Combining EEG Oscillation Analysis and Explainable Artificial Intelligence for Characterizing Visuospatial Attention

Combining EEG oscillation analysis and explainable artificial intelligence for characterizing visuospatial attention

Elisa Magosso[0000−0002−4673−2974], Paolo Bruno[0009-0008-6399-3405], and Davide Borra[0000−0003−3791−8555]

Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI), University of Bologna, Cesena Campus, Cesena, Italy
davide.borra2@unibo.it

Abstract. Covert visuospatial attention is the ability of human brain to voluntarily direct the attentional focus in the visual space without moving eyes. EEG studies support the involvement of alpha rhythm, mainly acting on posterior regions, in mediating visuospatial attention mechanisms. However, most of the studies considered a small set of regions, focusing on only one or just a few posterior regions, selected a priori. Thus, it remains unclear how other parietal-occipital regions contribute to visuospatial attention. Here, we collected EEG signals during a covert attention-orienting task, requiring to orient attention either to left or right hemifield. We examined EEG-source level signals, performing a whole cortex analysis. Two different approaches were employed to evidence the parietal occipital regions that have a key role in attention-related alpha mechanisms. The first is based on more traditional analyses, evaluating modulation of both alpha power in cortical regions and of alpha-band interregional cortical connectivity, depending on the attended hemifield. The second is based on explainable artificial intelligence (XAI), combining a convolutional neural network and an explanation technique, to reveal the regions more relevant for discriminating left vs. right attention orienting. Both approaches point to superior parietal cortex as a key node in both hemispheres for visual spatial attention, concurrently with other occipital regions (e.g., lateral occipital cortex). This study, although preliminary, shows the potentiality of combining a XAI approach with more traditional methods for increasing the robustness of EEG-source analysis. The same combined approach can be easily transposed to investigate other cognitive tasks.

Keywords: Electroencephalography, Visuospatial attention, Alpha rhythm, Granger causality, Convolutional neural network, Layerwise relevance propagation.

1 Introduction

In everyday life, we need to efficiently cope with an overwhelming stream of visual information. A fundamental ability of human brain is to voluntarily direct the attentional focus in the visual space without moving eyes (covert visuospatial attention), in

order to prioritize and enhance the processing of visual input coming from specific spatial locations, relevant for current or near-future goals, and suppress processing of distracting input coming from task-irrelevant locations.

Studies based on magnetoencephalography (MEG) and electroencephalography (EEG) support a functional role of alpha band oscillations (8-12 Hz) in covert visuospatial attention. Indeed, there is converging evidence [1–5] that alpha power is topographically modulated in parietal-occipital visual sites depending on the direction of visual attention, whit alpha power decrease (alpha desynchronization) in posterior brain areas devoted to processing visual stimuli from the attended locations, and alpha power increase (alpha synchronization) in posterior areas devoted to processing the unattended locations. For example, in case of covert orienting of attention to either hemifield, posterior alpha desynchronization was observed in the hemisphere contralateral to the attended hemifield, accompanied by posterior alpha synchronization in the ipsilateral one [1, 3, 5, 6].

The prominent view in the current literature is that alpha rhythm exerts neuronal excitatory/inhibitory effects [7, 8]. Alpha suppression is thought to reflect an increase in neuronal excitability (or a release of inhibition) thus facilitating the processing of information coming from the attended part of the visual field. On the contrary, alpha enhancement is associated to decreased excitability (or increase of inhibition), thus filtering out possibly interfering visual inputs coming from the unattended locations.

Alpha power in parietal-occipital sensory cortices is thought to be under the control of top-down signals from higher-order brain areas via long-range connectivity, as supported by numerous studies. These include studies that use simultaneous EEG-fMRI recordings [9] to investigate correlations between BOLD signals in frontoparietal areas and attention-related modulation in occipital alpha power. Other studies combine EEG/MEG with TMS to produce transient virtual lesion of frontal or parietal areas and analyze the effects on alpha modulation in occipital parietal cortex [10–12]. Functional interregional interactions within alpha band in visuospatial attention are also investigated by estimating frequency-resolved connectivity from EEG or MEG [5, 13–16]; in particular, the use of directed measures of connectivity, such as Granger Causality (as in [5]) or Transfer Entropy (as in [14]) is especially relevant to assess directionality of causal influences acting on posterior regions. Overall, this body of studies points to alpha-band influences from frontoparietal to posterior parietal and occipital areas as crucial for the control of visuospatial attention.

However, most of these studies limit to examining small sets of regions, focusing on only one or just a few parietal and occipital regions, selected a priori, as loci to assess alpha power modulation and/or alpha communication influences. Thus, it remains unclear whether the selected regions are the most central ones or other parietal occipital regions are more relevant as to alpha mechanisms related to visuospatial attention.

Here, EEG signals were recorded while participants performed a covert attention-orienting task, directing attention either to left or right hemifield depending on a displayed cue. We examined EEG-source level signals by using two different approaches, to evidence the parietal occipital regions that have a key role in attention-related alpha mechanisms. We did not make any a priori selection of the cortical sources and we performed a whole cortex analysis, by parcellating the entire cortex into regions of

interest (ROIs). One adopted approach is based on more traditional analyses, consisting of alpha power computation in ROIs and estimation of interregional cortical connectivity via spectral Granger Causality [17, 18], and it is complemented with Graph Theory [19, 20] to compute centrality indices of the ROIs. The other is an explainable artificial intelligence (XAI) approach (as adopted in [21]), based on a convolutional neural network (CNN) combined with an explanation technique (ET), to reveal the ROIs most relevant in driving the CNN classification between the two attention conditions (attention towards left or right), and thus the ROIs likely more modulated by attention orientation mechanisms. The reason for using two different approaches is twofold. First, we want to show, for the first time, that a CNN can be used for contrasting and analyzing EEG-source level signals related to covert attentional orienting, besides more traditional techniques. Second, we wish to provide a proof-of-principle of how different approaches can be used together to increase robustness of EEG-source analysis.

2 Materials and Methods

2.1 Participants and experimental protocol

Thirteen healthy volunteers (3 female, age: M = 24 year, SD = 2.38 year) participated in the study, which was conducted in accordance with the Declaration of Helsinki and approved by the Bioethics Committee of the University of Bologna (protocol number 29146, year 2019). Written informed consent was obtained from all participants before the beginning of the experiment. All data were analyzed and reported anonymously.

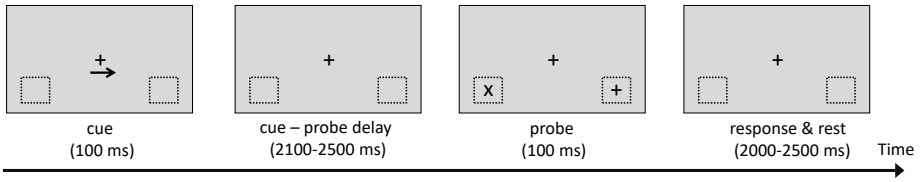

Fig. 1. Timeline of each trial of the covert visuospatial attention task.

The participants seated in a quite dark room and performed a cued visuospatial attention task (Fig. 1). Black visual stimuli were presented onto a grey background on a monitor in front of the participants. Throughout all the task, the participants were instructed to maintain fixation on a cross placed centrally. Each trial started with the brief presentation (100 ms) of an either left- or right-pointing arrow cue, indicating the participant to direct covert attention to the box located in the lower left or lower right visual field. The two directional arrows were equiprobable. After an interval of random duration (2100 - 2500 ms), the probe occurred, consisting in the brief presentation (100 ms) of a symbol ('+' or 'x') either in only one box or in both boxes (8 possible probe configurations, each equiprobable). The trial ended with an interval of a random duration (2000 - 2500 ms), during which the participants were instructed to press a keyboard key only if they detected a '+' stimulus in the cued location. This paradigm aimed at maximizing the attention orienting to the cued location, since the participants never had

to respond to stimuli in the uncued location, suppressing their possible interference. The entire experiment included 10 blocks of trials, with 80 trials per block (40 trials for each directional cue), for a total of 400 trials for each directional cue. The order of presentation of the directional cue and of probe configuration was chosen pseudo-randomly in each block.

We focused the analysis on the cue-probe interval during which the participants oriented the covert attention towards either the left or right visual space according to the preceding cue, and we contrasted the two conditions (in the following denoted as *Attend Left* and *Attend Right* condition) using the two different approaches.

2.2 EEG acquisition, pre-processing and cortical source estimation

For each participant, 64-channel EEG was recorded at 512 Hz sampling rate, using g.HIAMP80 amplifier (g.tec Medical Engineering GmbH, Schiedlberg, UA, Austria), and active electrodes (g.Scarabeo) placed according to the 10-10 system; the reference electrode was on the right earlobe and the ground electrode in AFz. Electrode impedances were kept below 50 kΩ. A notch digital filter (stopband of 48-52 Hz) was applied during recording.

The following EEG pre-processing steps were applied using Matlab custom scripts.

a) Linear detrending of signals belonging to each block.

b) Band-pass filtering (1-60 Hz) of signals belonging to each block.

c) Bad channels identification in each block via random sample consensus method [22].

d) Concatenation of the signals across the blocks and removal of all channels labelled as bad in step c).

e) Removal of ocular and cardiac artifacts via independent component analysis (ICA) [23] applied to the concatenated signals resulting from step d).

f) Spherical spline interpolation of bad channels removed in step c).

g) Average re-referencing.

h) Epoching of the concatenated signals into 1.5s-length epochs, from –1500 ms to 0 ms, with 0 ms denoting the onset of the probe stimuli. Thus, each epoch corresponded to the last 1.5 s of the cue-probe interval. Elimination of the first part of the cue-probe interval from the extracted epochs ensured that the examined signals were not contaminated by the evoked potential activity elicited by cue appearance and disappearance.

At the end of pre-processing, for each participant we had 400 1.5s-epochs for each condition (*Attend Left* and *Attend right*), with 64 scalp electrode signals per epoch.

Cortical source activity was estimated from pre-processed EEG signals using Matlab toolbox Brainstorm, adopting a template head model (ICBM152 MNI template) with the source space restricted to the cortex and discretized into 15002 vertices. The inverse problem was solved using sLORETA (standardized Low Resolution Electromagnetic Tomography) [24] with the dipole source orientation constrained to be perpendicular to the cortex, thus obtaining one source signal for each cortical vertex. Cortical vertices were grouped into 68 cortical ROIs according to the Desikan-Killiany Atlas (see Table 1 and Fig. 2). For each epoch, a waveform representative of the neural activity of each

ROI was derived, by averaging all signals of the vertices belonging to that ROI. At the end of this step, for each participant we had 400 1.5s-epochs for each condition (*Attend Left* and *Attend right*), with 68 cortical ROI signals per epoch. The following analyses (Sections 2.3 and 2.4) were applied at the cortical-ROI level.

Table 1. List of the name of cortical ROIs and their abbreviations (Abbr.). The Desikan Killiany Atlas (34 ROIs, both in left (L) and right (R) hemisphere) was adopted here to parcellate the cortex (lobes are termed as T: temporal, F: frontal, P: parietal, O: occipital).

ROI name	Abbr.	ROI name	Abbr.
Banks superior temporal sulcus (T)	BK	*Parahippocampal gyrus (T)*	PH
Caudal anterior-cingulate cortex (F)	cAC	*Pars opercularis (F)*	pOP
Caudal middle frontal gyrus (F)	cMF	*Pars orbitalis (F)*	pOR
Cuneus cortex (O)	CU	*Pars triangularis (F)*	pTR
Entorhinal cortex (T)	EN	*Pericalcarine cortex (O)*	PCL
Frontal pole (F)	FP	*Postcentral gyrus (P)*	POC
Fusiform gyrus (T)	FU	*Posterior-cingulate cortex (P)*	PCG
Inferior parietal cortex (P)	IP	*Precentral gyrus (F)*	PRC
Inferior temporal gyrus (T)	IT	*Precuneus cortex (P)*	PCU
Insular cortex	IN	*Rostral ant. cingulate cortex (F)*	rAC
Isthmus–cingulate cortex (P)	IST	*Rostral middle frontal gyrus (F)*	rMF
Lateral occipital cortex (O)	LO	*Superior frontal gyrus (F)*	SF
Lateral orbital frontal cortex (F)	lOF	*Superior parietal cortex (P)*	SP
Lingual gyrus (O)	LG	*Superior temporal gyrus (T)*	ST
Medial orbital frontal cortex (F)	mOF	*Supramarginal gyrus (P)*	SMG
Middle temporal gyrus (T)	MT	*Temporal pole (T)*	TP
Paracentral lobule (F)	PAC	*Transverse temporal cortex (T)*	TT

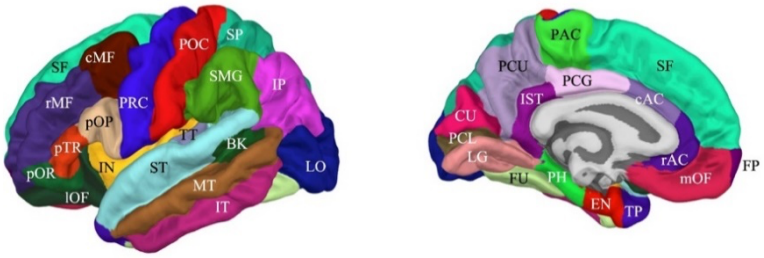

Fig. 2. Cortex parcellation according to the Desikan-Killiany Atlas (left: lateral view, right: medial view), showing cortical ROIs location. The abbreviation of each ROI is reported in Table 1.

2.3 Approach 1: Alpha power, connectivity and centrality indices

Alpha-band power. For each participant and each attention condition, power spectral density (PSD) was computed at each ROI by concatenating all 400 epochs (Welch's periodogram method with Hamming window of 0.5 s, 0.4 s overlap and 4 s zero-padding). Then, from the resulting PSDs, alpha-band power (8-12 Hz) was obtained for each cortical ROI in *Attend Left* (*Pow Attend Left*) and in *Attend Right* (*Pow Attend Right*). We then computed, ROI by ROI, an index of alpha power modulation contrasting the two conditions, according to Eq. (1) as similarly done in other studies (e.g.[1, 5, 25])

$$alpha\ modulation\ index = \frac{Pow\ Attend\ Left - Pow\ Attend\ Right}{Pow\ Attend\ Left + Pow\ Attend\ Right} \cdot 100 \quad (1)$$

Of course, positive values of this index correspond to *Pow Attend Left>Pow Attend Right*, while negative values to *Pow Attend Left<Pow Attend Right*. For each ROI, two one-tailed non-parametric permutation t-tests (5000 permutations) were applied to identify cortical regions having this index significantly greater or less than 0, and False Discovery Rate correction for multiple comparisons was applied.

Alpha-band connectivity. We employed spectral Granger Causality (GC) [17, 18] to estimate the directional causal influences between each pair of ROIs in the alpha band, and to assess differences depending on the attention condition. We used autoregressive processes of order 30 as done in other recent studies [21, 26, 27]. For each participant and each attention condition, the spectral GC ($GC_{i\rightarrow j}(f)$) was computed, from each ROI i to each ROI j, over the concatenated epochs belonging to that condition, at each frequency f); then, alpha-band GC was obtained for each pair of ROIs by averaging the spectral GC over the range of alpha frequencies. This resulted in two 68×68 non-symmetric connectivity matrices for each participant (one for *Attend Left* one for *Attend Right*). In each matrix, the off-diagonal ij-th value quantified the alpha-band causal influence from ROI i to ROI j. From these matrices, we derived sparse connectivity matrices via statistical comparison between the two attention conditions, in order to reduce the noise. To this aim, a two-tailed permutation t-test (5000 permutations) was employed, and significant connections were defined as having an uncorrected p-value less than 0.05. Consequently, among the possible 68 × 68 connections, only the significantly different connections between the two conditions were retained for further analysis, while all connections resulting non-significant (uncorrected p-value higher than 0.05) were set to zero in each participant. In this way, two sparse connectivity matrices were obtained for each participant, one per attention condition, and used for computing centrality indices.

Centrality indices from Graph Theory. For each participant, the sparse connectivity matrix of each attention condition can be interpreted as an adjacency matrix A (68 x 68) representing a directed graph. The nodes of the graph are the 68 cortical ROIs and the element A_{ij} of the matrix represents the weight of the edge from node i to node j. Under this representation, we can compute centrality indices for each node, derived from Graph Theory. Centrality indices assign a score to each node based on some aspect of importance. Here we used *authority* (a) and *hubness* (h), which measure importance of each node in relation to inflow and outflow of information, respectively. These indices are strictly interdependent and have the following mathematical formulation.
Authority of node i (a_i) is the sum of the weights of edges entering into the node, with each weight multiplied by the *hubness* of the node the edge originates from

$$Authority\ (a_i) = \sum_j A_{ji} h_j \quad (2)$$

Hubness of node i (h_i) is the sum of the weights of edges exiting from the node, with each weight multiplied by the authority of the node the edge originates from

$$Hubness\ (h_i) = \sum_j A_{ij} a_j \tag{3}$$

These indices are computed trough an iterative procedure until convergence.

Hubness and authority can be interpreted as measuring which nodes in a graph are primarily 'broadcasters' and which are primarily 'receivers', respectively. However, they do not take into account only the number and strengths of the exiting connections and of the entering connections, but also the centrality of the targeting nodes and of the sending nodes. So, for example, a strong entering connection coming from a node which is a poor broadcaster may be of little importance compared to a lower connection coming from an important broadcaster. Thus, these measures better summarize the effective significance of the inflow and outflow of a node compared to simply the sum of entering and exiting connections. It is clear that a strong hub points to many strong authorities and a strong authority receives from many strong hubs.

For each participant, hubness and authority of each ROI were computed for *Attend Left* and *Attend Right* condition. Then, separately for each centrality index, two one-tailed permutation t-tests were applied to assess which ROIs have the index greater or less in *Attend Left* compared to *Attend Right*; for each test, False Discovery Rate correction for multiple comparisons (68 comparisons) was applied.

2.4 Approach 2: Neural network and explanation technique

Convolutional neural network. The CNN adopted here was previously used in Borra et al. [21] to perform a CNN-based analysis of the cortex-level activity in a different cognitive task (involving face and body inversion effect). The network consists of a shallow architecture, trained here for discriminating among *Attend Left* and *Attend Right* condition from each single-trial source-level activity. The input to the network is denoted by the matrix $X \in \mathbb{R}^{68\times T}$ (68 is the number of ROIs and T the number of time steps, see below). To reduce the computational cost, in this analysis the input neural activity was downsampled to 128 Hz.

A brief description of the network is provided below; further details can be found in the reference publication. First, the network processes the input X using 2D convolutions in time to learn how to optimally filter each ROI signal (kernel size of (1,33)); then, 2D convolutions in space are applied separately for each filtered version of the input (spatial depthwise convolution) to learn how to optimally recombine the information across the different ROIs. Sixteen different temporal kernels (i.e., 16 different filtered versions of the input), and 2 different spatial kernels for each filtered version of the input (thus, 32 in total) were learned. Neurons are activated via rectified linear functions. The output from the previous processing is downsampled in time (2D average pooling, with a pooling size and stride of (1,10) and (1,2), respectively) to reduce the computational cost. Finally, neuron activations are provided to a fully-connected layer with 2 output neurons, one per class (i.e., *Attend Left* and *Attend Right*). Output neurons are activated via softmax function, for obtaining the probability that the input trial contains the neural response associated to *Attend Left* and *Attend Right* condition. To

improve generalization, batch normalization and dropout are used after each convolutional layer (with a dropout rate of 0.1). Overall, the network accepts as input the cortex-level activity, and provides as output the probability that it contains the response elicited while attending the left vs. right hemifield in the spatial attention task.

A within-subject training strategy was employed for training networks, thus designing subject-specific neural networks (13 total networks). Each participant-specific dataset consisted of 800 cortex-level epochs (400 per spatial attention condition), and was divided into a training and test set using a fixed 80-20% random split. Moreover, we reserved 20% of the training set as validation set, thus resulting into 512, 128, and 160 training, validation and test epochs. To increase the number of training examples, we adopted a sliding window decoding approach along the time axis, that is, we extracted 0.5s-length chunks of neural activity from the 1.5s-length epoch, with a stride of 0.1 s. Therefore, the input time steps T processed by the network was equal to 64 (0.5 s at 128 Hz). With this procedure, the dataset was augmented 11 times, thus we used 5632, 1408 and 1760 training, validation and test examples.

The network trainable parameters (4.6k in total) were optimized up to 500 training epochs using the cross-entropy as loss function and Adam as optimizer (learning rate of 1e-4, mini-batch size of 64). Early stopping was employed, returning the network parameters at the epoch with maximum validation accuracy.

Explanation technique. Once the networks were trained for discriminating among the two attentional states, the knowledge learned by the decoder – embedded in the trainable parameters – was exploited for deriving representations that quantify the relevance of each input ROI and time sample for maximally separating the spatial attentional states. Specifically, we computed the layerwise relevance propagation (LRP) [28]. This technique operates by propagating the activation of output neurons (preceding the softmax activation function) back to the input layer, by exploiting local propagation rules applied at each layer forming the network. This explanation technique was successfully applied in prior studies for explaining decision of networks applied to EEG and EEG-derived cortex activity [21, 29, 30]. LRP was applied for each test example and provided as output a 2D relevance representation with the same shape of the input (i.e., $\in \mathbb{R}^{68\times T}$), quantifying how much each input ROI and time sample contributed to the discrimination of the output attentional states. In general, LRP provides both positive and negative values in its maps, associated to positive and negative evidence for the model prediction. As the decoding problem was a binary classification task, we considered just the LRP map associated to the *Attend Right* condition, as the other LRP map (associated to *Attend Left* condition*)* simply presents a reverted relevance. Moreover, as we were interested in analyzing the absolute relevance, we computed the absolute value of LRP maps and we then averaged maps together across test examples and across time samples, obtaining an absolute relevance array $\in \mathbb{R}^{68}$ that quantifies the relevance of each ROI, regardless the sign. To identify the group of the most relevant ROIs, we performed the same strategy employed in [21]. The most relevant ROI, on average across subjects, was identified. Then, a pairwise Wilcoxon signed-rank test was performed, comparing the relevance distribution of the most relevant ROI with each other ROI. P-values were corrected for multiple tests (67 in total). The group of the most

relevant ROIs was formed by the ROIs that do not significantly differ from the most relevant ROI.

3 Results and Discussion

3.1 Approach 1

Fig. 3 shows the alpha modulation index in all ROIs, averaged across participants. ROIs contralateral to a visual hemifield exhibited a relative alpha power increase in case of non-attending vs. attending the hemifield. This is more consistent (see asterisks) in posterior parietal and occipital ROIs and also in some right temporal ROIs. The modulation is maximal in bilateral superior parietal, inferior parietal and lateral occipital ROIs. Overall, this topological alpha power modulation related to visuospatial attention orienting is in line with previous EEG/MEG studies at source level reporting alpha modulation spreading beyond the occipital cortex to parieto-temporal cortices [11, 13, 15, 25].

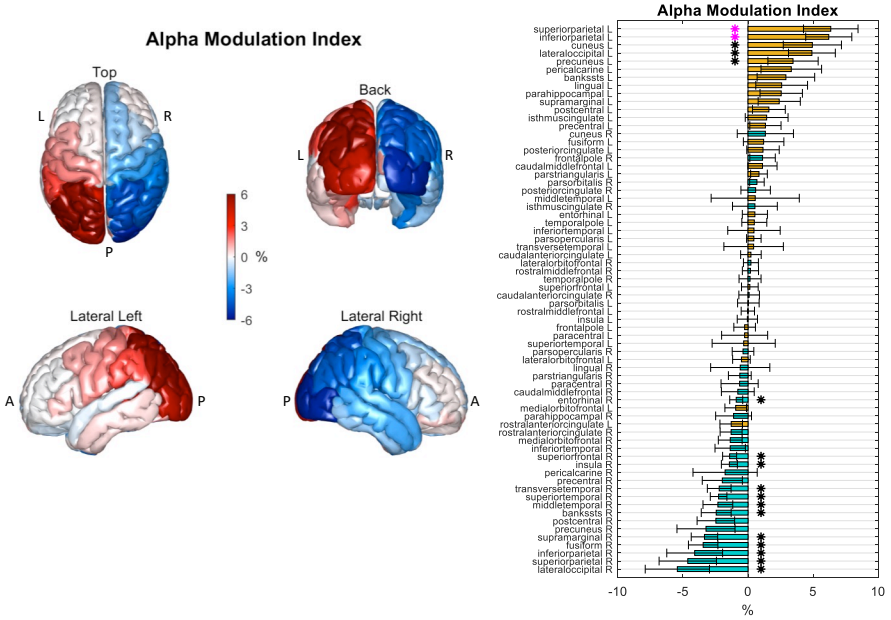

Fig. 3. *Left Panel*: Alpha Modulation Index in each ROI (mean across the participants) displayed over the cortex, according to different views. Red shades denote positive values, blue shades negative values. L: Left, R: Right; A: Anterior; P: Posterior. *Right Panel*: Alpha Modulation Index in each ROI (mean and standard error of the mean across the participants) displayed in a bar plot. Yellow bars correspond to left hemisphere ROIs, cyan bars to right hemisphere ROIs. Magenta and black asterisks denote ROIs with Alpha Modulation Index statistically different from 0, with and without correction for multiple comparisons, respectively.

As to alpha-band connectivity (Fig. 4), stronger values of connectivity were lateralized towards the ROIs, especially posterior and temporal, in the hemisphere

contralateral to the unattended hemifield. Thus, these ROIs were implicated in stronger alpha-band coupling than homologous ROIs contralateral to the attended hemifield. Hence, in addition to alpha-band power modulation, an evident strengthening of alpha-band communication in the hemisphere contralateral to the unattended hemifield was observed, similar to a prior study using undirected measure of connectivity [15].

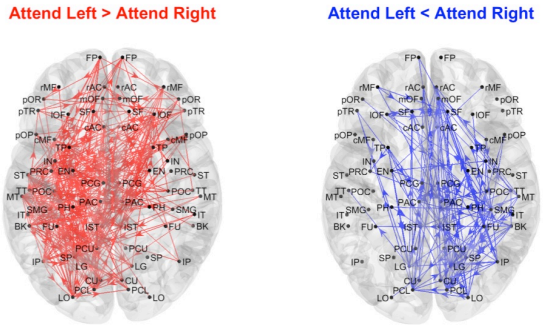

Fig. 4. Connections, represented over the cortex (top view), statistically different ($p<0.05$, uncorrected) between the two conditions (i.e., taken from the sparse connectivity matrices), stronger in Attend Left compared to Attend Right (red connections, left map) and stronger in Attend Right compared to Attend Left (blue connections, right map).

Fig. 5 shows the ROIs exhibiting significant difference in hubness and authority between the two attention conditions. Focusing on posterior ROIs, hubness and authority assumed larger values mainly in ROIs contralateral to the unattended hemifield. By further restricting the argument on authority (as here we are mainly interested in alpha-band influences exerted over posterior regions), *Attend Left* condition was associated with higher authority in left SP and bilateral cuneus, while *Attend Right* with higher authority in right SP and some right occipital ROIs (LG, PLC, LO). Interestingly, based on these results, only the superior parietal resulted to have the authority index oppositely modulated in the two hemispheres when attending to one relative to the other hemifield (e.g., when attending left hemifield compared to right hemifield, authority in left SP increased while in right SP decreased). The pattern of connectivity entering into the posterior authority-modulated ROIs shows that SP, both left and right, had several frontal bilateral sources, denoting a clear top-down frontoparietal pattern that governs the alpha-band inflow of SP. Conversely, the other posterior regions exhibiting authority modulation were not or little targeted by frontal regions, and were rather driven by temporal and parietal sources, including also SP.

Overall, this first approach points to SP as key node in both hemispheres involved in alpha mechanisms related to visuospatial attention, exhibiting strong alpha power modulation and direct top-down alpha-band influences from frontal regions. Occipital regions such as cuneus, lateral occipital, pericalcarine, appeared relevant too, exhibiting alpha-band modulation (in power and/or authority), but with little direct frontal influx. These results, although preliminary, are in line with the role of SP; the latter, together with intraparietal sulcus and frontal eye fields, is considered part of the dorsal attention

network, implicated in attentional functions including spatial orientation [31, 32]. Moreover, some studies [9, 10, 33] suggest that attention-related alpha activity in visual occipital regions are modulated by parietal regions rather than directly by frontal regions, while frontal regions would be directly coupled to parietal regions.

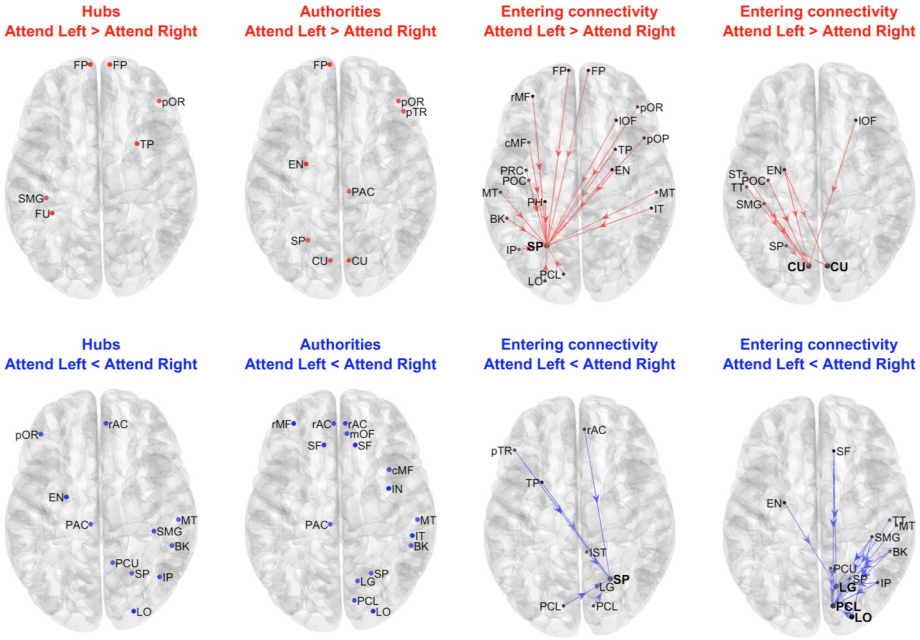

Fig. 5. *Upper panels (red color)* - ROIs exhibiting hubness and authority indices significantly higher ($p<0.05$ corrected) in Attend Left compared to Attend Right (first and second upper maps). Connections significantly higher in Attend Left compared to Attend Right entering into posterior ROIs with higher authority index (SP L, CU L and CU R), considering separately SP for sake of clarity. *Lower panels (blue color)* -The same as upper panels but with reference to quantity significantly higher in Attend Right Compared to Attend Left.

3.2 Approach 2

The neural network scored a decoding accuracy of 64.1 ± 3.3% (mean ± standard error of the mean across subjects), and the performance of each subject-specific decoder was above the chance level (50%). The relatively modest decoding performance obtained in this decoding problem could be related to the structure of the adopted CNN and the nature of the visuospatial cognitive task. CNN structures proposed in the literature, including the one adopted here in this study, are mainly designed for detecting event-related components at the single-trial EEG level [21, 34, 35]. On the other hand, the neural signatures that could help in the discrimination between attentional states are mainly related to modulations of cortical oscillations (e.g., alpha-band power and connectivity modulations). In this study, we show preliminary results by exploiting a decoding framework previously used in the literature for investigating modulations in

event-related components (specifically, N170 modifications related to face and body inversion effects). However, we speculate that in order to improve the decoding performance and thus the reliability of derived XAI representations, future studies should consider CNN designs tailored to the neural signatures under investigation, e.g., incorporating architectural elements learning features related to EEG oscillations.

Fig. 6 reports the relevance of each ROI in discriminating the spatial attentional states, obtained by exploiting the XAI approach. It is worth highlighting that this is the first time that an XAI approach is used for providing insights in covert visuospatial attention processes. From this figure, several bilateral ROIs were among the most relevant ones, including mostly occipital (LO, CU, PCL cortices), temporal (FU gyri), and parietal regions (SP cortices). Interestingly, although the network was trained using the entire cortex activity, i.e. using also frequency bands other than alpha-band, the ranking based on LRP-derived measures resembles the results obtained in Approach 1, suggesting that alpha-band, more than other bands, mainly drives network decision. The ROIs identified as more relevant by the network where not only LO and SP, which resulted maximally modulated in alpha power (Fig.3), but also bilateral CU and PCL that appeared less modulated in alpha-power, but involved in authority modulation (Fig.4). It may be possible that the XAI approach has an enhanced capability of detecting more subtle alpha-band changes compared to traditional power spectral analysis, such as changes related to connectivity modulation. Of course, it may also be possible that the XAI approach is sensitive to other frequency bands modulations and the relevance of the ROI reflects these further influences.

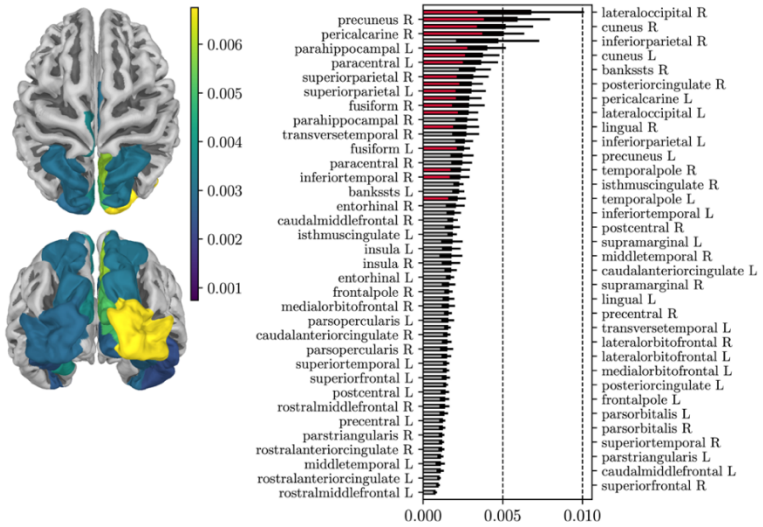

Fig. 6. Relevance of each ROI for discriminating spatial attentional states. On the left panels, the relevance is mapped onto the cortex (dorsal and caudal views displayed) in its mean value across subjects; only the group of the most relevant ROIs is highlighted, as resulting from the performed statistical analysis. On the right panel, the same information is reported as a ranking, to facilitate the individuation of the most relevant ROIs. Bar heights and error bars denote the mean value and the standard error of the mean across subjects, respectively. The group of the most relevant ROIs, obtained with the performed statistical analysis, is colored in red.

4 Conclusions

In conclusion, we analyzed EEG correlates of covert visuospatial attention at the whole cortex-level by applying two approaches: one (Approach 1) based on more traditional but still powerful methods, and one (Approach 2) based on more recent explainable AI technique. The first approach – considering power results and connectivity-derived results – points to the superior parietal as a key node in both hemispheres involved in alpha mechanisms related to visuospatial attention; this node receives direct frontal top-down alpha influences, and contributes to redistribute them to occipital regions (e.g. CU, PCL, LO). Utilizing the second approach, we also found – independently from the first approach – that the superior parietal, but also the other occipital ROIs, are among the ones encoding most for spatial attention, further supporting the results found with more traditional analyses. Of course, the results presented here should be intended as preliminary, given the limited number of participants, and clearly require to be further validated on a larger sample in the future. However, despite the small sample size, interesting and promising results were obtained. Overall, we showed that an explainable AI framework can be conveniently used together with traditional analyses for increasing the robustness of EEG-source analysis. Moreover, as being application-independent, the adopted analysis framework can be easily transposed by neuroscientists to investigate other cognitive tasks.

Acknowledgments. This work is supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) — A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wöstmann, M., Maess, B., Obleser, J.: Orienting auditory attention in time: Lateralized alpha power reflects spatio-temporal filtering. NeuroImage. 228, 117711 (2021). https://doi.org/10.1016/j.neuroimage.2020.117711.

2. Sauseng, P., Klimesch, W., Stadler, W., Schabus, M., Doppelmayr, M., Hanslmayr, S., Gruber, W.R., Birbaumer, N.: A shift of visual spatial attention is selectively associated with human EEG alpha activity. European Journal of Neuroscience. 22, 2917–2926 (2005). https://doi.org/10.1111/j.1460-9568.2005.04482.x.

3. Thut, G., Nietzel, A., Brandt, S.A., Pascual-Leone, A.: Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. J Neurosci. 26, 9494–9502 (2006). https://doi.org/10.1523/JNEUROSCI.0875-06.2006.

4. Rihs, T.A., Michel, C.M., Thut, G.: Mechanisms of selective inhibition in visual spatial attention are indexed by alpha-band EEG synchronization. Eur J Neurosci. 25, 603–610 (2007). https://doi.org/10.1111/j.1460-9568.2007.05278.x.

5. Wang, C., Rajagovindan, R., Han, S.-M., Ding, M.: Top-Down Control of Visual Alpha Oscillations: Sources of Control Signals and Their Mechanisms of Action. Front Hum Neurosci. 10, 15 (2016). https://doi.org/10.3389/fnhum.2016.00015.

6. Worden, M.S., Foxe, J.J., Wang, N., Simpson, G.V.: Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. J Neurosci. 20, RC63 (2000).

7. Jensen, O., Mazaheri, A.: Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition. Front Hum Neurosci. 4, 186 (2010). https://doi.org/10.3389/fnhum.2010.00186.

8. Foxe, J.J., Snyder, A.C.: The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention. Front. Psychol. 2, (2011). https://doi.org/10.3389/fpsyg.2011.00154.

9. Liu, Y., Bengson, J., Huang, H., Mangun, G.R., Ding, M.: Top-down Modulation of Neural Activity in Anticipatory Visual Attention: Control Mechanisms Revealed by Simultaneous EEG-fMRI. Cereb Cortex. 26, 517–529 (2016). https://doi.org/10.1093/cercor/bhu204.

10.Capotosto, P., Babiloni, C., Romani, G.L., Corbetta, M.: Differential contribution of right and left parietal cortex to the control of spatial attention: a simultaneous EEG-rTMS study. Cereb Cortex. 22, 446–454 (2012). https://doi.org/10.1093/cercor/bhr127.

11.Marshall, T.R., O'Shea, J., Jensen, O., Bergmann, T.O.: Frontal Eye Fields Control Attentional Modulation of Alpha and Gamma Oscillations in Contralateral Occipitoparietal Cortex. J. Neurosci. 35, 1638–1647 (2015). https://doi.org/10.1523/JNEUROSCI.3116-14.2015.

12.Peylo, C., Hilla, Y., Sauseng, P.: Cause or consequence? Alpha oscillations in visuospatial attention. Trends Neurosci. 44, 705–713 (2021). https://doi.org/10.1016/j.tins.2021.05.004.

13.Siegel, M., Donner, T.H., Oostenveld, R., Fries, P., Engel, A.K.: Neuronal synchronization along the dorsal visual pathway reflects the focus of spatial attention. Neuron. 60, 709–719 (2008). https://doi.org/10.1016/j.neuron.2008.09.010.

14.Doesburg, S.M., Bedo, N., Ward, L.M.: Top-down alpha oscillatory network interactions during visuospatial attention orienting. Neuroimage. 132, 512–519 (2016). https://doi.org/10.1016/j.neuroimage.2016.02.076.

15.Lobier, M., Palva, J.M., Palva, S.: High-alpha band synchronization across frontal, parietal and visual cortex mediates behavioral and neuronal effects of visuospatial attention. Neuroimage. 165, 222–237 (2018). https://doi.org/10.1016/j.neuroimage.2017.10.044.

16.Van der Lubbe, R.H.J., Panek, B., Jahangier, I., Asanowicz, D.: Lateralized connectivity in the alpha band between parietal and occipital sources when spatial attention is externally and internally directed. Front. Cognit. 2, (2023). https://doi.org/10.3389/fcogn.2023.1145854.

17.Ding, M., Chen, Y., Bressler, S.L.: Granger Causality: Basic Theory and Application to Neuroscience. In: Handbook of Time Series Analysis. pp. 437–460. John Wiley & Sons, Ltd (2006). https://doi.org/10.1002/9783527609970.ch17.

18.Chicharro, D.: On the spectral formulation of Granger causality. Biol Cybern. 105, 331–347 (2011). https://doi.org/10.1007/s00422-011-0469-z.

19.Sporns, O.: Graph theory methods: applications in brain networks. Dialogues Clin Neurosci. 20, 111–121 (2018).

20.Farahani, F.V., Karwowski, W., Lighthall, N.R.: Application of Graph Theory for Identifying Connectivity Patterns in Human Brain Networks: A Systematic Review. Front Neurosci. 13, 585 (2019). https://doi.org/10.3389/fnins.2019.00585.

21.Borra, D., Bossi, F., Rivolta, D., Magosso, E.: Deep learning applied to EEG source-data reveals both ventral and dorsal visual stream involvement in holistic processing of social stimuli. Sci Rep. 13, 7365 (2023). https://doi.org/10.1038/s41598-023-34487-z.

22.Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM. 24, 381–395 (1981). https://doi.org/10.1145/358669.358692.

23.Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7, 1129–1159 (1995). https://doi.org/10.1162/neco.1995.7.6.1129.

24.Pascual-Marqui, R.D.: Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. Methods Find Exp Clin Pharmacol. 24 Suppl D, 5–12 (2002).

25.Marshall, T.R., Bergmann, T.O., Jensen, O.: Frontoparietal Structural Connectivity Mediates the Top-Down Control of Neuronal Synchronization Associated with Selective Attention. PLOS Biology. 13, e1002272 (2015). https://doi.org/10.1371/journal.pbio.1002272.

26.Magosso, E., Ricci, G., Ursino, M.: Alpha and theta mechanisms operating in internal-external attention competition. J Integr Neurosci. 20, 1–19 (2021). https://doi.org/10.31083/j.jin.2021.01.422.

27.Pirazzini, G., Starita, F., Ricci, G., Garofalo, S., di Pellegrino, G., Magosso, E., Ursino, M.: Changes in brain rhythms and connectivity tracking fear acquisition and reversal. Brain Struct Funct. 228, 1259–1281 (2023). https://doi.org/10.1007/s00429-023-02646-7.

28.Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE. 10, e0130140 (2015). https://doi.org/10.1371/journal.pone.0130140.

29.Sturm, I., Lapuschkin, S., Samek, W., Müller, K.-R.: Interpretable deep neural networks for single-trial EEG classification. Journal of Neuroscience Methods. 274, 141–145 (2016). https://doi.org/10.1016/j.jneumeth.2016.10.008.

30.Korda, A.I., Ventouras, E., Asvestas, P., Toumaian, M., Matsopoulos, G.K., Smyrnis, N.: Convolutional neural network propagation on electroencephalographic scalograms for detection of schizophrenia. Clin Neurophysiol. 139, 90–105 (2022). https://doi.org/10.1016/j.clinph.2022.04.010.

31.Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci. 3, 201–215 (2002). https://doi.org/10.1038/nrn755.

32.Petersen, S.E., Posner, M.I.: The Attention System of the Human Brain: 20 Years After. Annu Rev Neurosci. 35, 73–89 (2012). https://doi.org/10.1146/annurev-neuro-062111-150525.

33.Meehan, T.P., Bressler, S.L., Tang, W., Astafiev, S.V., Sylvester, C.M., Shulman, G.L., Corbetta, M.: Top-down cortical interactions in visuospatial attention. Brain Struct Funct. 222, 3127–3145 (2017). https://doi.org/10.1007/s00429-017-1390-6.

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

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