Understanding the role of self-attention in a Transformer model for the discrimination of SCD from MCI using resting-state EEG

Elena Sibilano, Domenico Buongiorno, Michael Lassi, Antonello Grippo, Valentina Bessi, Sandro Sorbi, Alberto Mazzoni, Vitoantonio Bevilacqua, and Antonio Brunetti

***Abstract*— The identification of EEG biomarkers to discriminate Subjective Cognitive Decline (SCD) from Mild Cognitive Impairment (MCI) conditions is a complex task which requires great clinical effort and expertise. We exploit the self-attention component of the Transformer architecture to obtain physiological explanations of the model's decisions in the discrimination of 56 SCD and 45 MCI patients using resting-state EEG. Specifically, an interpretability workflow leveraging attention scores and time-frequency analysis of EEG epochs through Continuous Wavelet Transform is proposed. In the classification framework, models are trained and validated with 5-fold cross-validation and evaluated on a test set obtained by selecting 20% of the total subjects. Ablation studies and hyperparameter tuning tests are conducted to identify the optimal model configuration. Results show that the best performing model, which achieves acceptable results both on epochs' and patients' classification, is capable of finding specific EEG patterns that highlight changes in the brain activity between the two conditions. We demonstrate the potential of attention weights as tools to guide experts in understanding which disease-relevant EEG features could be discriminative of SCD and MCI.**

***Index Terms*— Alzheimer's disease, Multi-Head Attention, Interpretability, Transformer, Resting-state EEG**

"This study was supported by BRIEF - Biorobotics Research and Innovation Engineering Facilities - Missione 4, 'Istruzione e Ricerca' - Componente 2, 'Dalla ricerca all'impresa' Linea di investimento 3.1, 'Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione', funded by European Union - NextGenerationEU, CUP: J13C22000400007" and by Tuscany Region — PRedicting the EVolution of SubjectIvE Cognitive Decline to Alzheimer's Disease With machine learning—PREVIEW - CUP: D18D20001300002. *(Corresponding author: Vitoantonio Bevilacqua)*

E. Sibilano, D. Buongiorno, V. Bevilacqua, and A. Brunetti are with the Department of Electrical and Information Engineering, Polytechnic University of Bari, Bari, Italy (e-mail: elena.sibilano@poliba.it; domenico.buongiorno@poliba.it; vitoantonio.bevilacqua@poliba.it; antonio.brunetti@poliba.it).

M. Lassi and A. Mazzoni are with The BioRobotics Institute, Scuola Superiore Sant'Anna, 56025 Pisa, Italy (e-mail: michael.lassi@santannapisa.it; alberto.mazzoni@santannapisa.it).

A. Grippo and S. Sorbi are with the IRCCS Fondazione Don Carlo Gnocchi, 50143 Florence, Italy (e-mail: antonello.grippo@unifi.it; sorbi@unifi.it).

V. Bessi and S. Sorbi are with the Department of Neuroscience, Psychology, Drug Research and Child Health, University of Florence, Azienda Ospedaliera Careggi, Florence, Italy (e-mail: valentina.bessi@unifi.it; sorbi@unifi.it).

I. Introduction

DEMENTIA disorders stand as the primary contributor to disability in the elderly population, since they encompass a variety of cognitive and behavioral symptoms that disrupt the capacity to carry out daily activities [1]. These conditions are progressive, and often preceded by less severe stages of impairment. Subjective Cognitive Decline (SCD) refers to a self-reported worsening of cognitive abilities experienced by an individual, without objective evidence of impairment assessed through standardized cognitive tests [2], while Mild Cognitive Impairment (MCI) is a clinical condition that results in a noticeable decline in cognitive or behavioural domains, assessed by standardized tests used for diagnostic purposes and to differentiate it from normal age-related cognitive changes [3]. SCD and MCI can be caused by various underlying factors. However, both conditions may represent early manifestations of dementia disorders, particularly Alzheimer's Disease (AD), as they have been associated with an increased likelihood of biomarkers' abnormalities consistent with AD pathology and with a higher risk for future AD [4], [5]. While some cases may progress to more severe impairment, other cases may remain stable or even move toward the recovery, or at least the non-worsening, of impaired domains. Hence, the characterization and identification of biomarkers related to SCD and MCI is essential to monitor individuals which will eventually develop the disease [5]. Additionally, with the advance of research for disease-modifying therapies for AD [6], the ability to accurately identify patients in the initial stages of the disease is pivotal for timely and targeted interventions to potentially alter the disease's course [7]. From a functional perspective, the discrimination between SCD and MCI is essential to target populations of subjects in the preclinical and prodromal states of AD, before irreversible neurodegeneration occurs. However, since morphological and functional changes in the brain between the two conditions are variable and partly overlapping, this task requires significant effort from domain experts.

Recently, there has been a growing interest in using Deep Learning (DL) techniques to automatically identify AD from neuroimaging data, such as magnetic resonance imaging (MRI) and positron emission tomography (PET) [8], but more prominently from electroencephalography (EEG) data [9],

[10]. Indeed, EEG has the advantages of non-invasiveness, ease of use and cost-effectiveness, representing a promising and more reproducible alternative to traditional diagnostic methods [11]. In particular, resting-state EEG (rsEEG) has been confirmed as a biomarker related to disease progression [12]–[14] which can be used to discriminate early stages of AD [15]. In this context, rsEEG-based DL models have also shown remarkable results in the identification of MCI [16], [17] and preclinical AD [18], [19].

However, a well-known limitation concerning the application of DL algorithms to clinical practice is the underlying black-box behaviour of the models, which can reduce their reliability and generate mistrust about their results. Explainability and visualization methods, such as GradCAM or LIME, have already been employed in tasks for the classification of biological signals [20], [21]; however, a trustworthy understanding of DL algorithms supporting decisions in healthcare is essential and still needed [22], [23]. The concept of *interpretability* could represent a valid approach to deal with this problem. An interpretable model is able to show why a decision is made for a specific input [24], by exposing the inner mechanisms through human-understandable explanations. The design and improvement of methods to enhance the interpretability aspects of DL models are currently open research topics [8], [22], [23].

Transformers [25] and Vision Transformers [26] have introduced a new approach to the interpretability of deep networks in the fields of Natural Language Processing and Computer Vision through the mechanism of self-attention. Indeed, while most of the explainability methods focus on giving information about how a model processes data or how it represents data internally, attention-based architectures generate explanation-producing systems by directly revealing which information flows through the network [27]. Specifically, attention can help access a model's inherent processes by showing how it assigns different weights to different inputs and parts of the input [28].

Various interpretability methods have been proposed for models based on Transformers [29]. Nonetheless, one effective approach is to leverage raw attention scores to visualize the portions of the input on which the model focused the most during the decision process [30], [31], particularly when working with time series [32].

In this context, the potential of explanations obtained from attention weights has been explored in the field of EEG classification. Recently, several works employing Transformers for EEG-based automatic sleep staging have highlighted the importance of understanding which physiologically interpretable patterns are detected by the model and how different parts of the input influence the classification outcome [33]–[35]. In particular, by mapping attention scores to the raw EEG signals, they all found that the Transformers attended more to sleep-related features, such as K-complexes and spindles, to classify specific sleep stages. Similarly, a study using EEG for emotion recognition employed the output weights of attention modules to visualize the importance of different EEG channels in emotional activities [36] and showed that the channels with higher attention focus were located in the temporal and prefrontal lobes, which play a major role in human emotion regulation. Likewise, Baghdadi *et al.* performed an EEG-based seizure type classification task and observed that the attention block in their model assigned larger weights to channels and time blocks that contributed to characterize specific seizure types [37].

Although several studies have demonstrated the potential of Transformers for the classification of AD and its transitional stages using MRI [38]–[41] and PET images [42], [43], a very limited number of works have employed these models with EEG data [44]–[46] and none of them has investigated the potential of attention scores as interpretability tools to find pathological biomarkers of AD.

In our previous study, we proposed an innovative approach to classify SCD and MCI subjects using rsEEG with a deep model based on the Transformer architecture [47]. Specifically, the model processed rsEEG epochs of 10 seconds, predicted a label for each epoch, and subsequently classified each subject based on the class predicted for most of the corresponding epochs. We demonstrated that the Transformer model was able to distinguish between SCD and MCI subjects with good accuracy levels, comparable to those obtained with state-of-the-art DL models employed for EEG signal classification tasks.

In this work, we argue that our model allows to provide explanations for its decisions and support the identification of alterations in the brain activity of SCD and MCI patients by detecting patterns of interest in the input signals. In addition, we aim to provide a further analysis of our method by tuning parameters and performing ablation studies on different modules of the Transformer in order to highlight the role of the self-attention component in the classification process. The main contributions of this paper are summarised in the following:

- we propose an interpretability framework that leverages attention scores to qualitatively and quantitatively identify peculiar changes in the rsEEG signals between SCD and MCI patients. The proposed approach allows the automatic attention-guided segmentation of EEG epochs to locate the most meaningful parts of the signal;
- we integrate statistical and time-frequency analyses to correlate the decisions made by the Transformer with the physiological basis of cognitive impairment in the specific classification task;
- we investigate how choosing multi-head attention over traditional self-attention impacts the explainability capabilities of the model.

To the best of our knowledge, this is the first work that proposes a systematic approach based on resting-state EEG signals exploiting the attention weights to guide the detection of biomarkers in the context of AD detection and classification.

The remainder of this study is organized as follows: Section II describes the materials, the classification framework and the experimental details. In Section III, the findings of the interpretability approach are presented and discussed, along with the results obtained from the experiments conducted on the model architecture. Finally, conclusions and future works are discussed in Section IV.

II. Materials and Methods

A. EEG data acquisition and preprocessing

The dataset is composed of resting-state EEG recordings of 56 SCD and 45 MCI subjects collected at *IRCCS Don Carlo Gnocchi* in Florence, Italy, as part of the "PRedicting the EVolution of SubjectIvE Cognitive Decline to Alzheimer's Disease With machine learning (PREVIEW)" project [48]. Patients were classified as SCD according to the terminology proposed by the Subjective Cognitive Decline Initiative (SCD-I) Working Group [2], and as MCI according to the National Institute on Aging-Alzheimer's Association (NIA-AA) work-groups criteria for the diagnosis of MCI [49]. Fundamental demographic and clinical information of the patients are reported in Table I. More detailed information about the PREVIEW study design and participants can be found in [48].

TABLE I: Clinical-demographic characteristics of the study population reported as mean ± standard deviation. SCD: Subjective Cognitive Decline; MCI: Mild Cognitive Impairment; MMSE: Mini-Mental State Examination; TIB: Italian Brief Intelligence Test

	SCD	**MCI**
Age	66.26 ± 8.72	74.26 ± 8.20
Females *(%)*	78.3	54.3
Age onset	55.15 ± 8.04	62.09 ± 9.97
Years of Education	12.58 ± 3.47	10.18 ± 4.17
MMSE	27.48 ± 2.28	27.52 ± 2.13
TIB	107.22 ± 20.48	111.00 ± 6.01

All subjects were recruited in accordance with the Declaration of Helsinki and with the ethical standards of the Committee on Human Experimentation of Careggi University Hospital (Florence, Italy). The study was approved by the local Institutional Review Board ("Comitato Etico di Area Vasta Centro" reference 15691oss).

EBNeuro's GalNt system (EBNeuro, Florence, Italy) with 64 channels, including two EOG channels and ECG, and a sampling rate of 512 Hz was used to collect the signals. Electrodes were positioned according to the 10–10 international system. The acquisition protocol was designed to include both eyes-closed and eyes-open conditions, resulting in recordings lasting about 20 minutes. Only the eyes-closed portions of the signal were retained for analysis, since they constituted the largest part of the acquisition protocol.

Raw data were preprocessed using a standardized pipeline, the PREP pipeline [50], which was implemented in Matlab R2019b (The Mathworks, Natick, MA, USA) with the EEGLAB toolbox v.2021.0. After that, signals were additionally filtered with a 50 Hz notch filter to ensure the removal of line noise components.

Then, Independent Component Analysis (ICA) was performed and a semi-automatic method employing EEGLAB's ICLabel [51] was used to exclude components relative to noise and artifacts. Lastly, residual artifactual segments were manually removed.

EOG and ECG were excluded from the analyses. Of 61 recorded channels, 19 were retained, namely Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2, which we already proved to ensure sufficient information for this specific classification task; in fact, the choice of this subset of channels is not new in the context of EEG analysis for AD detection [12], [52]. In our previous work, we compared the classification results of our Transformer model when using both the reduced and complete set of channels and found that the overall performances of our model worsened when all channels were employed due to redundancy in the input signal [47].

Subsequently, the signals were bandpass filtered between 1 Hz and 30 Hz to include the EEG frequency bands of interest, i.e. delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz) and beta (12-30 Hz), and then normalized using the z-score normalization, according to Equation 1, where X^* is the normalized output signal, X is the reference signal, μ is its mean value and σ is the standard deviation for each subject. Mean and standard deviation have been computed per channel.

$$X^* = \frac{X - \mu}{\sigma} \tag{1}$$

For each subject, the signal was split into n epochs with the same time duration L and each epoch was labelled with the class of the corresponding subject. Since the EEG recordings had a different duration for each subject, and the preprocessing led to the removal of a variable number of segments of the input signal, n differed per subject.

B. Transformer architecture

As described in [47], the inputs to our Transformer model are EEG epochs of channels C and length L, while the output is the probability for each epoch to belong to one of the two classes, i.e. SCD or MCI.

Figure 1 shows the architecture of the Transformer. The model is composed of three main modules, namely patch embedding, positional encoding and the self-attention module, which is included in an encoder block. Lastly, it comprises a classification module constituted by a fully-connected layer with the *softmax* activation function.

1) Patch embedding: Each EEG epoch is split into patches and linearly projected by a single convolution operation, like in Song *et al.* [53]. Embedding dimension is set to *emb*, kernel size is set to (C, k), and the stride is the same as kernel width, i.e. $(1, k)$. This allows us to compress the input signal into $(L * f/k)$ single-channel patches, where f is the EEG sampling rate. Classification (CLS) tokens are prepended to the projected epochs and used to predict the epoch class after being updated by attention, as in the Vision Transformer [26].

2) Positional encoding: As in the original architecture, our Transformer model employs the position information of the input patches to make decisions. After patch embedding, the positions of all patches in the sequence are encoded in a vector, which is linearly added to the input sequence. The obtained array is then given as input to the first encoder block containing the attention module.

3) Self-attention module: The encoder block includes one module of attention, as well as a feed-forward module, normalization layer and dropout (Figure 2). In our model, the

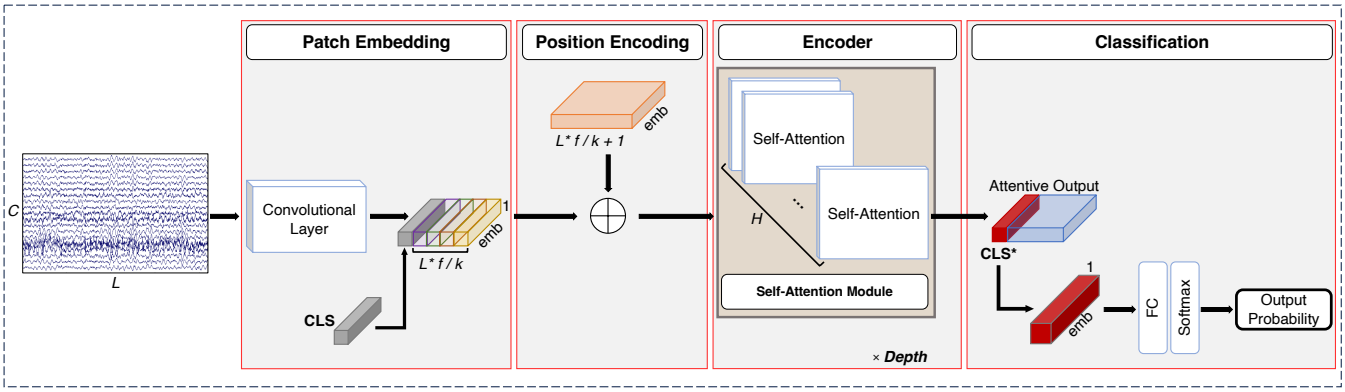


Fig. 1: Representation of the modules composing the proposed Transformer. C is the number of EEG channels, L is the length of the input epoch (in s), f is the EEG sampling rate (in Hz), k is the kernel size, emb is the embedding dimension and H is the number of attention heads. The classification token is denoted as CLS; the classification token updated after the Attention module is denoted as CLS*.

encoder block is replicated $Depth$ times. The module is based on the concept of self-attention, which computes an attention function of the input elements to retrieve the dependencies of each element with respect to the others [25]. Specifically, the input vector is firstly projected by a linear transformation into three different vectors: the query vector q, the key vector k and the value vector v, with dimensions $dq = dk = dv$. Vectors derived from the different input patches are then merged together into three different matrices, namely Q, K and V. Subsequently, the function computes scores between each pair of input patches, meaning that these values impact how much attention is given to other input patches when encoding the current one. These scores are normalized for gradient stability and then translated into probabilities using the *softmax* function, according to Equation 2. Finally, V is multiplied by scores to obtain attention values (Equation 3).

$$Scores = softmax(\frac{Q \cdot K^T}{\sqrt{dk}}) \quad (2)$$

$$Attention = Scores \cdot V \quad (3)$$

To investigate how the traditional self-attention and Multi-Head Attention (MHA) strategies could affect the classification performances, we varied the number of heads (H) per encoder. Depending on H, MHA generates different representations of the input into the q, k and v spaces. Then, the scaled dot-product is performed on these mapped queries, keys, and values simultaneously, which are then concatenated, as in Equation 4, where Q_i, K_i, V_i denote the query, key, and value obtained by linear transformation of divided token in the i-th head.

$$MHA(Q, K, V) = Concat(head_1, \ldots, head_H), \quad (4)$$
$$head_i = Attention(Q_i, K_i, V_i)$$

The attentive output, generated with self-attention or by MHA, is linearly projected. Finally, the output is then processed by the classification block.

C. Training setup and details

To avoid overevaluation of model performance, a test set was generated using 20 % of total subjects with a stratified random sampling approach. A stratified 5-fold cross-validation was employed on the remaining subjects, i.e. 43 SCD and 37 MCI, to train and validate the classification model. Using this technique, the data is divided subject-wisely into five equally-sized subsets, and the model is iteratively trained on four of these subsets and validated on the remaining one. Each subset is used as validation set exactly once.

Models were trained using Adam optimizer ($lr = 10e-4$, $\beta 1 = 0.9$, $\beta 2 = 0.999$, $eps = 1e-08$), which is the most employed method when training Transformer-based architectures [25] since it has faster convergence than non-adaptive algorithms such as SGD [54]. The value of lr was chosen by reducing it by a factor of 10 until finding an optimum in the validation set accuracy, starting from $10e-2$. Cross-Entropy was used as loss function. Batch size was set to 16 and the number of training iterations was equal to 250. In the proposed model, emb is set to 32 and $Depth$ is set to 2, resulting in 56194 trainable parameters.

It should be noted that the number of input signal epochs per subject varied depending on the duration of the single EEG recording. Thus, in the training phase, the number of EEG epochs in the majority class was reduced by randomly sampling to match the number of those in the minority class. During inference and performance evaluation, all epochs in the test set for both classes were considered.

Since our model directly classifies epochs either as SCD or MCI, a hard voting approach was then employed in order to predict a label for each subject. This means that each subject was classified based on the most frequent class predicted for

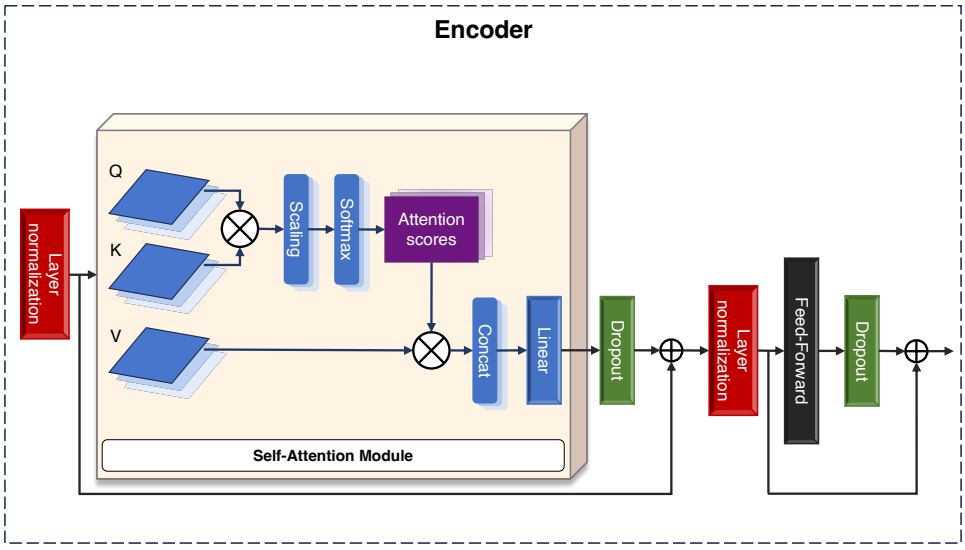


Fig. 2: Detailed architecture of the Transformer encoder block. Dot-product in the self-attention layer is represented by $\otimes$.

their corresponding epochs. If the frequency of predictions was equal for both classes, the subject was classified as MCI.

For each fold in the cross-validation, we calculated the Area Under Curve (AUC) of the Receiver-Operating Characteristics (ROC) curves on training and validation sets. We then selected the model at the training iteration which allowed to obtain the highest AUC score on the validation set and evaluated the performances on the test set. Results in Table II are reported in terms of mean ($\pm$ standard deviation) accuracy (Equation 5), sensitivity (Equation 6), specificity (Equation 7), precision (Equation 8) and F1-score (Equation 9) for epochs' classification on the test set. AUC values are also reported, given the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). All metrics were computed considering the MCI class as the positive class and the SCD as the negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$F1\text{-}Score = \frac{2 * TP}{2 * TP + FN + FP} \tag{9}$$

D. Interpretability workflow via Self-attention

To understand the behavior of the model for the investigated classification task, it is important to know which parts of the input the model pays more attention to. To this end, we extracted weights from each attention layer of the trained models in order to identify the signal patch that contributed the most to the classification of each EEG epoch. As described above, the classification of an EEG epoch is made upon the updated representation of the CLS token, i.e. CLS* (see Fig. 1). Thus, for each attention matrix, we considered the first row of values that correspond to the scaled dot-product attention of the CLS* token on the representations corresponding to the non-overlapping patches of the raw signal. This gives attention weights for each patch of the input epoch, helping evaluate their impact on the prediction.

It is worth noting that our Transformer, in its configuration with $H > 1$, uses a multi-depth and multi-head attention mechanism, which can produce different attention patterns that can be challenging to visualize [28]. We averaged the attention scores across attention heads in order to retain all information produced by the attention module. On the other hand, we extracted different results for the first and the second encoder blocks to evaluate the contribution of each attention layer separately.

For all subjects in the dataset, we identified n patches, corresponding to n epochs of the raw signals with the highest attention weights. This means that for each epoch of length L, a patch of signal with dimension k datapoints was obtained, where k is the dimension of the kernel in the convolutional layer employed for patch embedding. In order to uphold the assumption that the highest attention weights are representative of significant changes in the EEG activity between SCD and MCI groups, we collected and concatenated 1-second long windows of the signal centered on the previously identified patches, obtaining a new set of signals for each class. Epochs belonging to the same class were then merged in a single time series. To validate the significance of the results through a comparison with a reference, we also segmented the complete signals with windows of 1 second and, once again, concatenated epochs of the same class to obtain one SCD and one MCI time series. Statistical analysis was performed on EEG data using Matlab's Letswave 7 tool. We applied the non-parametric cluster-based permutation Student's t-test for unpaired data [55] to compare the signals' epochs of the

two groups, both for the attention-based set and the reference, which allowed us to handle the multiple comparisons problems. The cluster significant threshold was set to 0.05 and the number of permutations was set to 2000. Multi-sensor analysis was performed in order to consider both temporal and spatial adjacency of the samples.

Finally, to gain a physiological interpretation of the results, we performed time-frequency analysis by applying Continuous Wavelet Transform (CWT) to the EEG epochs and averaging the results across each group. Complex Morlet wavelet with bandwidth of 1 Hz and central frequency of 1.5 Hz was used as mother wavelet.

III. Results and Discussion

In this section, we extensively illustrate the results of interpretability analysis for visualizing the focus of the model on specific EEG patterns. Then, we provide results of parameter tuning tests for choosing the best model configuration and demonstrate the efficacy of the attention module through ablation studies.

A. Interpretability analysis

Following the approach proposed in section II-D, the interpretability analysis was performed on the model which obtained the highest values of mean accuracy and AUC on the test set, i.e. the Transformer configuration with $L = 30$ s, $k = 64$ and $H = 8$. This configuration achieves mean accuracy of 65.4 % (95 % CI [0.637 - 0.671], p-value [Accuracy $>$ No Information Rate] = 0.00026) on epochs' classification, and 65.7 % of accuracy for subjects' classification through hard voting.

As a first attempt to visualize the attention focus, we present heat maps of attention scores on the raw EEG signals for both SCD and MCI classes. Figure 3 shows two examples of 5-s-long windows extracted from the corresponding 30-s epochs of one correctly classified SCD (Fig. 3a) and one correctly classified MCI (Fig. 3b) subject of the test set. For clarity purposes, the normalized and non-normalized signals of one channel, namely T3, have been plotted for both samples. Attention scores are plotted over patches of k datapoints, with dark red indicating areas with higher focus, and light yellow indicating areas with lower focus.

To quantitatively evaluate the contribution of the attention scores on the final classification outcome, we show results of the nonparametric cluster-based permutation Student's t-test and the corresponding time-frequency analysis with the aim of highlighting differences between the two groups. We considered channels with clustered p-value < 0.01 to be significant.

When comparing epochs of 1 s centered on patches with the highest attention, the most significant differences between the SCD and MCI signals are, indeed, located in the time interval that corresponds to those patches, i.e. from 437 ms to 562 ms since the epoch start. For instance, when considering the results of the first Transformer attention block (Fig. 4a) it can be noted that most statistically significant inter-group differences can be found in the central part of the time window, as shown by the corresponding scalp topographies representing clustered p-values. The most significant changes occur on clusters including the following channels: Fp1, Fp2, F3, F7, Fz, F4, F8, C3, Cz, C4, P3, P4, Pz, T5, T6, O1 and O2.

This evidence is strengthened by the results obtained on the second attention block (Fig. 4b). In this case, almost all statistically significant differences correspond to the highest attention scores which are located in the middle of the considered time window. Scalp topographies of clustered p-values show that the significant clusters include the Fp1, Fp2, F7, F3, F4, F8, C4, Cz, T3, P3, Pz and P4 channels.

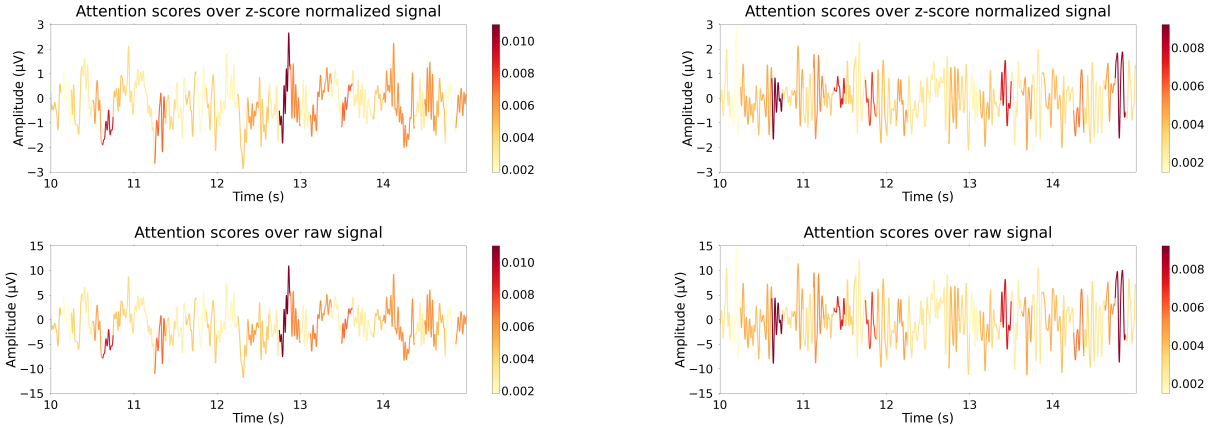
The clusters found in both cases indicate brain regions that are congruent with scientific evidence from cross-sectional and longitudinal studies on the cognitive spectrum of AD. As reported in [56], the left posterior parietal and left and right temporo-occipital regions (which are represented by P3, P4, T6 and O2 electrodes) were consistently described as the most discriminative brain areas between controls, MCI and AD, while the left posterior temporal region and fronto-central midline (corresponding to T5, Fz, Cz and Pz channels) as important in the prediction of clinical progression in patients with SCD.

On the other hand, statistical analysis performed on the reference dataset, i.e. considering all epochs of 1 s extracted from the input signal, regardless of weights attributed by attention, found no significant channels at any time instant ($p > 0.01$). This result confirms that, although mean classification accuracy on the test set is not optimal, the Transformer is able to capture global temporal dependencies of the signal that allow the classification of each epoch with good discrimination capability.

However, differently from other studies that applied an interpretability approach based on attention scores to EEG signals in the context of sleep stage classification [33], [34] or motor imagery paradigms [57], [58], these features are not easily detectable and do not provide enough explanations in the time domain.

Hence, on the basis of the findings derived from the statistical analysis, we report scalp topographies of the average power CWT for SCD and MCI subject groups based on the results of the first Transformer block ($Depth = 1$). In particular, Fig. 6 shows CWT maps averaged across the whole 1-s interval (first and third row) and the interval of interest (second and fourth row) for delta (Fig.6a) and alpha (Fig.6b) frequency bands, respectively. Of notice, differences between the groups are once again more evident when considering the time interval corresponding to the highest attention scores, rather than the entire time window. The maps confirm that subjects belonging to the MCI group show a lower power in high frequencies and higher power in low frequencies in accordance with state-of-the-art results in the context of AD characterization from rsEEG [15], [56], [59]. In addition, these explanations keep with expectations of our previous work [47].

Additionally, we compared these maps with the ones obtained on the reference dataset, and found that in the latter the differences between the groups do not correspond to specific time intervals, in accordance with the results of the aforementioned statistical analysis.



(a) SCD (b) MCI

Fig. 3: Sample plots of two 5-s long EEG epochs with relative attention scores for one SCD (a) and one MCI (b) subject of the test set. Both normalized and non-normalized signals are shown.

To further understand the role of the multi-head attention mechanism, we repeated the analyses for the baseline model with single-head self-attention, which achieves a mean accuracy of 59.5 % (95 % CI [0.577 - 0.612], p-value [Accuracy $>$ No Information Rate] $>$ 0.05) on epochs' classification and 61.9 % on patients' classification. As expected, and as found by [57], the attention activation of a single head is similar to the one obtained by averaging multiple heads, with significantly different patches (p-value $<$ 0.01) between SCD and MCI groups corresponding to the highest attention scores, but resulting in more sparse and less consistent channel clusters, particularly when considering the results obtained on the first Transformer block. For $Depth = 1$, significant clusters include Fp2, F4, F7, F8, Cz, C3, C4, P3, Pz, T3, T4, T5, T6, and O2 (Fig. 5a). For $Depth = 2$, significant channels are Fp1, Fp2, F3, F7, F8, Fz, Cz, C3, C4, Pz, P4, T3, T4, T5, T6, O1 and O2 (Fig. 5b). This is explained considering that the baseline model, for the same model depth, has lower performances which do not reach the statistical significance in terms of accuracy; such a result is in line with [60], who report that single-head attention necessitates deeper models to prove more effective than MHA, but increasing the model complexity. Thus, the attention focus is less indicative of discriminative EEG features. Consequently, the spectral analysis obtained with CWT shows similar outcomes, with changes in activation between groups mostly gathered in the central part of the window, but being less enhanced, especially in the lower frequencies, for both $Depths$.

B. Hyperparamter tuning

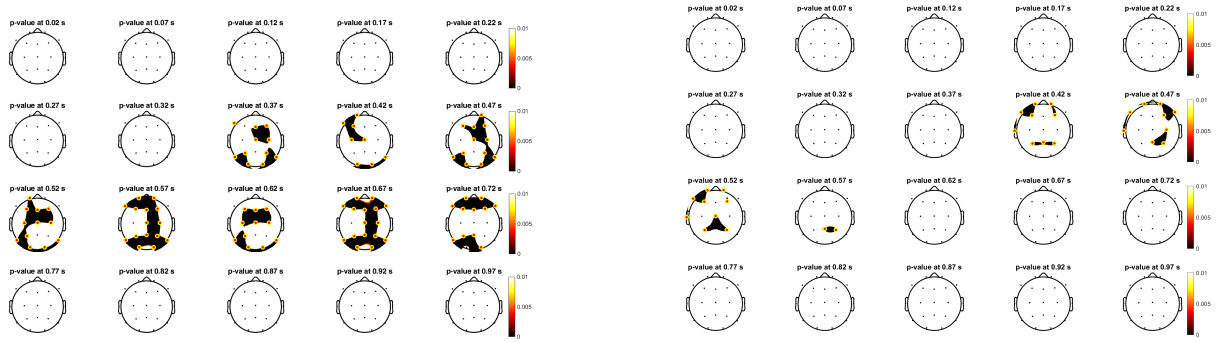
We conducted experiments to identify the best model's parameters to achieve optimal classification performances. We varied two parameters that influence the construction of the input, namely the time duration of input EEG epochs and the design of the convolutional kernel, and also investigated the influence of the number H of attention heads in the attention layer, known to impact feature learning.

In particular, three different lengths of input epochs (10, 30 and 60 seconds) and five different kernel sizes (16, 32, 64, 128 and 512) were tested and compared to identify the combination with the highest classification performance. Table II reports mean results for all the considered metrics on epochs' and patients' classification. The highest levels of mean accuracy are reached with a kernel size of 64, with values of 65.4 % and 63.0 % for epochs of 30 s and 60 s respectively, and a kernel size of 32 on epochs of 10 s with a value of 63.4 %.

By contrast, the lowest results are yielded when using kernel sizes of 512 (52.4 % on epochs of 60 s) and 16 (54.3 % on epochs of 30 s). Although the differences are not significant ($p > 0.05$), in accordance with Song *et al.* [57], we found that large kernel sizes tend to flatten temporal features and reduce the learning of global dependencies, while small kernels produce tokens that do not contain enough information for the model to perceive local changes in the signal. On the other hand, the length of the input EEG signal seems to impact the performances of our model to a lesser extent. However, as a general remark, using very long epoch lengths (i.e. 60 s) results in a smaller dataset size which increases the risk of lowering the performance of the classification model.

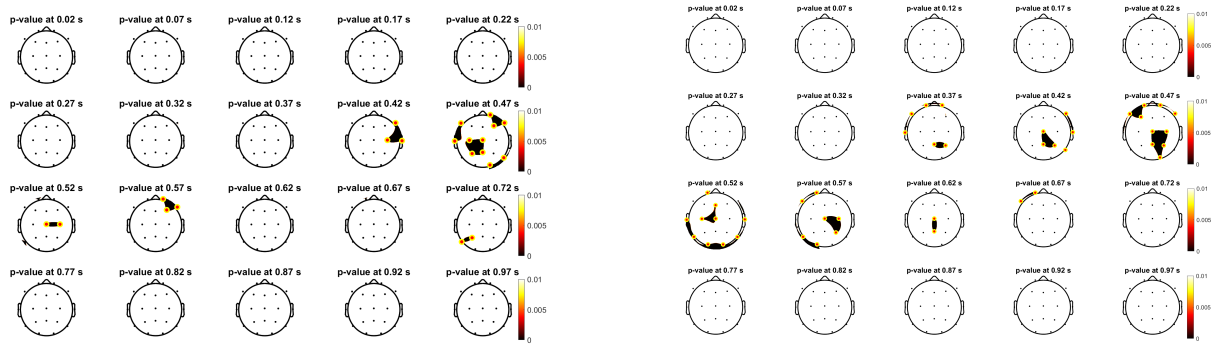
We also compared the impact of choosing different numbers of heads for the attention layer, performing experiments by varying H in [1, 2, 4, 8, 16, 32]. Since each head projects the input onto a subspace of dimension $dim = \frac{emb}{H}$ to compute the context [25], the values of H were chosen based on the embedding dimension.

The results reported in Fig. 7 show that the effects on the performance of the model follow no evident trend ($p > 0.05$), but the highest accuracy of 65.4 % is obtained with $H = 8$, compared to 59.5 % with $H = 1$, 62.8 % with $H = 2$, 57.8 % with $H = 4$, 56.9 % with $H = 16$ and 59.3 % with $H = 32$. Also, as shown by the error bars in the same figure, setting the number of heads to 8 allowed to obtain the smallest 95 % confidence interval. Conversely, the highest confidence intervals were derived from configurations with 1 and 32



(a) (b)

Fig. 4: The results of the cluster permutation Student's t-test for multi-head attention model, $Depth = 1$ (a) and $Depth = 2$ (b). Clustered p-values over time are plotted on scalp maps at 50 ms intervals. Significant channels are yellow-circled and highlighted.



(a) (b)

Fig. 5: The results of the cluster permutation Student's t-test for single-head self-attention model, $Depth = 1$ (a) and $Depth = 2$ (b). Clustered p-values over time are plotted on scalp maps at 50 ms intervals. Significant channels are yellow-circled and highlighted.

heads. This suggests that while employing a greater number of heads enables the model to identify more meaningful features, a progressive increase in the number of heads results in shorter feature lengths within each head. This, in turn, contributes to a marginal decrease in performance. This result confirms previous evidence from another study [61].

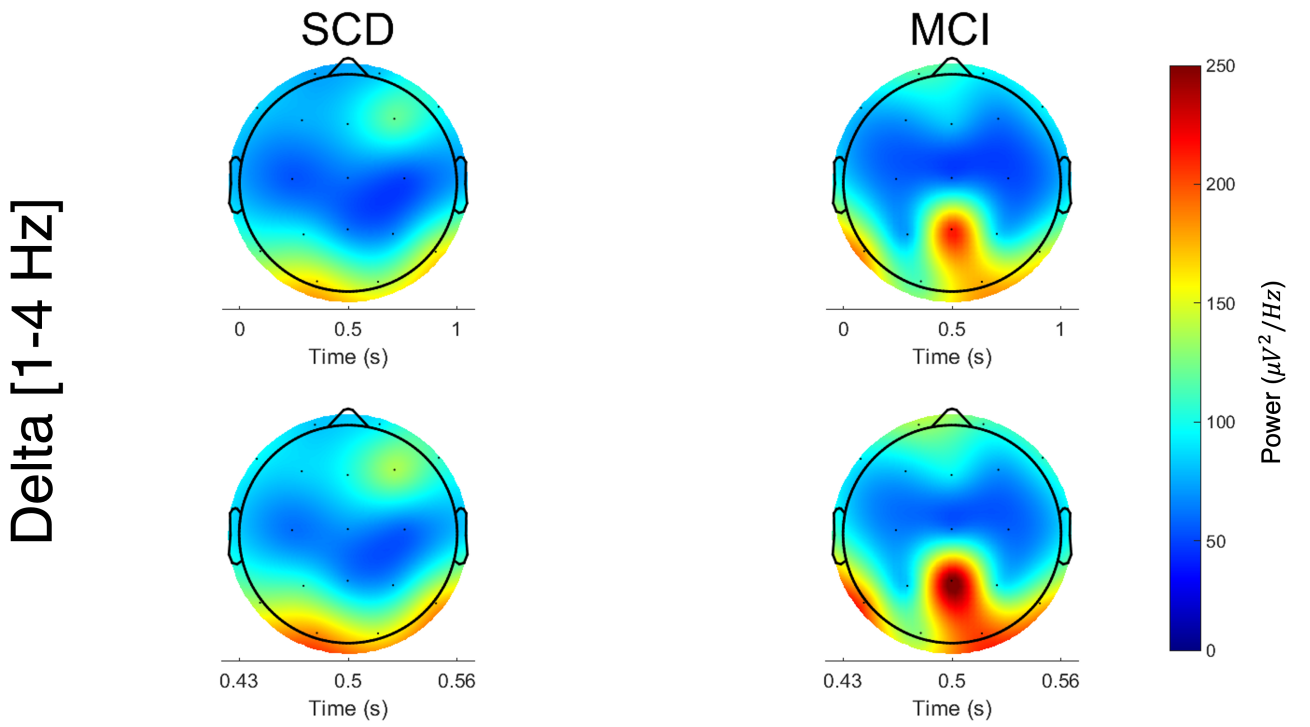
C. Ablation Study

In this section, we systematically analyze the importance of two key components of our model, namely the attention-based Transformer encoder module and the positional encoding module. An ablation study was conducted by firstly removing the Transformer encoder, i.e. the classification was performed on the input signal after convolution without applying any attention strategy. Then, we reintroduced the Transformer encoder module and dismissed the positional encoding, so that the model had no information about the position of each patch in the input sequence when performing classification. Lastly, we removed both the Transformer and the positional encoding blocks. In the study, we included results for both MHA and single-head self-attention models.

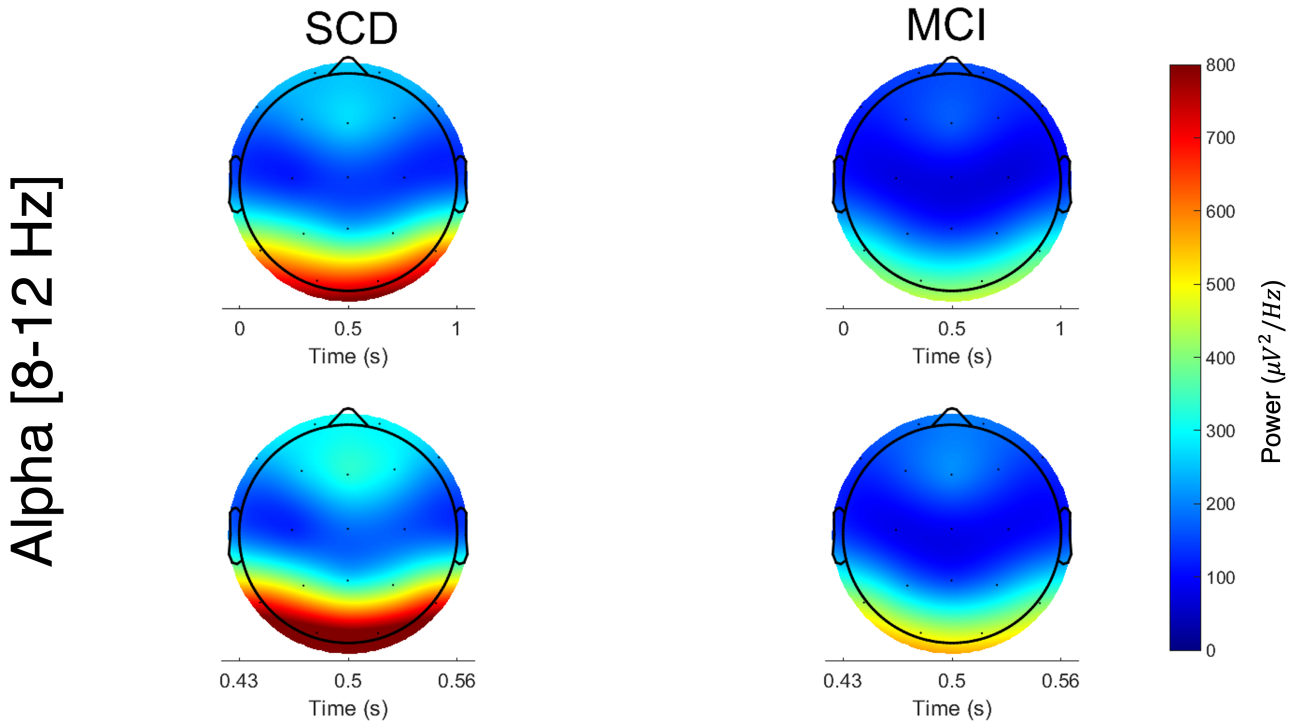
As depicted in Figure 8, and as already shown in Fig. 7, for the same input configuration, the model employing multiple heads has overall better performances than the model employing the traditional self-attention layers, which does not reach statistical significance in classification accuracy on the test set and shows high variability over the folds.

Nevertheless, the effectiveness of using an attention mechanism is confirmed by the results obtained when the Transformer block is removed, in which the mean accuracy on the test set drops significantly in the epochs' classification, decreasing by 16 % ($p = 0.004$) for the MHA configuration and by 10.5 % ($p > 0.05$) for the single-head self-attention configuration including it. Also, in patients' classification it reduces significantly by 19 % ($p = 0.009$) in the first case and by 15.2 % ($p > 0.05$) in the latter.

The removal of the positional encoding has a different impact on the two models. For the model employing MHA, the mean accuracy over the folds decreases by 1.5% ($p > 0.05$). Although the difference is not significant, these results



(a)



(b)

Fig. 6: Scalp topographies of Average Continuous Wavelet Transform of EEG signals segmented based on attention scores of the first Transformer block for SCD and MCI groups. (a) Average CWT in delta band (1-4 Hz) across the whole second interval (first row) and the interval of interest (second row). (b) Average CWT in alpha band (8-12 Hz) across the whole second interval (first row) and the interval of interest (second row).

TABLE II: Classification results on the epochs' test set for different input configurations, expressed as mean ± standard deviation.

Epoch length	Kernel	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
10	16	0.61 ± 0.01	0.59 ± 0.12	0.62 ± 0.09	0.48 ± 0.01	0.52 ± 0.04	0.60 ± 0.02
	32	0.63 ± 0.03	0.63 ± 0.13	0.64 ± 0.03	0.51 ± 0.03	0.56 ± 0.07	0.63 ± 0.05
	64	0.62 ± 0.06	0.54 ± 0.06	0.66 ± 0.07	0.50 ± 0.08	0.52 ± 0.06	0.60 ± 0.06
	128	0.47 ± 0.08	0.54 ± 0.15	0.42 ± 0.19	0.36 ± 0.05	0.43 ± 0.07	0.52 ± 0.06
	512	0.57 ± 0.03	0.53 ± 0.04	0.59 ± 0.02	0.44 ± 0.03	0.48 ± 0.03	0.59 ± 0.04
30	16	0.56 ± 0.06	0.57 ± 0.08	0.55 ± 0.05	0.44 ± 0.05	0.49 ± 0.06	0.56 ± 0.06
	32	0.62 ± 0.05	0.61 ± 0.11	0.62 ± 0.09	0.50 ± 0.06	0.54 ± 0.06	0.61 ± 0.05
	64	$\mathbf{0.65 \pm 0.05}$	0.58 ± 0.11	0.70 ± 0.06	0.54 ± 0.06	0.56 ± 0.07	$\mathbf{0.64 \pm 0.06}$
	128	0.62 ± 0.12	0.46 ± 0.15	$\mathbf{0.72 \pm 0.26}$	$\mathbf{0.57 \pm 0.14}$	0.48 ± 0.08	0.62 ± 0.12
	512	0.55 ± 0.03	0.57 ± 0.11	0.54 ± 0.09	0.43 ± 0.03	0.48 ± 0.05	0.58 ± 0.05
60	16	0.60 ± 0.08	0.64 ± 0.16	0.58 ± 0.12	0.48 ± 0.09	0.55 ± 0.11	0.61 ± 0.09
	32	0.59 ± 0.12	0.56 ± 0.14	0.62 ± 0.19	0.49 ± 0.12	0.51 ± 0.10	0.59 ± 0.11
	64	0.63 ± 0.09	$\mathbf{0.64 \pm 0.12}$	0.62 ± 0.09	0.51 ± 0.10	$\mathbf{0.57 \pm 0.11}$	0.63 ± 0.10
	128	0.59 ± 0.08	0.57 ± 0.20	0.60 ± 0.13	0.46 ± 0.07	0.50 ± 0.13	0.60 ± 0.11
	512	0.53 ± 0.05	0.49 ± 0.13	0.55 ± 0.10	0.40 ± 0.05	0.43 ± 0.07	0.52 ± 0.06

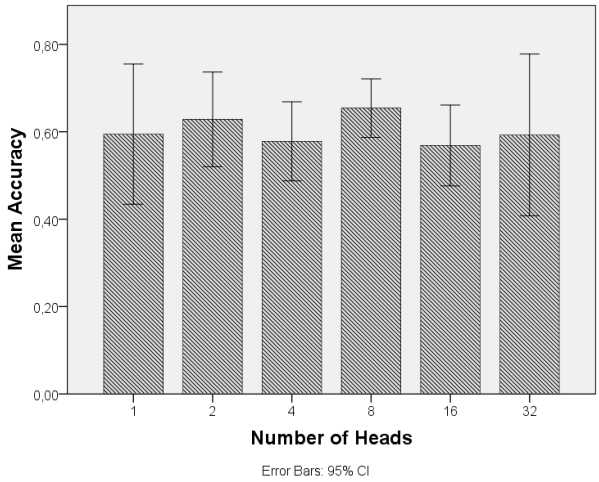


Fig. 7: The impact of different numbers of attention heads on the mean accuracy over folds for epoch-wise classification on the test set.

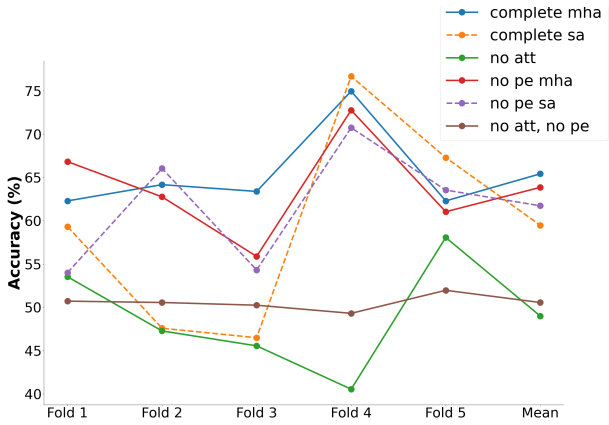


Fig. 8: The results of ablation study for epoch-wise classification on the test set. Accuracy values are plotted for single folds and as mean values over folds. In the legend, att is the attention module, pe is the positional encoding, mha is the multi-head attention and sa is the traditional self-attention with one head.

suggest that this model makes use of positional encoding in an informative way, but is still able to compensate for it with the attention module. Additionally, this consideration is supported by the results of the last ablation test, in which both the Transformer and the positional encoding modules are removed. In this case, the mean performances of the model are slightly higher than the case in which only the Transformer is removed, by 1.5 % epoch-wise ($p = 0.02$), proving that the positional encoding module is useful when combined with multi-head attention, but has a negative impact on the results when added to a convolutional-based model. In fact, positional information could be inherently learned by a convolutional layer with a sufficient receptive field size [62] and thus the information provided by the positional encoding in this case could produce redundancy. On the other hand, the ablation of the positional encoding module in the single-head self-attention model also seems to impact positively on the classification performances, by increasing accuracy of 2.2 %, but not significantly ($p > 0.05$), which further proves that the attention module is capable of learning positional information by itself [63]. However, this result needs further understanding [64].

D. General remarks

The complexity of EEG signals poses a challenge in the identification of biomarkers that can accurately discriminate between SCD and MCI conditions. Here, we demonstrate that MHA can be used in an end-to-end Transformer model to automatically locate time windows of the resting-state EEG that may account for significant changes in the brain activity. The interpretability analysis shows a higher global efficacy of MHA compared to traditional self-attention approaches. Indeed, although we previously found that the MHA-based Transformer does not outperform other investigated DL methods for the specific task, it allows to highlight significant differences between the groups which could not be explained otherwise. In addition, the ablation study we conducted confirmed the effectiveness of introducing Transformer blocks in our model, in particular when coupled with the encoding of

the positions of patches in the input.

Based on the present results, we speculate that our framework could serve not only to enhance the interpretability of a black-box model which achieves state-of-the-art classification performances, thus addressing the problem of the trade-off between accuracy and trustworthiness [65], but also as a guide for experts to facilitate the extraction of rsEEG markers of cognitive decay. A recent work employed the attention mechanism to design an EEG channel interpolation algorithm [66]. Similarly, our method could be exploited also in different applications to select relevant domain-specific information by taking into account short and long temporal dependencies of the signal.

Future work will be aimed at addressing the limitations of the proposed approach. First, the workflow for interpreting the Transformer could be extended by extracting physiological EEG biomarkers guided by attention scores. The validity of our method should also be assessed on larger EEG datasets, and alternative methods for attention visualization aimed at increasing the interpretability of the model should be analyzed. Lastly, variants of the proposed Transformer model should be further explored to improve classification performances.

IV. Conclusion

The functional characterization of SCD and MCI conditions using non-invasive rsEEG signals constitutes a fundamental step to support an early diagnosis of cognitive decline. Building upon our previous results, in this work we constructed an interpretability framework leveraging the mechanism of attention and found that the focus of our Transformer model, which corresponds to the highest attention scores on specific signal patches, is representative of hallmark EEG patterns that could allow to discriminate SCD from MCI.

The source code is publicly available at `https://github.com/LabInfInd/SCD_MCI_Transformer`.

References

[1] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and M. Prina, "World alzheimer report 2015. the global impact of dementia: an analysis of prevalence, incidence, cost and trends." Alzheimer's disease international, Tech. Rep., 2015.

[2] F. Jessen, R. E. Amariglio, M. van Boxtel, M. Breteler, M. Ceccaldi, G. Chételat, B. Dubois, C. Dufouil, K. A. Ellis, W. M. van der Flier, L. Glodzik, A. C. van Harten, M. J. de Leon, P. McHugh, M. M. Mielke, J. L. Molinuevo, L. Mosconi, R. S. Osorio, A. Perrotin, R. C. Petersen, L. A. Rabin, L. Rami, B. Reisberg, D. M. Rentz, P. S. Sachdev, V. de la Sayette, A. J. Saykin, P. Scheltens, M. B. Shulman, M. J. Slavin, R. A. Sperling, R. Stewart, O. Uspenskaya, B. Vellas, P. J. Visser, M. Wagner, and S. C. D. I. S.-I. W. Group, "A conceptual framework for research on subjective cognitive decline in preclinical alzheimer's disease," *Alzheimer's & dementia*, vol. 10, no. 6, pp. 844–852, 2014.

[3] O. L. Lopez, "Mild cognitive impairment," *Continuum: Lifelong Learning in Neurology*, vol. 19, no. 2 Dementia, p. 411, 2013.

[4] A. Li, L. Yue, S. Xiao, and M. Liu, "Cognitive function assessment and prediction for subjective cognitive decline and mild cognitive impairment," *Brain imaging and behavior*, vol. 16, no. 2, pp. 645–658, 2022.

[5] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack Jr., J. Kaye, T. J. Montine, D. C. Park, E. M. Reiman, C. C. Rowe, E. Siemers, Y. Stern, K. Yaffe, M. C. Carrillo, B. Thies, M. Morrison-Bogorad, M. V. Wagster, and C. H. Phelps, "Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[6] J. Cummings, Y. Zhou, G. Lee, K. Zhong, J. Fonseca, and F. Cheng, "Alzheimer's disease drug development pipeline: 2023," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 9, no. 2, p. e12385, 2023.

[7] F. L. Guest, H. Rahmoune, and P. C. Guest, "Early diagnosis and targeted treatment strategy for improved therapeutic outcomes in alzheimer's disease," *Reviews on New Drug Targets in Age-Related Disorders*, pp. 175–191, 2020.

[8] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," *Frontiers in neuroscience*, vol. 14, p. 779, 2020.

[9] A. I. Triggiani, V. Bevilacqua, A. Brunetti, R. Lizio, G. Tattoli, F. Cassano, A. Soricelli, R. Ferri, F. Nobili, L. Gesualdo, M. R. Barulli, R. Tortelli, V. Cardinali, A. Giannini, P. Spagnolo, S. Armenise, F. Stocchi, G. Buenza, G. Scianatico, G. Logroscino, G. Lacidogna, F. Orzi, C. Buttinelli, F. Giubilei, C. Del Percio, G. B. Frisoni, and C. Babiloni, "Classification of healthy subjects and alzheimer's disease patients with dementia from cortical sources of resting state eeg rhythms: a study using artificial neural networks," *Frontiers in neuroscience*, vol. 10, p. 604, 2017.

[10] S. Gunes, Y. Aizawa, T. Sugashi, M. Sugimoto, and P. P. Rodrigues, "Biomarkers for alzheimer's disease in the current state: A narrative review," *International Journal of Molecular Sciences*, vol. 23, no. 9, p. 4962, 2022.

[11] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[12] C. Babiloni, S. Lopez, C. Del Percio, G. Noce, M. T. Pascarelli, R. Lizio, S. J. Teipel, G. González-Escamilla, H. Bakardjian, N. George, E. Cavedo, S. Lista, P. A. Chiesa, A. Vergallo, P. Lemercier, G. Spinelli, M. J. Grothe, M.-C. Potier, F. Stocchi, R. Ferri, M.-O. Habert, F. J. Fraga, B. Dubois, and H. Hampel, "Resting-state posterior alpha rhythms are abnormal in subjective memory complaint seniors with preclinical alzheimer's neuropathology and high education level: the insight-pread study," *Neurobiology of Aging*, vol. 90, pp. 43–59, 2020.

[13] C. Babiloni, X. Arakaki, H. Azami, K. Bennys, K. Blinowska, L. Bonanni, A. Bujan, M. C. Carrillo, A. Cichocki, J. de Frutos-Lucas, C. Del Percio, B. Dubois, R. Edelmayer, G. Egan, S. Epelbaum, J. Escudero, A. Evans, F. Farina, K. Fargo, A. Fernández, R. Ferri, G. Frisoni, H. Hampel, M. G. Harrington, V. Jelic, J. Jeong, Y. Jiang, M. Kaminski, V. Kavcic, K. Kilborn, S. Kumar, A. Lam, L. Lim, R. Lizio, D. Lopez, S. Lopez, B. Lucey, F. Maestú, W. J. McGeown, I. McKeith, D. V. Moretti, F. Nobili, G. Noce, J. Olichney, M. Onofrj, R. Osorio, M. Parra-Rodriguez, T. Rajji, P. Ritter, A. Soricelli, F. Stocchi, I. Tarnanas, J. P. Taylor, S. Teipel, F. Tucci, M. Valdes-Sosa, P. Valdes-Sosa, M. Weiergräber, G. Yener, and B. Guntekin, "Measures of resting state eeg rhythms for clinical trials in alzheimer's disease: Recommendations of an expert panel," *Alzheimer's & Dementia*, vol. 17, no. 9, pp. 1528–1553, 2021.

[14] P. Rossini, R. Di Iorio, F. Vecchio, M. Anfossi, C. Babiloni, M. Bozzali, A. Bruni, S. Cappa, J. Escudero, F. Fraga, P. Giannakopoulos, B. Guntekin, G. Logroscino, C. Marra, F. Miraglia, F. Panza, F. Tecchio, A. Pascual-Leone, and B. Dubois, "Early diagnosis of alzheimer's disease: the role of biomarkers including advanced eeg signal analysis. report from the ifcn-sponsored panel of experts," *Clinical Neurophysiology*, vol. 131, no. 6, pp. 1287–1310, 2020.

[15] M. Lassi, C. Fabbiani, S. Mazzeo, R. Burali, A. A. Vergani, G. Giacomucci, V. Moschini, C. Morinelli, F. Emiliani, M. Scarpino *et al.*, "Degradation of eeg microstates patterns in subjective cognitive decline and mild cognitive impairment: Early biomarkers along the alzheimer's disease continuum?" *NeuroImage: Clinical*, vol. 38, p. 103407, 2023.

[16] C. Ieracitano, N. Mammone, A. Bramanti, A. Hussain, and F. C. Morabito, "A convolutional neural network approach for classification of dementia stages based on 2d-spectral representation of eeg recordings," *Neurocomputing*, vol. 323, pp. 96–107, 2019.

[17] F. C. Morabito, M. Campolo, C. Ieracitano, J. M. Ebadi, L. Bonanno, A. Bramanti, S. Desalvo, N. Mammone, and P. Bramanti, "Deep convolutional neural networks for classification of mild cognitive impaired and alzheimer's disease patients from scalp eeg recordings," in *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*. IEEE, 2016, pp. 1–6.

[18] J. Park, S. Jang, J. Gwak, B. C. Kim, J. J. Lee, K. Y. Choi, K. H. Lee, S. C. Jun, G.-J. Jang, and S. Ahn, "Individualized diagnosis of preclinical alzheimer's disease using deep neural networks," *Expert Systems with Applications*, vol. 210, p. 118511, 2022.

[19] I. Lazarou, K. Georgiadis, S. Nikolopoulos, V. P. Oikonomou, A. Tsolaki, I. Kompatsiaris, M. Tsolaki, and D. Kugiumtzis, "A novel

connectome-based electrophysiological study of subjective cognitive decline related to alzheimer's disease by using resting-state high-density eeg egi ges 300," *Brain Sciences*, vol. 10, no. 6, p. 392, 2020.

[20] V. Jahmunah, E. Y. K. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, "Explainable detection of myocardial infarction using deep learning models with grad-cam technique on ecg signals," *Computers in Biology and Medicine*, vol. 146, p. 105550, 2022.

[21] Y. Li, H. Yang, J. Li, D. Chen, and M. Du, "Eeg-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-cam," *Neurocomputing*, vol. 415, pp. 225–233, 2020.

[22] A. Fernandez-Quilez, "Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability," *AI and Ethics*, vol. 3, no. 1, pp. 257–265, 2023.

[23] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3197–3234, 2022.

[24] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[28] J. Vig, "A multiscale visualization of attention in the transformer model," *arXiv preprint arXiv:1906.05714*, 2019.

[29] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 782–791.

[30] J. Vig and Y. Belinkov, "Analyzing the structure of attention in a transformer language model," *arXiv preprint arXiv:1906.04284*, 2019.

[31] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.

[32] L. Schwenke and M. Atzmueller, "Show me what you're looking for: visualizing abstracted transformer attention for enhancing their local interpretability on time series data," in *The International FLAIRS Conference Proceedings*, vol. 34, 2021.

[33] J. Pradeepkumar, M. Anandakumar, V. Kugathasan, D. Suntharalingham, S. L. Kappel, A. C. De Silva, and C. U. Edussooriya, "Towards interpretable sleep stage classification using cross-modal transformers," *arXiv preprint arXiv:2208.06991*, 2022.

[34] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2456–2467, 2022.

[35] Z. Chen, Z. Yang, L. Zhu, W. Chen, T. Tamura, N. Ono, M. Altaf-Ul-Amin, S. Kanaya, and M. Huang, "Automated sleep staging via parallel frequency-cut attention," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[36] J.-Y. Guo, Q. Cai, J.-P. An, P.-Y. Chen, C. Ma, J.-H. Wan, and Z.-K. Gao, "A transformer based neural network for emotion recognition and visualizations of crucial eeg channels," *Physica A: Statistical Mechanics and its Applications*, vol. 603, p. 127700, 2022.

[37] A. Baghdadi, R. Fourati, Y. Aribi, S. Daoud, M. Dammak, C. Mhiri, H. Chabchoub, P. Siarry, and A. Alimi, "A channel-wise attention-based representation learning method for epileptic seizure detection and type classification," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–16, 2023.

[38] S. Sarraf, A. Sarraf, D. D. DeSouza, J. A. Anderson, M. Kabia, and A. D. N. Initiative, "Ovitad: Optimized vision transformer to predict various stages of alzheimer's disease using resting-state fmri and structural mri data," *Brain Sciences*, vol. 13, no. 2, p. 260, 2023.

[39] M. Odusami, R. Maskeliūnas, and R. Damaševičius, "Pixel-level fusion approach with vision transformer for early detection of alzheimer's disease," *Electronics*, vol. 12, no. 5, p. 1218, 2023.

[40] Z. Hu, Z. Wang, Y. Jin, and W. Hou, "Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction," *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107291, 2023.

[41] G. M. Hoang, U.-H. Kim, and J. G. Kim, "Vision transformers for the prediction of mild cognitive impairment to alzheimer's disease progression using mid-sagittal smri," *Frontiers in Aging Neuroscience*, vol. 15, p. 1102869, 2023.

[42] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, and N. Jacobs, "Advit: Vision transformer on multi-modality pet images for alzheimer disease diagnosis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–4.

[43] H. Shin, S. Jeon, Y. Seol, S. Kim, and D. Kang, "Vision transformer approach for classification of alzheimer's disease using 18f-florbetaben brain images," *Applied Sciences*, vol. 13, no. 6, p. 3453, 2023.

[44] D. Kumar Ravikanti and S. Saravanan, "Eegalzheimer'snet: Development of transformer-based attention long short term memory network for detecting alzheimer disease using eeg signal," *Biomedical Signal Processing and Control*, vol. 86, p. 105318, 2023.

[45] J. Wei, W. Xiao, S. Zhang, and P. Wang, "Mild cognitive impairment classification convolutional neural network with attention mechanism," in *2020 IEEE 16th International Conference on Control & Automation (ICCA)*. IEEE, 2020, pp. 1074–1078.

[46] A. Miltiadous, E. Gionanidis, K. D. Tzimourta, N. Giannakeas, and A. T. Tzallas, "Dice-net: a novel convolution-transformer architecture for alzheimer detection in eeg signals," *IEEE Access*, 2023.

[47] E. Sibilano, A. Brunetti, D. Buongiorno, M. Lassi, A. Grippo, V. Bessi, S. Micera, A. Mazzoni, and V. Bevilacqua, "An attention-based deep learning approach for the classification of subjective cognitive decline and mild cognitive impairment using resting-state eeg," *Journal of Neural Engineering*, vol. 20, no. 1, p. 016048, 2023.

[48] S. Mazzeo, M. Lassi, S. Padiglioni, A. A. Vergani, V. Moschini, M. Scarpino, G. Giacomucci, R. Burali, C. Morinelli, C. Fabbiani, G. Galdo, L. G. Amato, S. Bagnoli, F. Emiliani, A. Ingannato, B. Nacmias, S. Sorbi, A. Grippo, A. Mazzoni, and V. Bessi, "Predicting the evolution of subjective cognitive decline to alzheimer's disease with machine learning: the preview study protocol," *BMC Neurology*, vol. 23, no. 1, p. 300, 2023.

[49] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder, M. C. Carrillo, B. Thies, and C. H. Phelps, "The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 270–279, 2011.

[50] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The prep pipeline: standardized preprocessing for large-scale eeg analysis," *Frontiers in neuroinformatics*, vol. 9, p. 16, 2015.

[51] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, 2019.

[52] R. Ferri, C. Babiloni, V. Karami, A. I. Triggiani, F. Carducci, G. Noce, R. Lizio, M. T. Pascarelli, A. Soricelli, F. Amenta, A. Bozzao, A. Romano, F. Giubilei, C. Del Percio, F. Stocchi, G. B. Frisoni, F. Nobili, L. Patanè, and P. Arena, "Stacked autoencoders as new models for an accurate alzheimer's disease classification support using resting-state eeg and mri measurements," *Clinical Neurophysiology*, vol. 132, no. 1, pp. 232–245, 2021.

[53] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial-temporal feature learning for eeg decoding," *arXiv preprint arXiv:2106.11170*, 2021.

[54] Y. Pan and Y. Li, "Toward understanding why adam converges faster than sgd for transformers," *arXiv preprint arXiv:2306.00204*, 2023.

[55] E. Maris and R. Oostenveld, "Nonparametric statistical testing of eeg- and meg-data," *Journal of neuroscience methods*, vol. 164, no. 1, pp. 177–190, 2007.

[56] A. A. Gouw, A. M. Alsema, B. M. Tijms, A. Borta, P. Scheltens, C. J. Stam, and W. M. van der Flier, "Eeg spectral analysis as a putative early prognostic biomarker in nondemented, amyloid positive subjects," *Neurobiology of Aging*, vol. 57, pp. 133–142, 2017.

[57] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[58] C.-C. Fan, H. Yang, Z.-G. Hou, Z.-L. Ni, S. Chen, and Z. Fang, "Bilinear neural network with 3-d attention for brain decoding of motor imagery movements from the human eeg," *Cognitive Neurodynamics*, vol. 15, pp. 181–189, 2021.

[59] R. Cassani, M. Estarellas, R. San-Martin, F. J. Fraga, and T. H. Falk,

"Systematic review on resting-state eeg for alzheimer's disease diagnosis and progression assessment," *Disease markers*, vol. 2018, 2018.

[60] L. Liu, J. Liu, and J. Han, "Multi-head or single-head? an empirical comparison for transformer training," *arXiv preprint arXiv:2106.09650*, 2021.

[61] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[62] J. Park and W. Sung, "Effect of adding positional information on convolutional neural networks for end-to-end speech recognition," in *INTERSPEECH*, 2020, pp. 46–50.

[63] A. Haviv, O. Ram, O. Press, P. Izsak, and O. Levy, "Transformer language models without positional encodings still learn positional information," *arXiv preprint arXiv:2203.16634*, 2022.

[64] W. Wei, Z. Wang, X. Mao, G. Zhou, P. Zhou, and S. Jiang, "Position-aware self-attention based neural sequence labeling," *Pattern Recognition*, vol. 110, p. 107636, 2021.

[65] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, and F. Hajamohideen, "Explainable artificial intelligence in alzheimer's disease classification: A systematic review," *Cognitive Computation*, vol. 16, no. 1, pp. 1–44, 2024.

[66] R. Liu and Z. Wang, "Assigning channel weights using an attention mechanism: an eeg interpolation algorithm," *Frontiers in Neuroscience*, vol. 17, p. 1251677, 2023.