a) and the probability density function of β (Fig. 3b) across all trained bandpass filters (i.e., filters of all the trained models, across all participants, cross-validation folds, and decoding problems). Interestingly, the shape parameter β was automatically selected by the network such that a frequency response closer to an ideal bandpass filter was described (i.e., higher β values), that is, closer to response modeled with a sinc-based functional form.

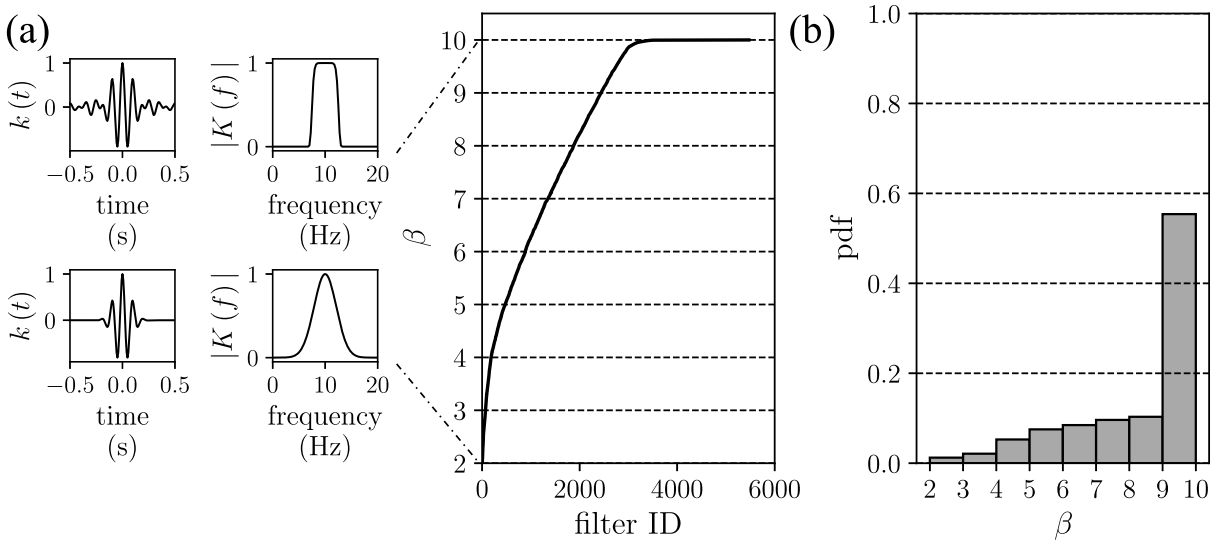

Fig. 3. Generalized Gaussian-based frequency response: shape parameter (β). Panel a – Distribution of β values across all the filters of the trained models (32 filters per model, overall 5472 filters). Here, both the impulse response and the frequency response for the upper and lower limits of β (i.e., $\beta = 2$ and $\beta = 10$) are also displayed (both normalized with respect to their maximum value). Panel b – Probability density function (pdf) of the parameter β.

In Fig. 4 we showcase a representative example of the multi-domain interpretable features (i.e., the features contained in θ) extracted from one trained network. From the frequency-domain EEG features (Fig. 4a), it is possible to directly infer how the trained bank of bandpass filters processed the input EEG. Here, a prominent involvement of alpha band (8–13 Hz) can be observed, with almost half of bandpass filters being learned with their central frequency within this range. The frequency-specific spatial filters (Fig. 4b), on the other hand, quantify the contribution of the EEG channels (specifically linked to each ith bandpass filter). Here, higher weights (either positive or negative) were linked to central electrode sites, especially in the alpha band (e.g., spatial filters with $i = 15$). In theta ($i = 0$, 4–8 Hz) and low-gamma ($i = 31$, 30–40 Hz) bands, higher weights mainly involved more frontal electrode sites. Finally, the frequency-specific temporal filters (Fig. 4c) quantify the contribution of time samples (specifically linked to each ith bandpass filter), separately for each output brain state. Here, a clear modulation of the contribution of different time samples can be observed, with time samples within the first half of the EEG trial being weighted more (either positively or negatively).

In the following, we present the results obtained by the processing of the learned interpretable features, thus showing the multi-domain characterization of motor imagery operated by the proposed framework.

Fig. 5 shows the probability that the different frequency components of the input EEG were exploited by the network to predict the motor imagery states (i.e., frequency-level probability). Alpha-band frequency components (at around 12 Hz) resulted the most frequently used ones, consistently across datasets. Beta-band frequency components (13–30 Hz, e.g., at around 25 Hz for BNCI2014-004 dataset) and low-gamma-band frequency components were used too but with a lower probability.

Moving to the spatial domain, Fig. 6 shows the spatial discriminatory power representations. The frequency-level representation (Fig. 6a) highlights a strong involvement of central and central-parietal sites consistently across datasets, especially in the alpha band, e.g., C3, C4, CP3 and CP4 for BNCI2014-001, C3 and C4 for BNCI2014-004, and C3, Cz and C4 for Lee2019-MI. This is also confirmed when considering the spatial discriminatory power (Fig. 6b) – obtained as the average value across frequencies – showing the highest contribution at central and central-parietal sites when imagining different actions.

Finally, Fig. 7 reports the temporal discriminatory power representations. Similarly to the previous results in the spatial domain, the frequency-level representation (Fig. 7a) shows a strong involvement of time samples processed within the alpha band, consistently across datasets. Moreover, a peculiar dynamic of the discriminatory power is observed, especially while analyzing the temporal discriminatory power (Fig. 7b), obtained as the average value across frequencies. In particular, the network weighted most the time samples within the first half of the trial in the first two datasets (BNCI2014-001 and BNCI2014-004), with a discriminatory power peaking at approximately 1.25 s after the motor imagery onset. On the other hand, the network weighted equally the time samples along almost the entire EEG trial length (i.e., a sustained importance of time samples along the trial) in the third dataset (Lee2019-MI).

In the previous 2-D representations (Figs. 6a and 7a), we reported the frequency-level discriminatory power separately in the spatial and temporal domains. Therefore, we separated the EEG signatures linked to the frequency, spatial, and temporal domains into two different 2-D representations, depicting the frequency-spatial and frequency-temporal planes. Importantly, these 2-D representations summarize the 3-D frequency-spatial–temporal representation derived from our approach, and were realized to conveniently visualize the EEG signatures as 2-D heatmaps, separately in different pairs of domains. To furnish an even more comprehensive overview of the feature analysis provided with our network, we computed a spatio-temporal representation quantifying the importance of spatial and temporal samples in the alpha band. To do so, we combined the previous frequency-level spatial and frequency-level temporal discriminatory powers in order to show the spatial and temporal discriminatory power averaged across the frequency bins of the alpha band (8–13 Hz). The resulting representations are reported in Fig. 8. The results reflect the spatial modulation and temporal dynamics of the discriminatory power in the alpha band, as already emerging separately in Figs. 6 and 7.

3.2. Decoding performance and comparison with state-of-the-art approaches

We accompany the investigation of the most salient multi-domain EEG characteristics of motor imagery with a comparative analysis between our fully-interpretable CNN (abbreviated as 'FINet', i.e., 'Fully-InterpretableNet', in Tables 2 and 3) and state-of-the-art neural networks proposed for EEG decoding (either interpretable and non -interpretable). Table 2 reports the decoding accuracy, as scored for

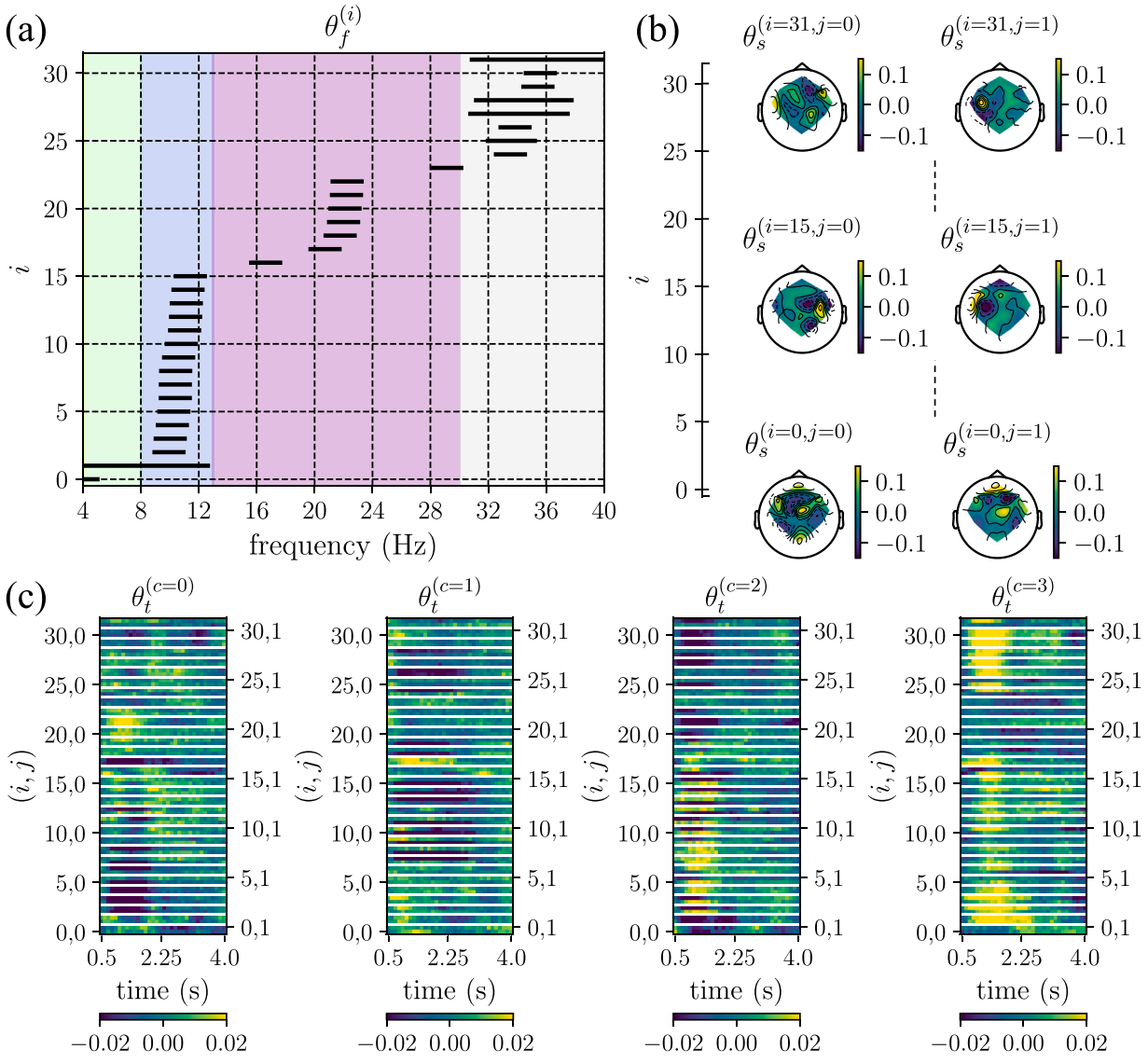

Fig. 4. Interpretable features extracted from a representative model. Here, the features associated to participant no. 9 of the BNCI2014-001 dataset for one cross-validation fold (i.e., the features relative to one model) are displayed. Panel a – Trainable bandpass filters. Horizontal bars denote the bandpass filters, with the endpoints representing the cutoff frequencies defined by the full-width half-maximum bandwidth h and central frequency f_c contained in $\theta_f^{(i)}$. EEG bands are color-coded as theta (4–8 Hz): green, alpha (8–13 Hz): blue, beta (13–30 Hz): purple, low-gamma (30–40 Hz): gray. Panel b – Frequency-specific spatial filters. For brevity, only three sets of spatial filters are visualized as heatmaps, linked to 3 (out of 32) bandpass filters, i.e., $\theta_s^{(i,j)}$, $i = 0$, $i = 15$, $i = 31$, $\forall j$. Panel c – Frequency-specific temporal filters. Here, the temporal filters are visualized for each bandpass filter and each spatial filter as a heatmap, separately for each motor imagery condition (representing the different brain states under analysis), i.e., $\theta_t^{(c)} = \left\{ k_t^{(i,j,c)} \right\}$, $\forall i$, $\forall j$.

each neural network separately in the BNCI2014-001 (4 classes, feet vs. left-hand vs. right-hand vs. tongue imagery), BNCI2014-001 (2 classes, left-hand vs. right-hand imagery), BNCI2014-004 (left-hand vs. right-hand imagery) and Lee2019-MI (left-hand vs. right-hand imagery) decoding problems. On the other hand, Table 3 reports the number of trainable parameters, the details about the feature interpretability (interpreted domains and percentage of interpretable parameters), and the training/inference times, for each network. Additionally, Fig. 9 graphically summarizes the comparative analysis between our network and the state-of-the-art interpretable networks, by representing, across the four decoding problems, the different neural networks inside the model performance-model size plane; here each network is represented by a dot and the circle size around each dot is modulated depending on the model interpretability.

Our fully-interpretable CNN – in all its three variants – performed significantly above the chance level ($p < 0.05$) in all decoding problems. Moreover, in each decoding problem, no significant differences ($p \geq 0.05$) were found between the three variants of our networks, with p-values of $p = 0.12$, $p = 0.09$, $p = 0.24$, $p = 0.54$, respectively for the four decoding problems. Therefore, despite the differences in the filter implementation of the first convolutional layer, the discriminative features relevant for motor imagery decoding were effectively captured by all the three network variants. As concerning the comparison with state-of-the-art interpretable networks, our networks mostly scored the highest accuracy scores when compared to the state-of-the-art, significantly outperforming state-of-the-art interpretable networks (see in Table 2 the markers *, †, ‡) for almost all the decoding problems (4) and the considered state-of-the-art networks (5), that is, 19 out of 20 times for our sinc-based and GG-based fully-interpretable CNNs, and 16 out of 20 times for the MW-based fully-interpretable CNN. Crucially, besides being superior in terms of decoding performance, our networks introduced the highest percentage of interpretable parameters (97.5%) compared to the state-of-the-art, which ranges from 0.4% to 37.4%.

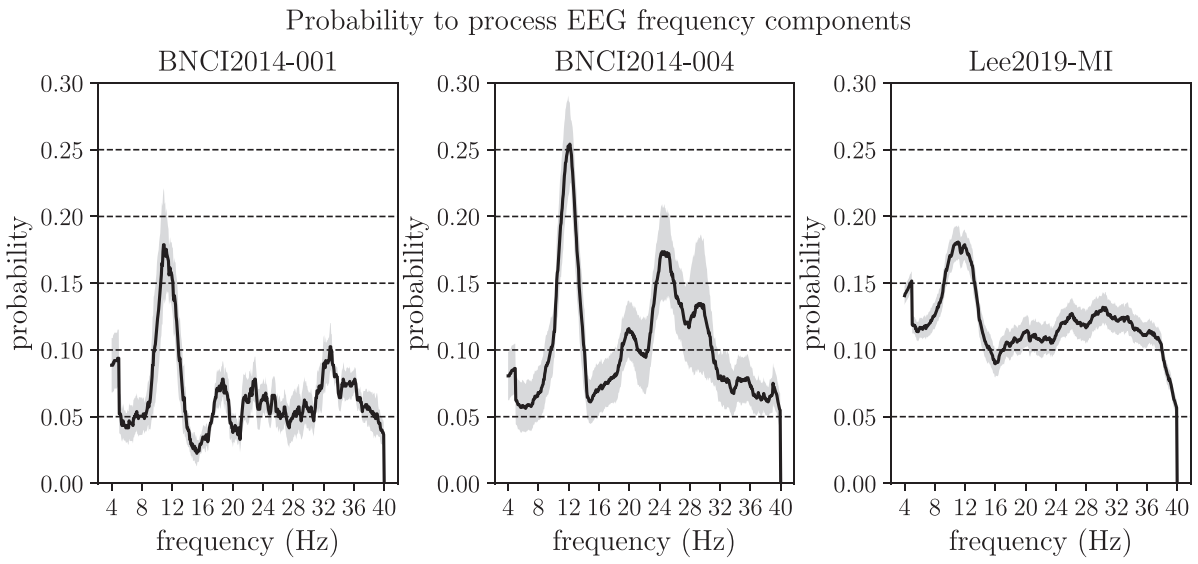

Fig. 5. Frequency-level probability. The figure displays the probability that each EEG frequency component was exploited by the network to predict motor imagery (i.e., a frequency component falling inside the bandwidth of the filters), separately for each dataset. The thick line represents the mean value across participants, while the shaded area denotes the standard error of the mean.

Table 2
Comparison with state-of-the-art neural networks for EEG decoding: decoding accuracy. The three variants of our proposed fully-interpretable CNN are labeled as: 'sinc-based FINet', 'MW-based FINet', and 'GG-based FINet', for brevity. Five interpretable neural networks are contrasted with our networks: SincEEGNet, WaSFCNN, MagEEGminer, CorrEEGminer, PLVEEGminer. Moreover, we also consider six non-interpretable neural networks in this comparative analysis: EEGNet, ShallowConvNet, DeepConvNet, EEGITNet, EEGInception, EEGConformer. The decoding accuracy scored by each network is listed (mean value across participants and standard error of the mean), separately for each decoding problem. The results from the conducted statistical analysis when comparing each of our networks with the state-of-the-art interpretable networks are reported too. Significant ($p < 0.05$) comparisons involving sinc-based, MW-based, and GG-based FINets are marked with the symbols *, †, ‡, respectively.

Neural network	Decoding accuracy			
	BNCI2014-001 (4)	*BNCI2014-001 (2)*	*BNCI2014-004*	*Lee2019-MI*
Sinc-based FINet (ours)	0.714 ± 0.044	0.841 ± 0.045	0.767 ± 0.046	0.652 ± 0.020
MW-based FINet (ours)	0.704 ± 0.046	0.834 ± 0.048	0.760 ± 0.049	0.657 ± 0.020
GG-based FINet (ours)	0.712 ± 0.041	0.829 ± 0.047	0.770 ± 0.047	0.657 ± 0.022
SincEEGNet [37]	0.621 ± 0.058 (*, †, ‡)	0.735 ± 0.058 (*, †, ‡)	0.743 ± 0.047 (*, ‡)	0.575 ± 0.014 (*, †, ‡)
WaSFCNN [38]	0.626 ± 0.065 (*, †, ‡)	0.778 ± 0.056	0.753 ± 0.045 (*, ‡)	0.616 ± 0.020 (*, †, ‡)
MagEEGminer [39]	0.467 ± 0.039 (*, †, ‡)	0.651 ± 0.038 (*, †, ‡)	0.716 ± 0.035 (*, ‡)	0.579 ± 0.012 (*, †, ‡)
CorrEEGminer [39]	0.495 ± 0.045 (*, †, ‡)	0.675 ± 0.041 (*, †, ‡)	0.665 ± 0.030 (*, †, ‡)	0.587 ± 0.011 (*, †, ‡)
PLVEEGminer [39]	0.452 ± 0.039 (*, †, ‡)	0.662 ± 0.040 (*, †, ‡)	0.570 ± 0.016 (*, †, ‡)	0.570 ± 0.012 (*, †, ‡)
EEGNet [20]	0.578 ± 0.058 (*, †, ‡)	0.755 ± 0.049 (*, †, ‡)	0.720 ± 0.043 (*, ‡)	0.613 ± 0.019 (*, †, ‡)
ShallowConvNet [19]	0.582 ± 0.050 (*, †, ‡)	0.760 ± 0.047 (*, †, ‡)	0.743 ± 0.046 (*, ‡)	0.643 ± 0.017
DeepConvNet [19]	0.518 ± 0.062 (*, †, ‡)	0.699 ± 0.067 (*, †, ‡)	0.761 ± 0.045	0.586 ± 0.016 (*, †, ‡)
EEGITNet [53]	0.599 ± 0.051 (*, †, ‡)	0.777 ± 0.041 (*)	0.771 ± 0.039	0.618 ± 0.017 (*, †, ‡)
EEGInception [30]	0.486 ± 0.047 (*, †, ‡)	0.697 ± 0.054 (*, †, ‡)	0.750 ± 0.038	0.615 ± 0.014 (*, †, ‡)
EEGConformer [33]	0.576 ± 0.064 (*, †, ‡)	0.718 ± 0.056 (*, †, ‡)	0.749 ± 0.046	0.664 ± 0.020

Moreover, our networks provided an increased interpretability in three distinct domains – that is, frequency, spatial, and temporal domains – while the state-of-the-art interpretable networks mainly limit the interpretability to the frequency domain (1-domain interpretability in 4 out of 5 state-of-the-art networks) and only one network (Sinc-EEGNet) focuses on both frequency and spatial domains (2-domain interpretability). The increased interpretability of our designs compared to the other interpretable networks, however, was associated with an increased number of trainable parameters (on overage across the four addressed decoding problems) and a general increase in the training and inference times. As concerning the comparison with state-of-the-art non-interpretable networks, the most remarkable result concerns the decoding accuracy. Indeed, also in this case our networks significantly outperformed the state-of-the-art most of the times (>50%), that is, 18 out of 24, 15 out of 24, and 17 out of 24 comparisons, respectively for the sinc-based, MW-based, and GG-based fully-interpretable CNNs. Importantly, in all other comparisons, the statistical differences resulted non-significant, thus, the decoding performance was statistically comparable. Therefore, our approaches resulted statistically superior or comparable to non-interpretable designs in terms of decoding performance.

Our experiments were conducted on datasets with different channel configurations, specifically 3, 22, and 62 EEG channels, in order to test the capabilities of our approach across different EEG layouts. However, we deemed it valuable to further inspect the effects of the channel layout, by conducting new dedicated experiments aimed at evaluating the predictive power of our approach when pruning channels starting from a layout with a large number of channels. Specifically, we considered the dataset with the highest number of EEG channels (Lee2019-MI, with 62 channels available) and two classes to be predicted (left vs. right hand motor imagery), and then we sampled a subset of EEG channels for solving the decoding problem. We considered six different subsets of EEG channels, ranging from 3 to 38 out of the 62 total channels available (Fig. 10 – middle). Three subsets were obtained using Cz, C3, C4 as seed channels, and including a progressively increasing number of adjacent channels, adopting a procedure similarly as in our previous

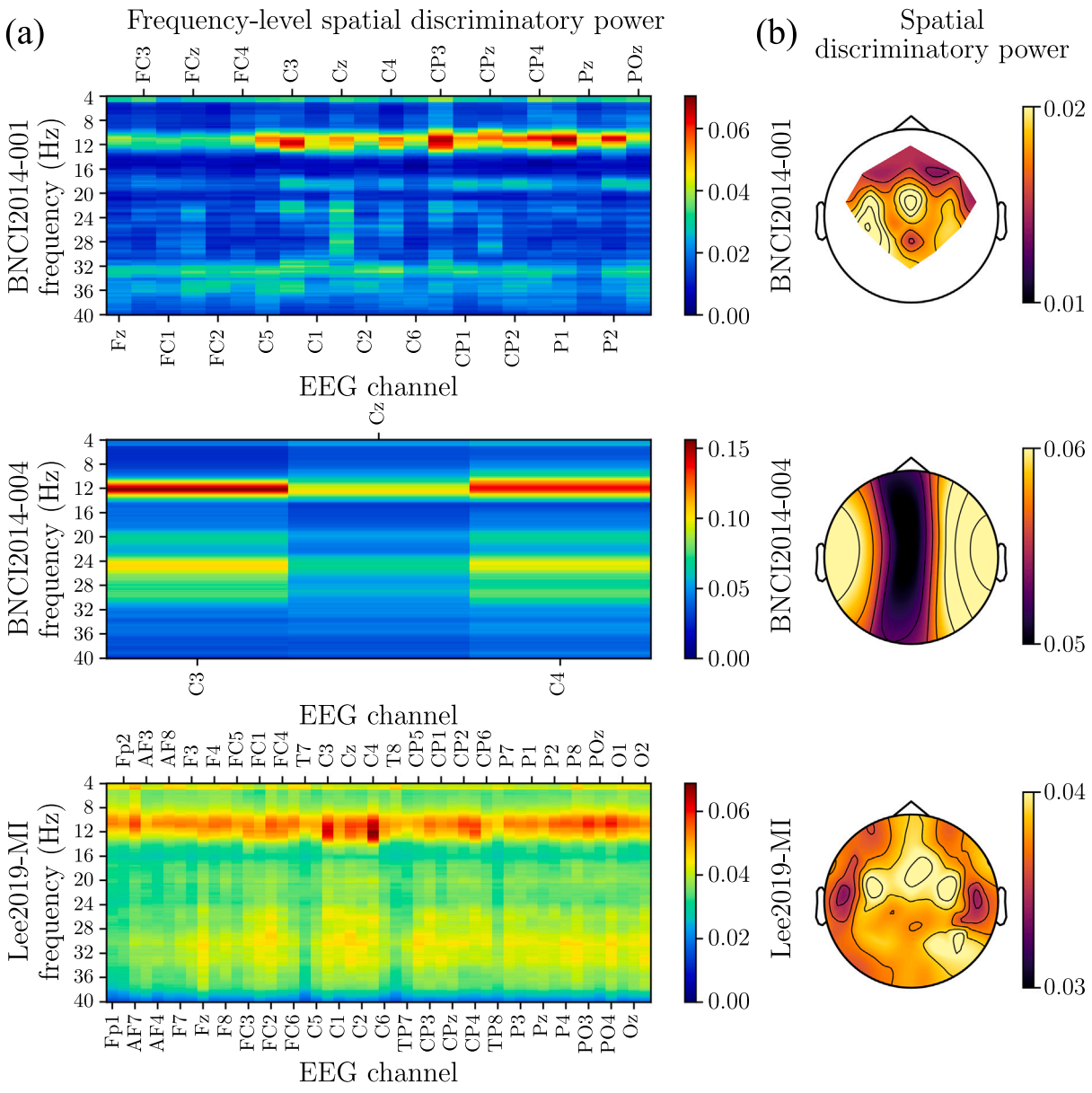

Fig. 6. Spatial discriminatory power representations. Panel a – Frequency-level spatial discriminatory power. Here, the contribution of each EEG channel to the discrimination of motor imagery states is reported, separately for each frequency component. For each dataset, the mean value across participants ($\in \mathbb{R}^{F \times C}$, where F are the frequency bins and C the EEG channels) is reported as a heatmap. Please note that for brevity, to improve the readability, for the Lee2019-MI dataset (comprising 62 channels), we removed the outermost channels from this visualization (while they remain present in panel b), that is, F10, F9, FT9, FT10, TP9, TP10, PO9, PO10. Panel b – Spatial discriminatory power. For each dataset, the frequency-level representation reported in panel a is averaged across the frequency components, and the mean value across participants is represented as a heatmap mapped onto the scalp.

study [18] (subsets with 3, 18, and 38 channels respectively); the other three subsets were obtained similarly but using only channel Cz as seed (subsets with 6, 18, and 37 channels respectively). These subsets were chosen as they approximately cover the sensorimotor cortex with a different coverage, exploiting from 4.8% to 61.3% of the total available EEG channels (62 in total). For each subset, our GG-based fully-interpretable network was trained and tested, as explained in Section 2.2.2, and the decoding performance (Fig. 10 – top) and number of trainable parameters (Fig. 10 – bottom) were computed for each subset. The number of trainable parameters varied from 4.8k to 7.1k depending on the number of channels in the subset, representing respectively 55.9% and 82.8% of the trainable parameters introduced when using all the 62 channels available (8.6k), thus, showing a considerable decrease. Furthermore, except for one configuration (6 channels around Cz) among the 6 configurations, the decoders trained with subsets of channels were statistically comparable with the decoder trained with all 62 channels (reference condition). Interestingly, using only 3 channels (4.8% of the 62 available channels) the neural decoder not only halved the number of trainable parameters relative to the reference configuration, but also provided statistically comparable performance.

4. Discussion

In this study, we propose a novel EEG analysis framework that exploits the automatic feature learning operated by a fully-interpretable

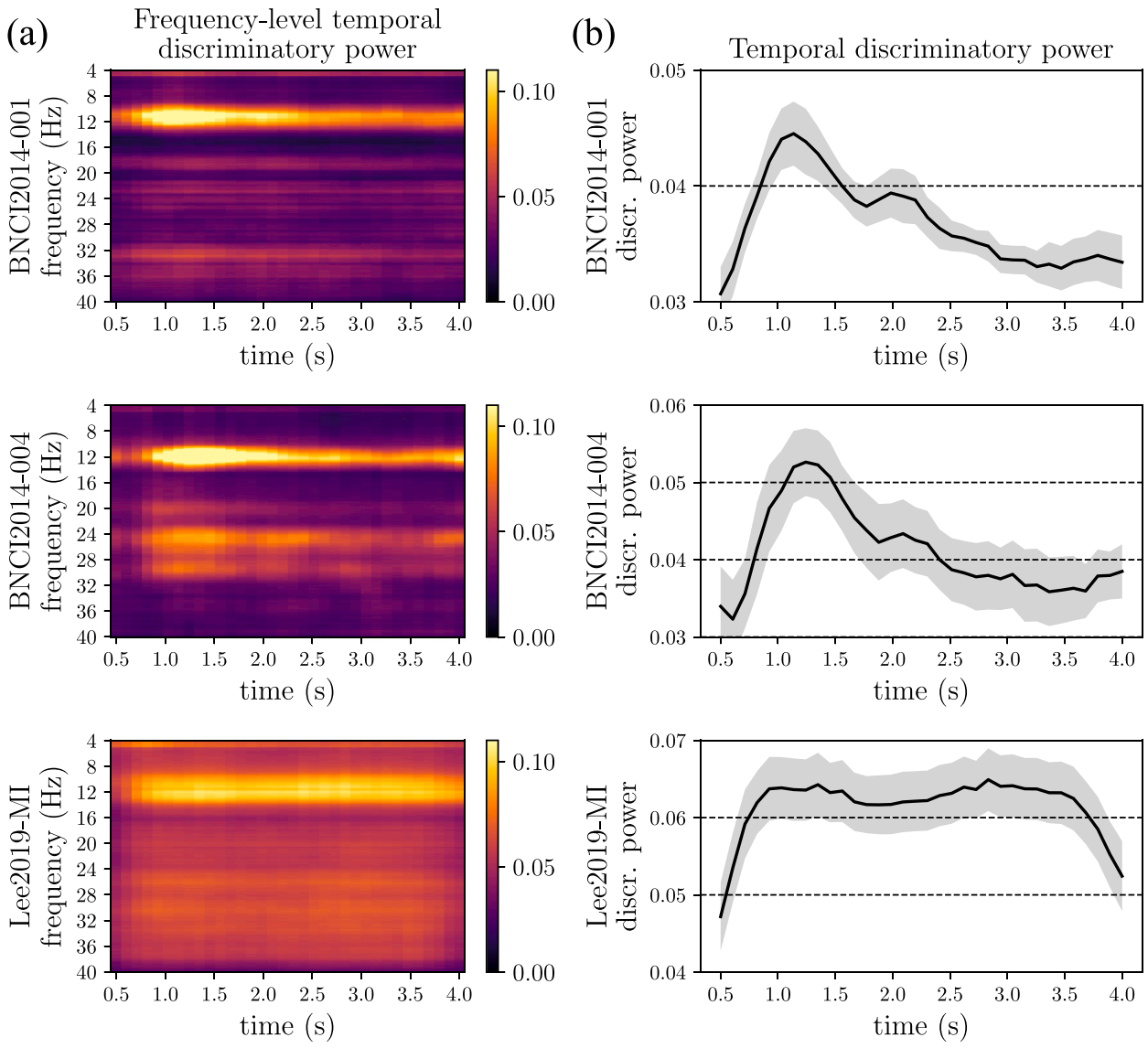

Fig. 7. Temporal discriminatory power representations. Panel a – Frequency-level temporal discriminatory power. Here, the contribution of each time sample to the discrimination of motor imagery states is reported, separately for each frequency component. For each dataset, the mean value across participants ($\in \mathbb{R}^{F \times T'}$, where F are the frequency bins and T' the temporal samples) is reported as a heatmap. Panel b – Temporal discriminatory power. For each dataset, the frequency-level representation reported in panel a is averaged across the frequency components. The thick line represents the mean value across participants, while the shaded area denotes the standard error of the mean. Please note that in both panels, the time axis is discretized starting from 0.5 s, as the network processed chunks of 500 ms in the last temporal convolutional layer, due to average pooling. For example, the first sample of the temporal convolutional kernel $k_t^{(i,j,c)}$ processes a portion from 0 to 500 ms of the input.

convolutional neural network, for characterizing the EEG oscillations during a cognitive task in the frequency, spatial, and temporal domains. Remarkably, to the best knowledge of the authors this is the first attempt of exploiting the feature learning operated by an interpretable network for characterizing neural correlates in multiple domains in an automatic, optimal, and end-to-end way. We tested the potentialities of our framework by applying it to real data acquired during a representative cognitive task (motor imagery). Finally, although the main focus of our study is on a CNN-based tool for EEG *analysis* rather than on a CNN-based tool for EEG *decoding*, for completeness, we also compared our fully-interpretable CNNs with the main state-of-the-art interpretable networks proposed for EEG decoding.

4.1. Multi-domain characterization of motor imagery

The proposed approach was tested on motor imagery, which was exploited as benchmark cognitive task. Indeed, in the literature significant research efforts have been devoted to EEG-based motor imagery decoding and analysis, due to the important implications of motor imagery for neurorehabilitation and assistive technologies. Different approaches, ranging from more traditional ones for example based on event-related desynchronization/synchronization (ERD/ERS) [54–56], to machine learning (e.g., see [45,57]) and deep learning techniques (e.g., see [19,40]), have been employed to characterize/decode motor imagery, and all these methodologies underline the importance of considering multi-domain information (temporal, spatial, spectral) for an effective characterization/classification. Therefore, we deemed motor imagery as an ideal test case for assessing the potential of our framework based on an interpretable CNN, specifically designed to automatically extract features of EEG oscillations in the frequency, spatial, and temporal domains in an end-to-end manner.

The frequency components that the network mainly exploited to discriminate the imagined actions were in the alpha band (see Fig. 5). Importantly, these were also the most salient frequency components (see the frequency-level discriminatory power representations in Figs. 6a and 7a). It is important to remark that no bias was introduced

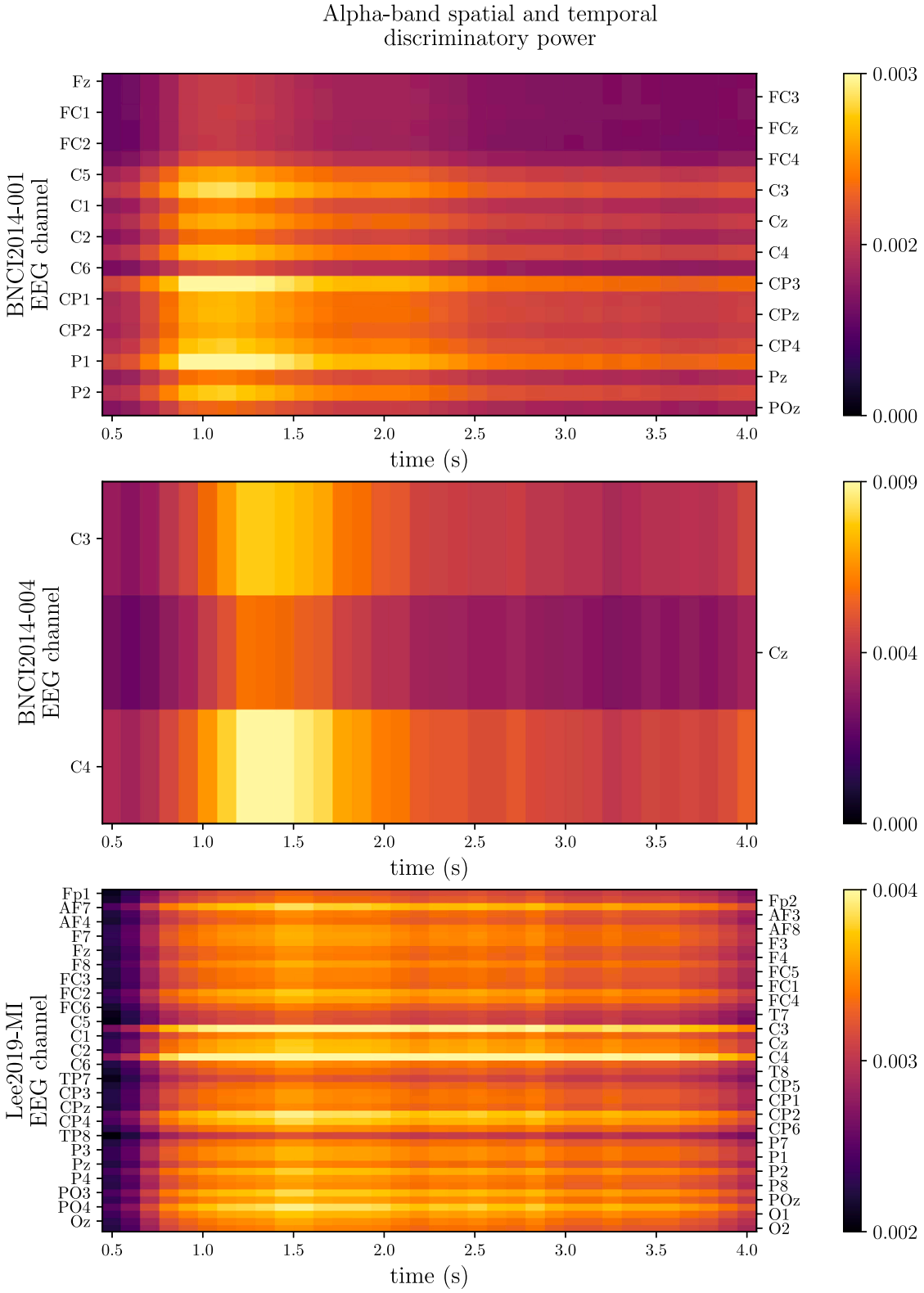

Fig. 8. Alpha-band spatial and temporal discriminatory power. Here, the contribution of each channel and time sample to the discrimination of motor imagery states, linked to alpha-band oscillations, is reported. For each dataset, the mean value across participants ($\in \mathbb{R}^{C \times T'}$, where C are the channels and T' the temporal samples) is reported as a heatmap. See also captions of Figs. 6 and 7 for further details.

Table 3
Comparison with state-of-the-art neural networks for EEG decoding: interpretability, model size and computational time. For each network, the domains in which it learns interpretable EEG features are specified. In addition, the number of trainable parameters (quantifying the model size), the percentage of interpretable trainable parameters (quantifying the interpretability), and the training and inference times (quantifying the computational time during training and testing) are also listed. Training and inference times are expressed as the time to process a mini-batch of examples (32 EEG epochs in this study) on GPU (NVIDIA Titan V), respectively during training and inference. For each quantity, the mean value across the addressed decoding problems is reported with the ranges of values displayed within the brackets.

Neural network	Interpretability			No. parameters	Interpretable param. (%)	Training time (ms/batch)	Inference time (ms/batch)
	frequency	*spatial*	*temporal*				
Sinc-based FINet (ours)	✓	✓	✓	8531 (4802–10372)	97.5 (96.0–98.1)	8.38 (4.60–18.34)	2.04 (1.40–3.28)
MW-based FINet (ours)	✓	✓	✓	8531 (4802–10372)	97.5 (96.0–98.1)	8.33 (4.50–18.26)	1.85 (1.21–3.12)
GG-based FINet (ours)	✓	✓	✓	8563 (4834–10404)	97.5 (96.0–98.2)	9.04 (5.53–19.09)	2.39 (1.79–3.65)
SincEEGNet [37]	✓	✓	✗	1783 (1154–2098)	22.9 (5.5–48.0)	5.92 (4.74–8.86)	1.69 (1.46–2.30)
WaSFCNN [38]	✓	✗	✗	21384 (4802–41677)	0.4 (0.1–1.0)	4.45 (2.48–5.54)	0.87 (0.52–1.62)
MagEEGminer [39]	✓	✗	✗	429 (44–870)	37.4 (33.0–42.8)	4.29 (3.91–5.23)	1.30 (1.03–1.81)
CorrEEGminer [39]	✓	✗	✗	5341 (44–15502)	13.1 (2.4–40.9)	5.95 (5.41–7.00)	1.69 (1.36–2.50)
PLVEEGminer [39]	✓	✗	✗	5183 (32–15136)	4.8 (0.04–18.8)	9.39 (7.79–14.05)	2.69 (1.94–4.33)
EEGNet [20]	✗	✗	✗	2271 (1642–2586)	0 (0–0)	4.28 (2.68–7.54)	0.88 (0.60–1.51)
ShallowConvNet [19]	✗	✗	✗	47963 (8082–102482)	0 (0–0)	4.70 (3.19–6.98)	1.87 (0.69–2.55)
DeepConvNet [19]	✗	✗	✗	281159 (265802–302677)	0 (0–0)	5.66 (5.08–5.87)	1.27 (1.07–1.84)
EEGITNet [53]	✗	✗	✗	4864 (3656–5660)	0 (0–0)	11.32 (9.50–16.39)	2.58 (2.33–3.20)
EEGInception [30]	✗	✗	✗	16073 (14818–17650)	0 (0–0)	12.47 (8.46–21.91)	2.43 (1.95–3.80)
EEGConformer [33]	✗	✗	✗	449779 (410946–505346)	0 (0–0)	24.94 (21.10–26.46)	5.93 (5.60–6.64)

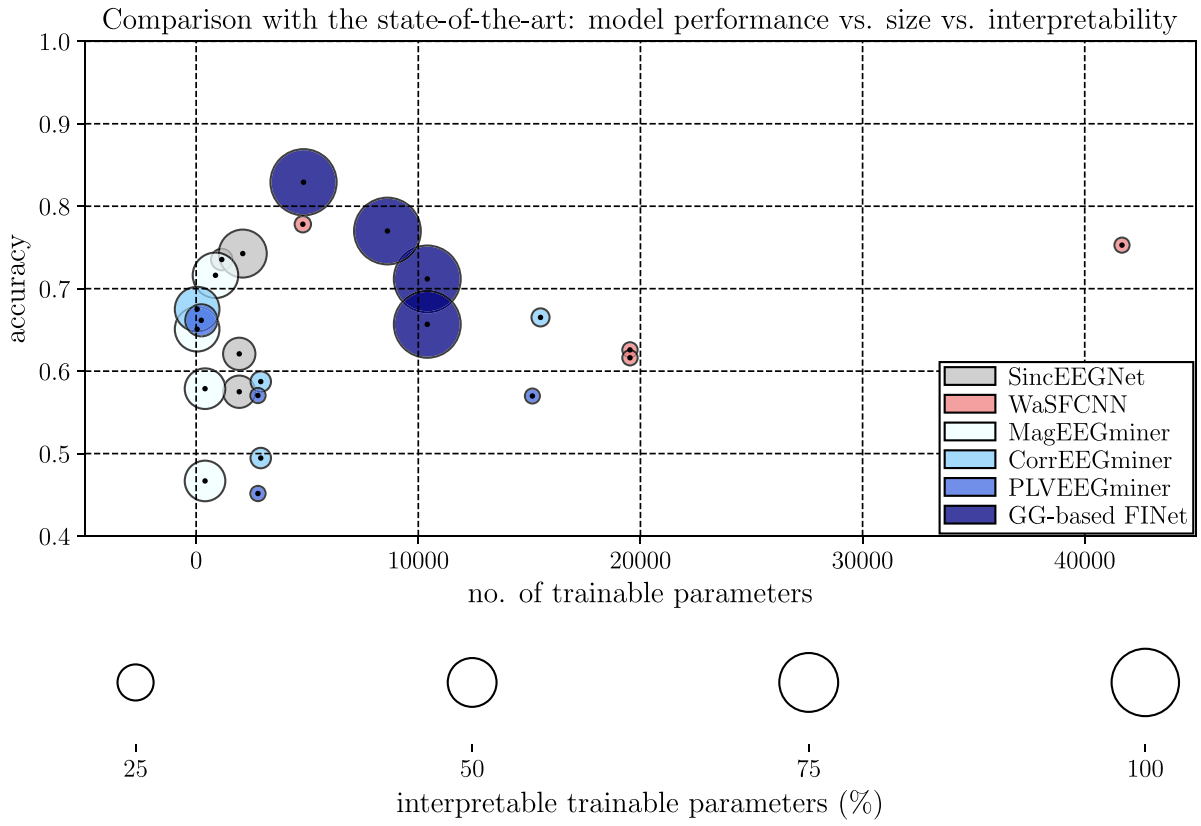

Fig. 9. Comparison with state-of-the-art interpretable neural networks for EEG decoding: model performance vs. size vs. interpretability. Here, the performance is measured with the accuracy, the model size with the number of trainable parameters, and the interpretability with the percentage of interpretable parameters. Each interpretable neural network is displayed in the performance-size plane as a black dot, with a surrounding circle. The circle radius is modulated depending on the amount of interpretability of the model. Here, for brevity and to improve the readability, only our GG-based fully-interpretable CNN is displayed (labeled as 'GG-based FINet' in the figure), as no significant differences were observed between the three variants of our network.

in the network towards the alpha band. Indeed, during training the bandpass filters in the first convolutional layer were not biased in any way towards selecting a specific frequency range in their passband, and rather the central frequency of each filter was randomly initialized over the entire useful frequency spectrum (4 to 40 Hz, as resulting from the pre-processing), see also Section 2.2.2. Therefore, the filters were free to learn a bandwidth centered over any possible frequency between 4 and 40 Hz, based only on the data without any prior

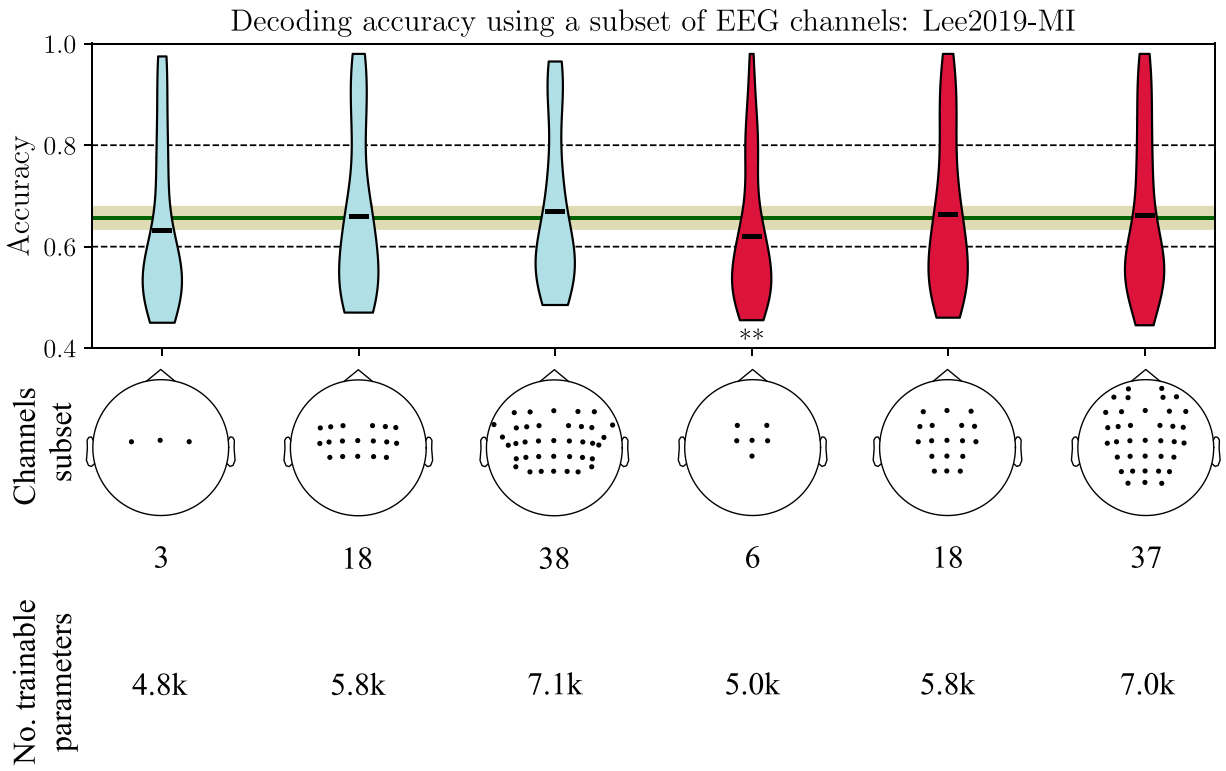

Fig. 10. Channel pruning analysis. The figure displays the configuration of the tested subsets of channels (middle) of the Lee2019-MI dataset (originally including 62 channels), the decoding accuracy obtained while applying our GG-based fully interpretable network to the subsets of channels (top), and the corresponding number of trainable parameters of the network (bottom). Six subsets are tested, ranging from 3 to 38 EEG channels. These include (from left to right in the figure): a configuration with only the 3 electrodes Cz, C3, C4; two configurations with a progressively increasing number of channels adjacent to Cz, C3, C4 (with 18 and 38 channels in total, respectively); three configurations with a progressively increasing number of channels adjacent to Cz (with 6, 18, 37 channels in total, respectively). For each channel subset, the decoding accuracy is displayed as a violin plot representing the accuracy distribution across participants (light-blue: subsets generated using Cz, C3, C4 as seeds; red: subsets generated using Cz as seed). The horizontal black line within each distribution denotes the mean value across participants. The green horizontal line and shaded area denote the mean value and standard error of the mean across participants, respectively, as scored by the network trained on all the available channels (62 channels), representing a reference condition. The accuracy distributions significantly different than the reference condition are marked (*$p < 0.05$, **$p < 0.01$).

constraint or initialization that would favor the alpha band. Hence, the emergence of the alpha-band frequency components as the most informative to predict the motor imagery states was exclusively a result of a data-driven learning process, reflecting genuine neural patterns rather than any architectural or training predisposition. Spatially, central and central/parietal electrode sites – placed approximately above the sensorimotor cortex – resulted the most salient (see the spatial discriminatory power in Fig. 6). These results are in line with the well known time–frequency power modulations observed in the EEG during motor imagery [54,55]. Indeed, the imagination of movements produces an ERD – i.e., a power reduction during the task vs. a baseline period – over the sensorimotor cortex, with a maximum ERD in the alpha band. This alpha-band ERD can be viewed as an electrophysiological signature of a network of cortical neurons ready to process motor information with an increased excitability (i.e., a neural correlate of motor readiness) [58]. Different alpha-band ERD patterns were found depending on the imagined action, involving mainly central and central/parietal channels contralateral to the imagined movement side for right-hand and left-hand motor imagery (overlying the contralateral primary sensorimotor area), midcentral channels for feet motor imagery (overlying the supplementary motor area), and midcentral and bilateral central channels for tongue motor imagery (overlying the supplementary motor area and the right and left primary sensorimotor areas) [54,58–60]. It is worth remarking that also other frequency ranges were found to be involved during motor imagery, such as beta and low-gamma bands [54,58,61–63], and were also found partially involved in our analyses (see fore example beta-band contribution for BNCI2014-004 in Figs. 5, 6a, and 7a); however, our framework suggests that alpha band is the most involved range in the motor imagery tasks analyzed. The weaker contribution of beta and low-gamma bands, which emerged automatically based on the data, might reflect smaller modifications of scalp oscillatory power in these bands compared to alpha band during motor imagery, as also reported in previous studies [55,64,65]. It should also be noted that the frequency- and spatial-domain signatures of motor imagery found with the proposed framework were consistent across datasets in our experiments, with small differences due to the different EEG channel setups (22 vs. 3 vs. 62 channels across the three datasets) and the different contrasted motor imagery conditions (e.g., BNCI2014-001 contains also feet and tongue imagery conditions in addition to left-hand and right-hand imagery).

As concerning the characterization in the time domain, the temporal discriminatory power – quantifying how much the different time samples are affected by motor imagery (see Fig. 7) – ramped up early after the motor imagery onset (0.5–1.0 s) and then, depending on the dataset processed, a different temporal dynamics was observed. In the first two datasets (BNCI2014-001 and BNCI2014-004) the discriminatory power peaked within the first half of the trial (1.0–1.5 s) and then slowly decreased over time; on the other hand, in the last dataset (Lee2019-MI) it maintained high levels for almost the entire trial duration. Therefore, in the last dataset, almost all time samples resulted strongly affected by motor imagery. These different temporal dynamics could be due to differences in the recording paradigm across datasets. All paradigms were based on the continuous imagination of actions during a relatively

long interval (from 4s-length to 4.5s-length motor imagery interval). However, the cue (e.g., a pointing arrow, encoding a specific motor imagery condition) lasted for a short interval in the BNCI2014-001 and BNCI2014-004 datasets, and during the motor imagery interval only a fixation cross was showed on the display; therefore, in this case the modulation could be more prone to vanish over time, as we move far from the cue. Conversely, in the Lee2019-MI dataset the cue was displayed also during the motor imagery interval, potentially strengthening the modulation during the entire trial duration. To further inspect these time-domain results, in Appendix A we conducted an ERD analysis on the same datasets processed with our framework; interestingly, besides the well known significant alpha-band ERD approximately over the sensorimotor cortex, the temporal dynamics in the ERD was similar to that found with our CNN-based framework, proving that the observed phenomenon is a genuine characteristics of the data and is not artificially introduced by our approach.

Overall, the results obtained with our framework match the known neurophysiology of motor imagery and the modulations observed in the literature with classic analyses based on time–frequency power modulations (ERD analysis). The quality of the multi-domain EEG features learned by our network – on top of which our analysis framework is built – was also ensured by the significantly higher decoding performance compared to the chance level. To further strengthen our results, it is important to remark that our framework was designed using three different variants of the proposed fully-interpretable architecture (i.e., sinc-based, MW-based, and GG-based), each exploiting a different way for designing the frequency response of the trainable bank of bandpass filters. Despite the intrinsic differences among the filter formulations, no significant difference was observed in the decoding accuracy across the three CNNs (based on the different filter types) in any of the four motor imagery decoding problems. This indicates that the network's ability to extract discriminative features for motor imagery classification is robust to the specific filter formulation, as long as the filter bank spans the relevant frequency bands. Furthermore, across the different variants of our framework (see Section 3.1 and Appendix B), similar frequency-, spatial-, and temporal- domain signatures of motor imagery were found, thus confirming our findings in a more sound way. As no statistical differences were found in the decoding performance among the variants of our network, and as they all provide the same level of interpretability (97.5%) and almost the same model size (from 8531 to 8563 trainable parameters), we suggest practitioners to use our framework based on the GG-based fully-interpretable network, as the functional form used for modeling the frequency response of bandpass filters exploits functions ranging from Gaussian (MW-based) to rectangular (sinc-based, ideal bandpass filter), thus being a generalization of the other variants (see Section 2.1.1).

4.2. Motor imagery decoding and comparison with state-of-the-art networks

We performed a comprehensive comparison of our fully interpretable networks against several interpretable and non-interpretable state-of-the-art networks. In terms of decoding accuracy, our interpretable architectures were either statistically superior or comparable to non-interpretable designs. This is a particularly remarkable result given that our networks were explicitly designed for achieving a high level of interpretability rather than to maximize performance, yet they still achieved overall superior results compared to state-of-the-art networks not constrained to be interpretable at all. Furthermore, in most of cases our networks resulted superior in decoding performance to the state-of-the art interpretable networks, too. The latter, although interpretable, limit their interpretability mainly to one domain (frequency) or at most to two domains (frequency and space), at variance with ours that are interpretable in three domains (frequency, space, and time). Thus overall, the proposed networks represent promising candidates as novel designs that combine maximization of interpretability with high performance. Compared to the other interpretable state-of-the-art networks, the extension of interpretability to all domains came at cost of a general increase in the number of trainable parameters (on average across the four motor imagery decoding problems) together with an increase in the computational times. As concerning the number of trainable parameters, useful cues for their reduction come from the pruning channel experiments. These show (see Fig. 10) that the trainable parameters can be dramatically reduced by considering a small sets of electrodes over the cortical regions more relevant for the addressed decoding problem (here the motor cortex) still preserving the same high level of decoding accuracy. This proved the feasibility to exploit the proposed designs to realize more efficient decoders in terms of number of parameters to fit (more convenient for devices with limited capabilities), exploiting more parsimonious EEG layouts (more convenient in laboratory routine), without worsening significantly the decoding accuracy and keeping interpretability in all domains. As concerning the computational times (training and inference time), the proposed networks settled approximately at halfway among all the other tested networks (both interpretable and non-interpretable). While computational times depend on several factors – e.g., depth of the network, type of convolutions adopted (for example, depthwise or not) – a factor that likely contributes to increase the computational times in our designs is related to the particular implementation adopted for the filters in the first convolutional layer. In this implementation, indeed, the filters are not described directly by the coefficients of the kernel; rather, they are defined by their frequency response and then the coefficients of the kernel are obtained by computing the inverse Fourier transform (obtaining the impulse response of the filter), windowed by a Hamming window. The additional operations required to compute the values of the kernel from the frequency response likely contribute to increase the computational times.

4.3. Impact

The main novelty point of our study regards the definition of a CNN-empowered analysis tool to investigate EEG oscillations. Classic analyses – from the pre-processing to processing – are generally conducted in a semi-automatic way and without optimally adapting the entire processing to the investigated cognitive task. Conversely, our proposed analysis framework is fully *automatic* and *optimized* on the cognitive task under analysis. Indeed, it is designed on top of an interpretable neural network that automatically learns the optimal EEG features for predicting the brain states that characterize the task, separately for each participant. Thus, since the features are not handcrafted but automatically learned from the data, the proposed analysis framework is application-agnostic and it can be easily adapted to any other cognitive tasks by simply re-training it on new data. Moreover, classic analyses process the EEG in the same way across participants, for example by computing in the same way the time–frequency power modulations (e.g., using the same mother wavelet, and/or focusing on the same time-intervals or frequency ranges). On the other hand, our approach automatically adapts itself to each participant's EEG signals, by revealing the most salient participant-specific neural signatures. This makes our approach particularly suitable for analyzing EEG signals in patients. Indeed, when examining pathological conditions, a patient-specific approach can be especially beneficial, due to the high inter-subject variability as the neural activity can be differently modulated depending on several factors, such as the size and location of the damaged brain area, the time from the onset of the lesion, and the severity of the neuropathology. Finally, our approach is applied in an *end-to-end* fashion to minimally pre-processed EEG, enabling a more direct link between the neural activity and the brain states under analysis with less factors influencing the quality of the results, otherwise influenced by specific EEG pre-processing and pro-

cessing choices in classic analyses. This aspect could have key role for potentially highlighting, in a data-driven way, new neural signatures related to EEG oscillations with minimal a priori assumptions, in both healthy participants and patients. Overall, the prospective implications of our framework for analyzing patients' signals could be useful, for example, for defining novel EEG-based biomarkers to quantify and track the neuropathology [66], for better targeting the neuronal targets for treatments involving non-invasive brain stimulation [67], and for developing diagnostic and pharmacotherapeutic strategies [2].

This study, besides proposing a CNN-empowered tool for EEG analysis – representing the primary contribution – it also offers a secondary contribution, which arises as a corollary of the primary one. Indeed, the fully-interpretable CNN design (integral to the definition of the analysis framework) could be exploited, in prospective, for realizing more interpretable deep learning-based neural decoders (e.g., for brain–computer interfacing), with 97.5% of the trainable parameters being interpretable, across the entire network. Indeed, even tough it is not the main aim of our study (as it is focused on leveraging network features for EEG analysis and not for EEG decoding), our networks turned out to significantly outperform existing interpretable solutions, with accuracy improvements spanning from 0.6% to 26.3%. Importantly, this was accompanied with an increased interpretability compared to existing approaches, extending the interpretation to all frequency, spatial, and temporal domains, and increasing from 1.6 to 243 times the percentage of the interpretable parameters learned by the network. This is a relevant aspect, as interpretable models are underexplored in the literature, which often considers non-interpretable neural networks and uses them as 'black boxes', generally focusing only on maximizing the network performance and ignoring the validation of the network features [16, 17]. Moreover, the studies that exploit interpretable models [37–39], mostly visualize and validate network features in the frequency domain or at most in the frequency and spatial domains. By adopting our neural network architecture or by proposing new networks based on our key architectural elements (see Section 2.1.1), practitioners could easily extract the network features in frequency, spatial, and temporal domains and compare them with the known neurophysiology/neuropathology characterizing the addressed EEG decoding problem, enriching the validation of deep learning-based neural decoders.

4.4. Limitations and future directions

This study is aimed at illustrating for the first time our novel CNN-based EEG analysis framework and at providing a first validation on real data recorded in a single cognitive task (motor imagery). The obtained results are extremely promising and motivate to further enrich the validation of our framework by extending its application to other cognitive tasks. Indeed, here we only used a single cognitive task (motor imagery) as reference, as it is characterized by a peculiar and well-characterized neurophysiology. The application of our framework to other cognitive tasks would enhance the robustness of our approach. Moreover, as the main strength of our framework relies on the participant-specific EEG processing, our approach will also be applied to patients' signals to enrich the characterization of neuropathological alterations and potentially define novel biomarkers to track the pathology. Finally, the higher training and inference times of the proposed interpretable networks compared to most of state-of-the-art networks limits the application of our approach in real-time implementations. Future studies aiming at using our approach in real-time should adopt solutions to improve the computational times of the interpretable layers. For example, after training, once learned the optimal frequency response parameters in the first interpretable layer – i.e., once determined the coefficients of the temporal convolutional kernels – to reduce inference times the kernel coefficients could be stored and directly used during inference using a traditional temporal convolution, instead of deriving them again from the frequency response.

5. Conclusions

In conclusion, here we propose a data-driven framework based on a fully-interpretable CNN, able to characterize EEG oscillations in an *automatic*, *optimal*, and *end-to-end* way. Our framework unveiled the frequency-, spatial-, and temporal-domain EEG signatures that characterize most the cognitive task under analysis, by taking advantage of the interpretable features learned by a fully-interpretable CNN. By exploiting an end-to-end and fully-interpretable feature learning, our approach can be conveniently applied to analyze minimally pre-processed EEG signals, automatically unveiling the signatures related to a cognitive task. Compared with traditional analyses, our approach could provide valuable insights about the most salient neural correlates especially in cognitive tasks that are under-looked, in which the signatures are still scarcely characterized. The results obtained on real data recorded during motor imagery tasks suggest that our CNN-empowered analysis approach is capable of revealing neurophysiologically plausible characteristics of EEG oscillations in the domains under investigation, by matching both the modulations known in the literature occurring during motor imagery and the results that can be obtained with classic analysis approaches (based on event-related spectral perturbations) on the same data. Therefore, neuroscientists could conveniently exploit our analysis pipeline in place of or in addition to classic analyses, for boosting the description and the characterization of brain oscillations during cognitive tasks in healthy participants and prospectively in patients, tracking their neuropathological alterations.

CRediT authorship contribution statement

Davide Borra: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Elisa Magosso:** Writing – review & editing, Validation, Supervision, Resources, Methodology.

Ethics approval

The authors declare that this research does not require an Ethics statement as it is based solely on publicly available data and does not involve direct interaction with human subjects.

Funding

This research was co-funded by the Italian Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, "DARE - DigitAl lifelong pRevEntion" initiative, code PNC0000002, CUP: B53C22006450001.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

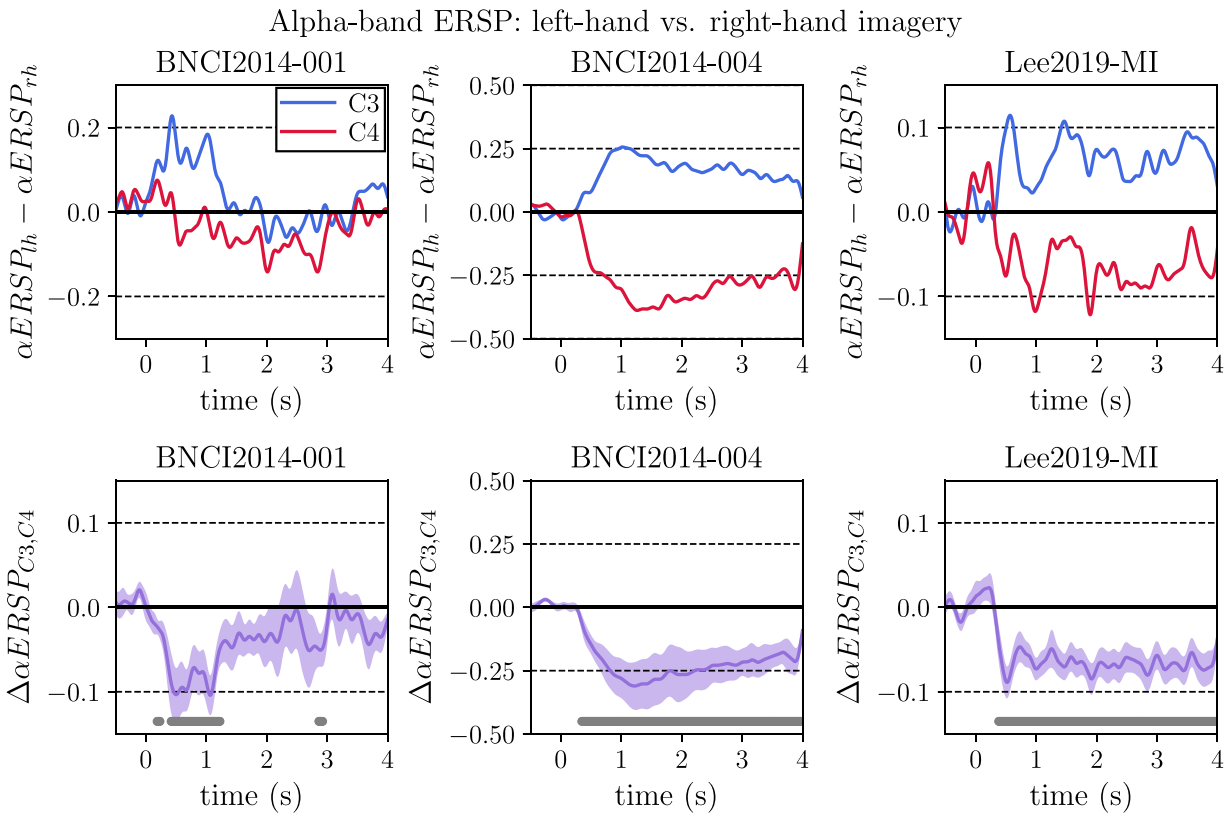

Fig. A.1. Alpha-band event-related spectral perturbations (ERSPs) at central electrode sites: left-hand vs. right-hand motor imagery. Top panels report the ERSP difference between the left-hand and right-hand motor imagery at C3 and C4 electrode sites, separately for each dataset. Here, the ERSP is reported as mean value across participants. Bottom panels report the modulation of the alpha-band ERSP at central electrode sites ($\Delta\alpha ERSP_{C3,C4}$), separately for each dataset. Here, the line denotes the mean value, while the shaded area the standard error of the mean across participants. The results of the performed statistical analysis are reported too, by marking with gray dots the time samples with a significant modulation of the alpha-band central ERSP.

Appendix A. Event-related spectral perturbations

For each participant, each trial, and each EEG channel, the same signals used as input of our interpretable CNN were processed as follows. The event-related spectral perturbations (ERSPs) were computed based on the continuous wavelet transform of the signal utilizing the complex Morlet wavelet as basis function. Wavelet transform coefficients were squared to obtain time–frequency power representations. Then, for each participant and each EEG channel, the power representations were averaged across trials, separately for each brain state (i.e., each motor imagery condition in this study), and were normalized using the power values of the rest interval from −1 s to 0 s as baseline (where 0 s represents the motor imagery onset). In particular, the ERSP was computed as the difference between the power at each time–frequency sample and the mean power across the time samples of the baseline interval at the same frequency, and divided by the same average baseline power. Therefore, the so computed ERSP quantifies the relative increase/decrease of the power during the motor imagery task with respect to the baseline period (here assumed as the rest period). Crucially, this quantifies the event-related desynchronization/synchronization (ERD/S, ERS: $ERSP > 0$, ERD: $ERSP < 0$).

ERSPs were considered for the left-hand and right-hand motor imagery conditions, being the common brain states across datasets. Then, we extracted the ERSP within the alpha band (8–13 Hz) at C3 and C4 channels, known to be modulated depending on the left-hand and right-hand motor imagery conditions, as being approximately over the left and right primary sensorimotor areas. As a stronger ERD (i.e., a negative ERSP) at EEG channels contralateral to the imagined action is expected, the *modulation of the alpha-band central ERSP* ($\Delta\alpha ERSP_{C3,C4}$) was quantified according to:

$$\begin{aligned}&\Delta\alpha ERSP_{C3,C4}\\&= \left[\left(\alpha ERSP_{lh,C4} - \alpha ERSP_{rh,C4}\right) + \left(\alpha ERSP_{rh,C3} - \alpha ERSP_{lh,C3}\right)\right]/2 =\\&= \left[\left(\alpha ERSP_{lh,C4} - \alpha ERSP_{rh,C4}\right) - \left(\alpha ERSP_{lh,C3} - \alpha ERSP_{rh,C3}\right)\right]/2.\end{aligned} \tag{A.1}$$

That is, for each central channel (C4 and C3) we computed separately the difference of the alpha-band ERSP between the imagery condition contralateral vs. ipsilateral to each channel, i.e between the left-/right-hand motor imagery for C4 ($\alpha ERSP_{lh,C4} - \alpha ERSP_{rh,C4}$) and the right-/left-hand motor imagery for C3 ($\alpha ERSP_{rh,C3} - \alpha ERSP_{lh,C3}$). Then these quantities were averaged to obtain $\Delta\alpha ERSP_{C3,C4}$. The more this quantity is negative the stronger the modulation of the alpha-band ERSP at the central electrodes C3 and C4. A permutation t test [68] was performed to assess if a statistically significant modulation between the contrasted motor imagery conditions occurs, by comparing $\Delta\alpha ERSP_{C3,C4}$ vs. the null distribution.

Fig. A.1 reports the alpha-band ERSP at central electrode sites and the modulation of the central alpha-band ERSP ($\Delta\alpha ERSP_{C3,C4}$), separately for each dataset. A significant modulation was observed (see bottom panels of Fig. A.1) in all datasets, with significant higher modulations in

early time samples for BNCI2014-001 and BNCI2014-004 datasets, and with significant modulations almost constant over time for the Lee2019-MI dataset.

Appendix B. Proposed EEG analysis framework: results from sinc-based and MW-based fully-interpretable neural networks

The results obtained with the proposed EEG analysis framework utilizing a sinc-based or Morlet wavelet-based fully-interpretable CNNs are reported in this section. Frequency-level probability representations are reported in Figs. B.1 and B.2, and frequency-level spatial and temporal discriminatory power representations are reported in Figs. B.3 and B.4.

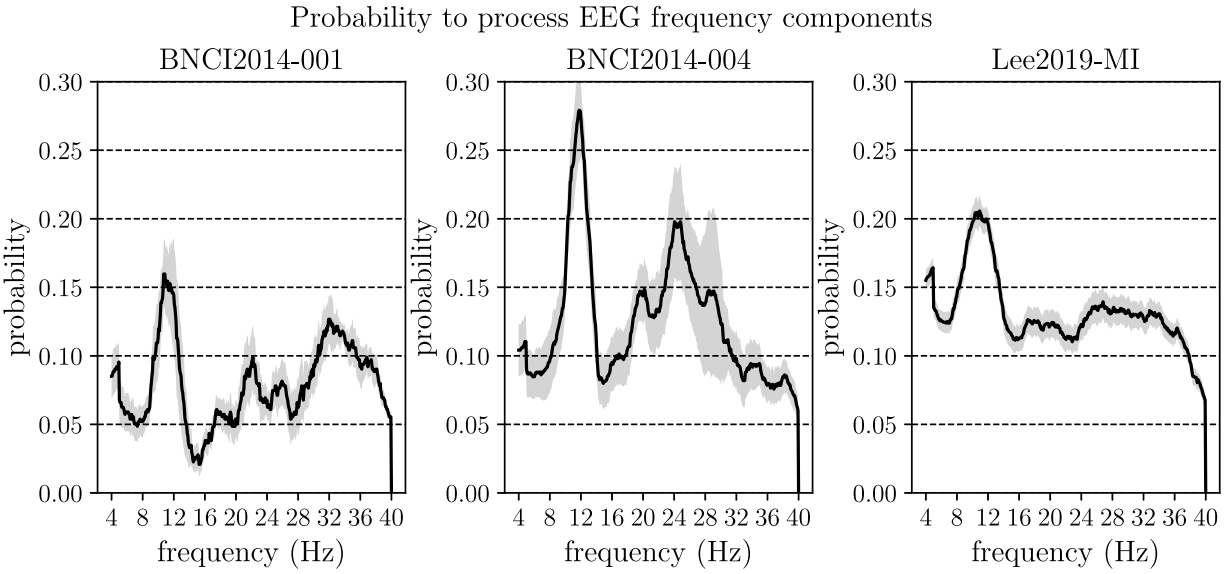

Fig. B.1. Frequency-level probability obtained with the proposed analysis framework utilizing a sinc-based fully-interpretable network. See caption of Fig. 5 for further details.

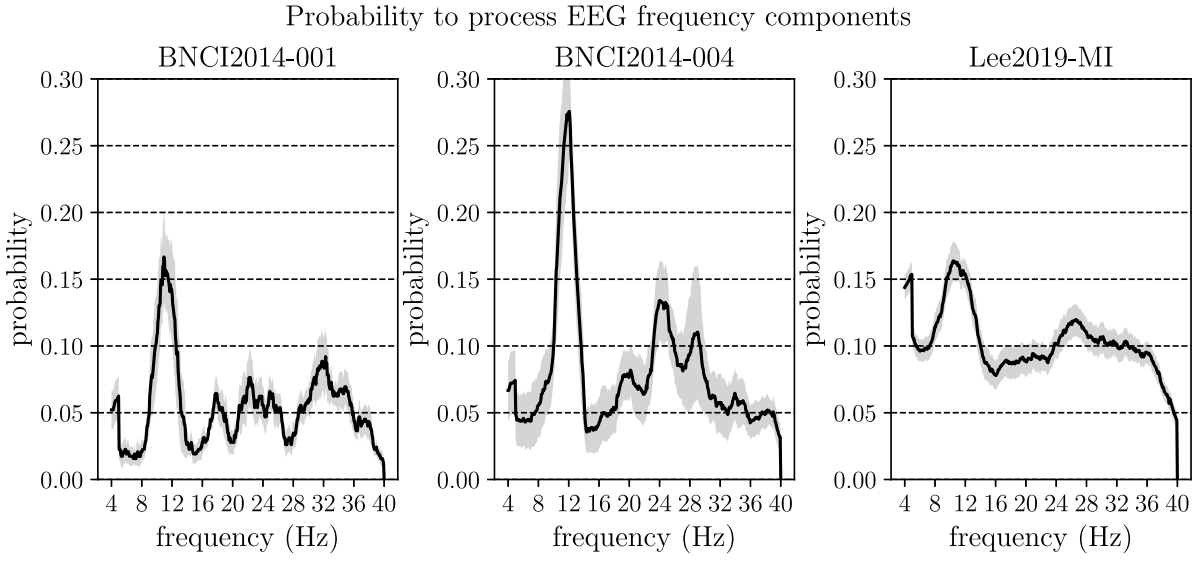

Fig. B.2. Frequency-level probability obtained with the proposed analysis framework utilizing a Morlet wavelet-based fully-interpretable network. See caption of Fig. 5 for further details.

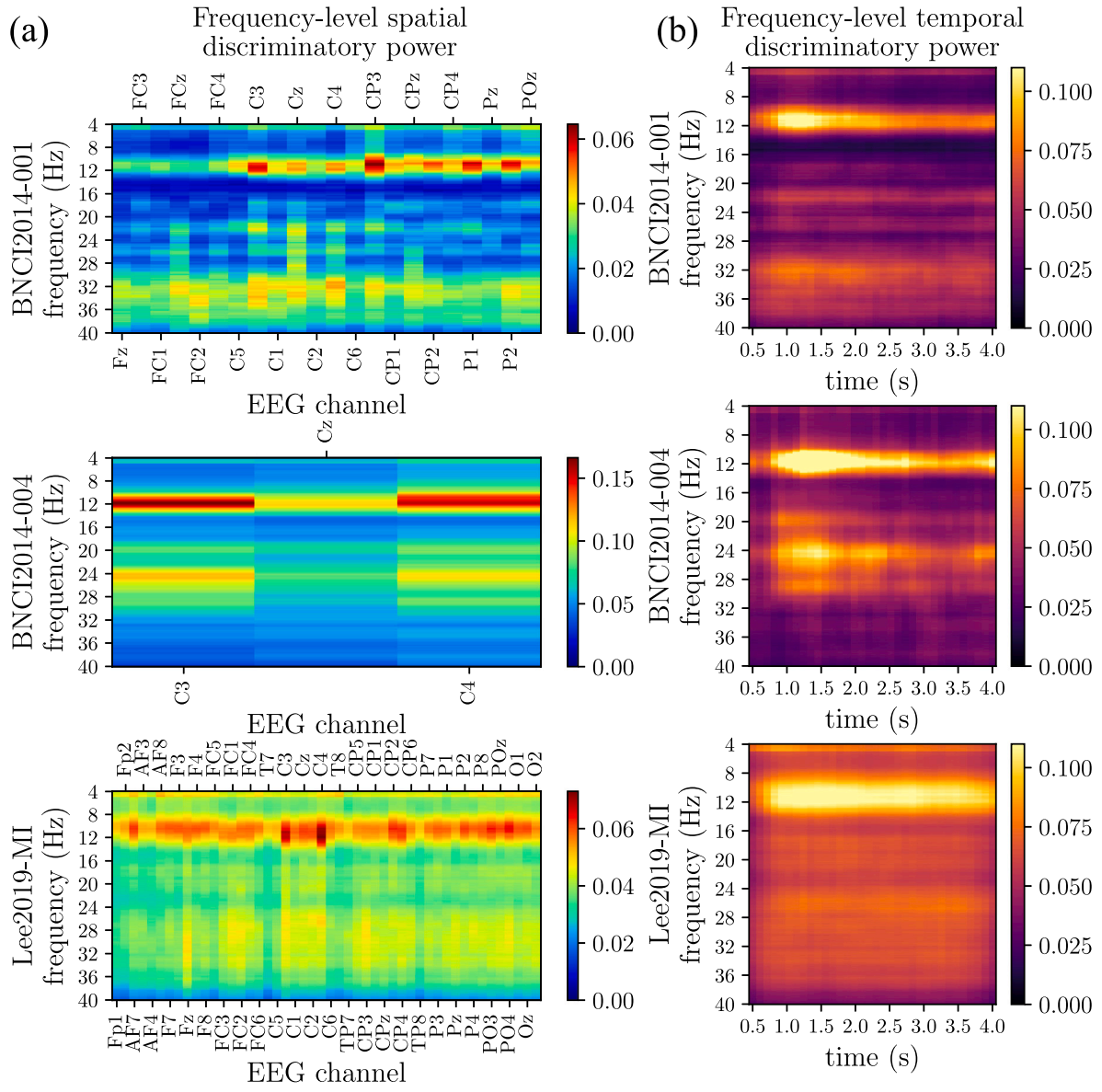

Fig. B.3. Frequency-level spatial (panel a) and temporal (panel b) discriminatory power representations obtained with the proposed analysis framework utilizing a sinc-based fully-interpretable network. See captions of Figs. 6a and 7a for further details.

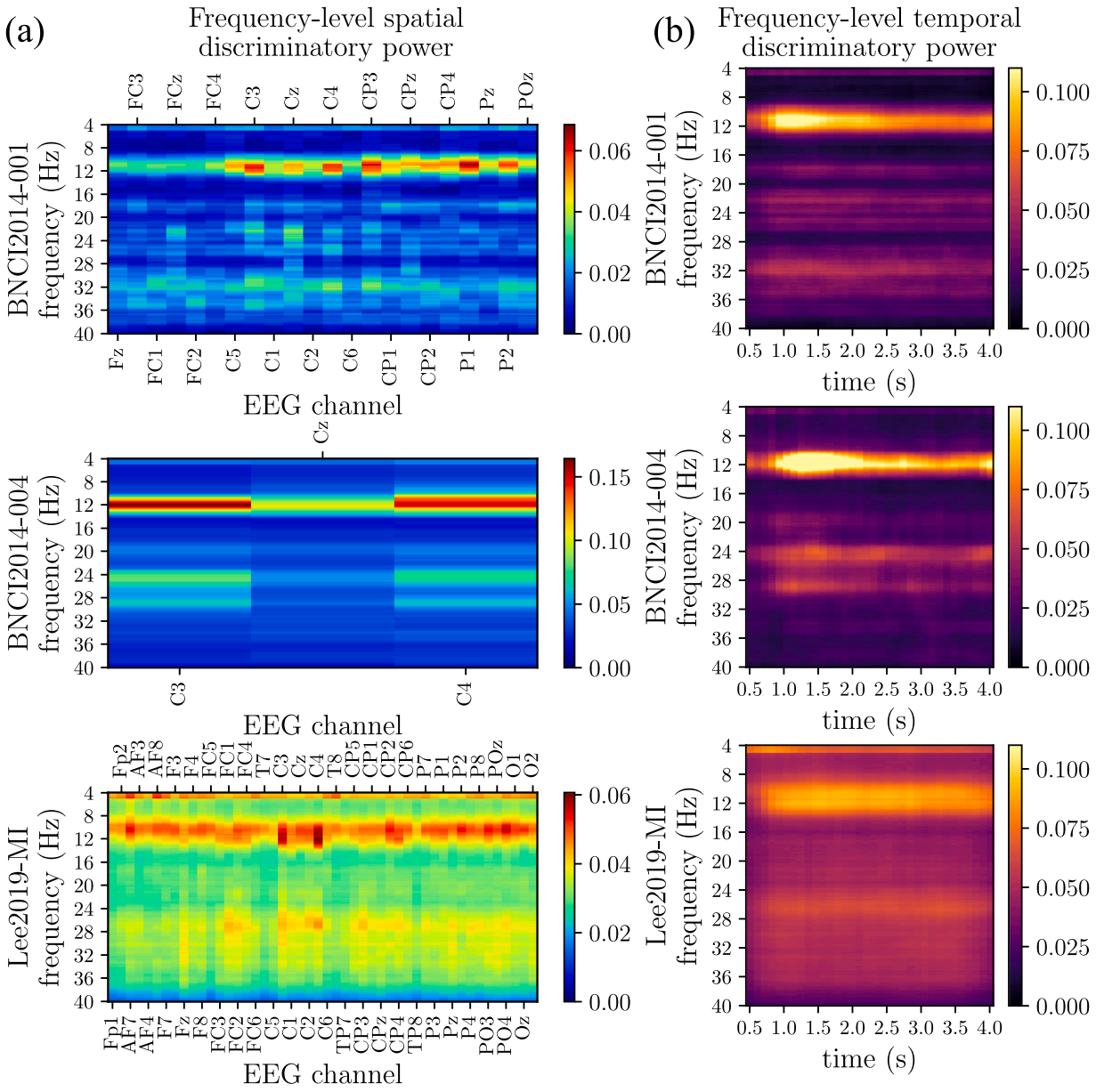

Fig. B.4. Frequency-level spatial (panel a) and temporal (panel b) discriminatory power representations obtained with the proposed analysis framework utilizing a Morlet wavelet-based fully-interpretable network. See captions of Figs. 6a and 7a for further details.

Data availability

The relevant codes for defining the interpretable neural networks presented in this study are publicly available in a Zenodo repository [69] at https://doi.org/10.5281/zenodo.16156955. The public datasets adopted in this study are available at https://www.bbci.de/competition/iv/ and at https://gigadb.org/dataset/100542.

References

[1] E. Niedermeyer, Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Lippincott Williams & Wilkins, 2011.

[2] E. Başar, Brain oscillations in neuropsychiatric disease, Dialogues Clin. Neurosci. 15 (3) (2013) 291–300, http://dx.doi.org/10.31887/dcns.2013.15.3/ebasar.

[3] R. Lizio, F. Vecchio, G.B. Frisoni, R. Ferri, G. Rodriguez, C. Babiloni, Electroencephalographic rhythms in alzheimer's disease, in: F. Ferrarelli (Ed.), Int. J. Alzheimer' s Dis. 2011 (1) (2011) http://dx.doi.org/10.4061/2011/927573.

[4] V.J. Geraedts, L.I. Boon, J. Marinus, A.A. Gouw, J.J. van Hilten, C.J. Stam, M.R. Tannemaat, M.F. Contarino, Clinical correlates of quantitative EEG in parkinson disease: A systematic review, Neurology 91 (19) (2018) 871–883, http://dx.doi.org/10.1212/wnl.0000000000006473.

[5] S.-P. Kim, Preprocessing of EEG, in: Computational EEG Analysis, Springer Singapore, 2018, pp. 15–33, http://dx.doi.org/10.1007/978-981-13-0908-3_2.

[6] J.A. Urigüen, B. Garcia-Zapirain, EEG artifact removal—state-of-the-art and guidelines, J. Neural Eng. 12 (3) (2015) 031001, http://dx.doi.org/10.1088/1741-2560/12/3/031001.

[7] K.A. Robbins, J. Touryan, T. Mullen, C. Kothe, N. Bigdely-Shamlo, How sensitive are EEG results to preprocessing methods: A benchmarking study, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (5) (2020) 1081–1090, http://dx.doi.org/10.1109/tnsre.2020.2980223.

[8] S. Coelli, A. Calcagno, C.M. Cassani, F. Temporiti, P. Reali, R. Gatti, M. Galli, A.M. Bianchi, Selecting methods for a modular EEG pre-processing pipeline: An objective comparison, Biomed. Signal Process. Control. 90 (2024) 105830, http://dx.doi.org/10.1016/j.bspc.2023.105830.

[9] S. Morales, M.E. Bowers, Time-frequency analysis methods and their application in developmental EEG data, Dev. Cogn. Neurosci. 54 (2022) 101067, http://dx.doi.org/10.1016/j.dcn.2022.101067.

[10] R. Grandchamp, A. Delorme, Single-trial normalization for event-related spectral decomposition reduces sensitivity to noisy trials, Front. Psychol. 2 (2011) http://dx.doi.org/10.3389/fpsyg.2011.00236.

[11] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, K.A. Robbins, The PREP pipeline: standardized preprocessing for large-scale eeg analysis, Front. Neuroinformatics 9 (2015) http://dx.doi.org/10.3389/fninf.2015.00016.

[12] A. Pedroni, A. Bahreini, N. Langer, Automagic: Standardized preprocessing of big EEG data, NeuroImage 200 (2019) 460–473, http://dx.doi.org/10.1016/j.neuroimage.2019.06.046.

[13] A. Keil, S. Debener, G. Gratton, M. Junghöfer, E.S. Kappenman, S.J. Luck, P. Luu, G.A. Miller, C.M. Yee, Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography, Psychophysiology 51 (1) (2013) 1–21, http://dx.doi.org/10.1111/psyp.12147.

[14] G. Niso, L.R. Krol, E. Combrisson, A.S. Dubarry, M.A. Elliott, C. François, Y. Héjja-Brichard, S.K. Herbst, K. Jerbi, V. Kovic, K. Lehongre, S.J. Luck, M. Mercier, J.C. Mosher, Y.G. Pavlov, A. Puce, A. Schettino, D. Schön, W. Sinnott-Armstrong, B. Somon, A.e. Šoškić, S.J. Styles, R. Tibon, M.G. Vilas, M. van Vliet, M. Chaumon, Good scientific practice in EEG and MEG research: Progress and perspectives, NeuroImage 257 (2022) 119056, http://dx.doi.org/10.1016/j.neuroimage.2022.119056.

[15] G.W. Lindsay, Convolutional neural networks as a model of the visual system: Past, present, and future, J. Cogn. Neurosci. 33 (10) (2021) 2017–2031, http://dx.doi.org/10.1162/jocn_a_01544.

[16] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, J. Neural Eng. 16 (5) (2019) 051001, http://dx.doi.org/10.1088/1741-2552/ab260c.

[17] A. Craik, Y. He, J.L. Contreras-Vidal, Deep learning for electroencephalogram (EEG) classification tasks: a review, J. Neural Eng. 16 (3) (2019) 031001, http://dx.doi.org/10.1088/1741-2552/ab0ab5.

[18] D. Borra, E. Magosso, M. Ravanelli, A protocol for trustworthy EEG decoding with neural networks, Neural Netw. 182 (2025) 106847, http://dx.doi.org/10.1016/j.neunet.2024.106847.

[19] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420, http://dx.doi.org/10.1002/hbm.23730.

[20] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013, http://dx.doi.org/10.1088/1741-2552/aace8c.

[21] M. Simões, D. Borra, E. Santamaría-Vázquez, M. Bittencourt-Villalpando, D. Krzemiński, A. Miladinović, T. Schmid, H. Zhao, C. Amaral, B. Direito, J. Henriques, P. Carvalho, M. Castelo-Branco, BCIAUT-P300: A multi-session and multi-subject benchmark dataset on autism for P300-based brain-computer-interfaces, Front. Neurosci. 14 (2020) http://dx.doi.org/10.3389/fnins.2020.568104.

[22] D. Borra, S. Fantozzi, E. Magosso, A lightweight multi-scale convolutional neural network for P300 decoding: Analysis of training strategies and uncovering of network decision, Front. Hum. Neurosci. 15 (2021) http://dx.doi.org/10.3389/fnhum.2021.655840.

[23] D. Xu, F. Tang, Y. Li, Q. Zhang, X. Feng, An analysis of deep learning models in SSVEP-based BCI: A survey, Brain Sci. 13 (3) (2023) 483, http://dx.doi.org/10.3390/brainsci13030483.

[24] N. Waytowich, V.J. Lawhern, J.O. Garcia, J. Cummings, J. Faller, P. Sajda, J.M. Vettel, Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials, J. Neural Eng. 15 (6) (2018) 066031, http://dx.doi.org/10.1088/1741-2552/aae5d8.

[25] Z. Ji, T. Xu, C. Chen, H. Yin, F. Wan, H. Wang, Subject-specific CNN model with parameter-based transfer learning for SSVEP detection, Biomed. Signal Process. Control. 103 (2025) 107404, http://dx.doi.org/10.1016/j.bspc.2024.107404.

[26] H. Cecotti, A. Graser, Convolutional neural networks for P300 detection with application to brain-computer interfaces, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 433–445, http://dx.doi.org/10.1109/tpami.2010.125.

[27] J. An, X. Chen, D. Wu, Algorithm contest of motor imagery BCI in the world robot contest 2022: A survey, Brain Sci. Adv. 9 (3) (2023) 166–181, http://dx.doi.org/10.26599/bsa.2023.9050011.

[28] D. Borra, V. Mondini, E. Magosso, G.R. Müller-Putz, Decoding movement kinematics from EEG using an interpretable convolutional neural network, Comput. Biol. Med. 165 (2023) 107323, http://dx.doi.org/10.1016/j.compbiomed.2023.107323.

[29] A. Vahid, M. Mückschel, S. Stober, A.-K. Stock, C. Beste, Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control, Commun. Biol. 3 (1) (2020) http://dx.doi.org/10.1038/s42003-020-0846-z.

[30] E. Santamaría-Vázquez, V. Martínez-Cagigal, F. Vaquerizo-Villar, R. Hornero, EEG-inception: A novel deep convolutional neural network for assistive ERP-based brain-computer interfaces, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (12) (2020) 2773–2782, http://dx.doi.org/10.1109/tnsre.2020.3048106.

[31] H. Wang, Z. Pei, L. Xu, T. Xu, A. Bezerianos, Y. Sun, J. Li, Performance enhancement of P300 detection by multiscale-CNN, IEEE Trans. Instrum. Meas. 70 (2021) 1–12, http://dx.doi.org/10.1109/tim.2021.3067943.

[32] H. Wang, L. Xu, A. Bezerianos, C. Chen, Z. Zhang, Linking attention-based multiscale CNN with dynamical GCN for driving fatigue detection, IEEE Trans. Instrum. Meas. 70 (2021) 1–11, http://dx.doi.org/10.1109/tim.2020.3047502.

[33] Y. Song, Q. Zheng, B. Liu, X. Gao, EEG conformer: Convolutional transformer for EEG decoding and visualization, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 710–719, http://dx.doi.org/10.1109/tnsre.2022.3230250.

[34] F.V. Farahani, K. Fiok, B. Lahijanian, W. Karwowski, P.K. Douglas, Explainable AI: A review of applications to neuroimaging data, Front. Neurosci. 16 (2022) http://dx.doi.org/10.3389/fnins.2022.906290.

[35] P. Rajpura, H. Cecotti, Y. Kumar Meena, Explainable artificial intelligence approaches for brain–computer interfaces: a review and design space, J. Neural Eng. 21 (4) (2024) 041003, http://dx.doi.org/10.1088/1741-2552/ad6593.

[36] A. Sujatha Ravindran, J. Contreras-Vidal, An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth, Sci. Rep. 13 (1) (2023) http://dx.doi.org/10.1038/s41598-023-43871-8.

[37] D. Borra, E. Magosso, M. Castelo-Branco, M. Simões, A Bayesian-optimized design for an interpretable convolutional neural network to decode and analyze the P300 response in autism, J. Neural Eng. 19 (4) (2022) 046010, http://dx.doi.org/10.1088/1741-2552/ac7908.

[38] D. Zhao, F. Tang, B. Si, X. Feng, Learning joint space–time–frequency features for EEG decoding on small labeled data, Neural Netw. 114 (2019) 67–77, http://dx.doi.org/10.1016/j.neunet.2019.02.009.

[39] S. Ludwig, S. Bakas, D.A. Adamos, N. Laskaris, Y. Panagakis, S. Zafeiriou, EEGminer: discovering interpretable features of brain activity with learnable filters, J. Neural Eng. 21 (3) (2024) 036010, http://dx.doi.org/10.1088/1741-2552/ad44d7.

[40] D. Borra, S. Fantozzi, E. Magosso, Interpretable and lightweight convolutional neural network for EEG decoding: Application to movement execution and imagination, Neural Netw. 129 (2020) 55–74, http://dx.doi.org/10.1016/j.neunet.2020.05.032.

[41] M.X. Cohen, A better way to define and describe morlet wavelets for time-frequency analysis, NeuroImage 199 (2019) 81–86, http://dx.doi.org/10.1016/j.neuroimage.2019.05.048.

[42] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015, http://dx.doi.org/10.48550/ARXIV.1502.03167.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (56) (2014) 1929–1958.

[44] A. Farahat, C. Reichert, C.M. Sweeney-Reed, H. Hinrichs, Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization, J. Neural Eng. 16 (6) (2019) 066010, http://dx.doi.org/10.1088/1741-2552/ab3bb4.

[45] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K.J. Miller, G.R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, B. Blankertz, Review of the BCI competition IV, Front. Neurosci. 6 (2012) http://dx.doi.org/10.3389/fnins.2012.00055.

[46] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, G. Pfurtscheller, Brain–computer communication: Motivation, aim, and impact of exploring a virtual apartment, IEEE Trans. Neural Syst. Rehabil. Eng. 15 (4) (2007) 473–482, http://dx.doi.org/10.1109/tnsre.2007.906956.

[47] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, S.-W. Lee, EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy, GigaScience 8 (5) (2019) http://dx.doi.org/10.1093/gigascience/giz002.

[48] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, http://dx.doi.org/10.48550/ARXIV.1412.6980.

[49] V. Jayaram, A. Barachant, MOABB: trustworthy algorithm benchmarking for BCIs, J. Neural Eng. 15 (6) (2018) 066011, http://dx.doi.org/10.1088/1741-2552/aadea0.

[50] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80, http://dx.doi.org/10.2307/3001968.

[51] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1) (1995) 289–300, http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[52] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Amer. Statist. Assoc. 32 (200) (1937) 675–701, http://dx.doi.org/10.1080/01621459.1937.10503522.

[53] A. Salami, J. Andreu-Perez, H. Gillmeister, EEG-ITNet: An explainable inception temporal convolutional network for motor imagery classification, IEEE Access 10 (2022) 36672–36685, http://dx.doi.org/10.1109/access.2022.3161489.

[54] C. Neuper, M. Wörtz, G. Pfurtscheller, ERD/ERS patterns reflecting sensorimotor activation and deactivation, in: Event-Related Dynamics of Brain Oscillations, Elsevier, 2006, pp. 211–222, http://dx.doi.org/10.1016/s0079-6123(06)59014-4.

[55] Y. Jeon, C.S. Nam, Y.-J. Kim, M.C. Whang, Event-related (de)synchronization (ERD/ERS) during motor imagery tasks: Implications for brain–computer interfaces, Int. J. Ind. Ergon. 41 (5) (2011) 428–436, http://dx.doi.org/10.1016/j.ergon.2011.03.005.

[56] G. Pfurtscheller, F. Lopes da Silva, Event-related EEG/MEG synchronization and desynchronization: basic principles, Clin. Neurophysiol. 110 (11) (1999) 1842–1857, http://dx.doi.org/10.1016/s1388-2457(99)00141-8.

[57] H. Wang, T. Xu, C. Tang, H. Yue, C. Chen, L. Xu, Z. Pei, J. Dong, A. Bezerianos, J. Li, Diverse feature blend based on filter-bank common spatial pattern and brain functional connectivity for multiple motor imagery detection, IEEE Access 8 (2020) 155590–155601, http://dx.doi.org/10.1109/access.2020.3018962.

[58] G. Pfurtscheller, F.L. Da Silva, Event-related EEG/MEG synchronization and desynchronization: basic principles, Clin. Neurophysiol. 110 (11) (1999) 1842–1857.

[59] A. Spiegler, B. Graimann, G. Pfurtscheller, Phase coupling between different motor areas during tongue-movement imagery, Neurosci. Lett. 369 (1) (2004) 50–54, http://dx.doi.org/10.1016/j.neulet.2004.07.054.

[60] G. Pfurtscheller, C. Brunner, A. Schlögl, F. Lopes da Silva, Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks, NeuroImage 31 (1) (2006) 153–159, http://dx.doi.org/10.1016/j.neuroimage.2005.12.003.

[61] C. Neuper, G. Pfurtscheller, Evidence for distinct beta resonance frequencies in human EEG related to specific sensorimotor cortical areas, Clin. Neurophysiol. 112 (11) (2001) 2084–2097, http://dx.doi.org/10.1016/s1388-2457(01)00661-7.

[62] G.R. Müller-Putz, D. Zimmermann, B. Graimann, K. Nestinger, G. Korisek, G. Pfurtscheller, Event-related beta EEG-changes during passive and attempted foot movements in paraplegic patients, Brain Res. 1137 (2007) 84–91, http://dx.doi.org/10.1016/j.brainres.2006.12.052.

[63] A. Korik, R. Sosnik, N. Siddique, D. Coyle, Decoding imagined 3D hand movement trajectories from EEG: Evidence to support the use of mu, beta, and low Gamma oscillations, Front. Neurosci. 12 (2018) http://dx.doi.org/10.3389/fnins.2018.00130.

[64] G. Pfurtscheller, C. Neuper, Motor imagery activates primary sensorimotor area in humans, Neurosci. Lett. 239 (2–3) (1997) 65–68, http://dx.doi.org/10.1016/s0304-3940(97)00889-6.

[65] A. Adham, B.T. Le, J. Bonnal, H. Bessaguet, E. Ojardias, P. Giraux, P. Auzou, Neural basis of lower-limb visual feedback therapy: an EEG study in healthy subjects, J. NeuroEngineering Rehabil. 21 (1) (2024) http://dx.doi.org/10.1186/s12984-024-01408-8.

[66] G.G. Yener, E. Başar, Brain oscillations as biomarkers in neuropsychiatric disorders, in: Application of Brain Oscillations in Neuropsychiatric Diseases - Selected Papers from “Brain Oscillations in Cognitive Impairment and Neurotransmitters” Conference, Istanbul, Turkey, 29 April–1 May 2011, Elsevier, 2013, pp. 343–363, http://dx.doi.org/10.1016/b978-0-7020-5307-8.00016-8.

[67] G. Assenza, F. Capone, L. di Biase, F. Ferreri, L. Florio, A. Guerra, M. Marano, M. Paolucci, F. Ranieri, G. Salomone, M. Tombini, G. Thut, V. Di Lazzaro, Oscillatory activities in neurological disorders of elderly: Biomarkers to target for neuromodulation, Front. Aging Neurosci. 9 (2017) http://dx.doi.org/10.3389/fnagi.2017.00189.

[68] T.E. Nichols, A.P. Holmes, Nonparametric permutation tests for functional neuroimaging: A primer with examples, Hum. Brain Mapp. 15 (1) (2001) 1–25, http://dx.doi.org/10.1002/hbm.1058.

[69] D. Borra, E. Magosso, Unveiling Multi-Domain Signatures of EEG Oscillations Using a Fully-Interpretable Convolutional Neural Network, Zenodo, 2025, http://dx.doi.org/10.5281/ZENODO.16156955, URL https://zenodo.org/doi/10.5281/zenodo.16156955.