

PAPER • OPEN ACCESS

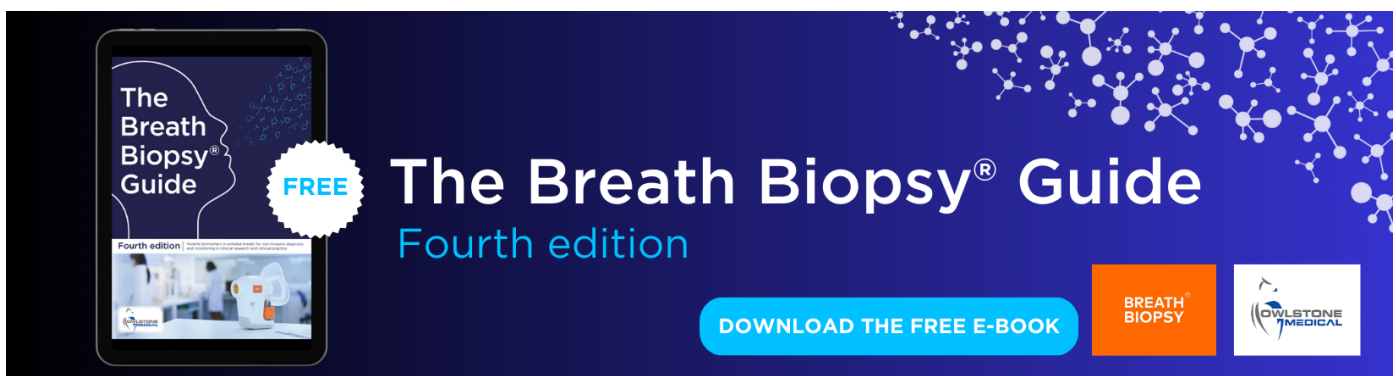
An attention-based deep learning approach for the classification of subjective cognitive decline and mild cognitive impairment using resting-state EEG

To cite this article: Elena Sibilano *et al* 2023 *J. Neural Eng.* **20** 016048

View the [article online](#) for updates and enhancements.

You may also like

- [From single layer to multilayer networks in mild cognitive impairment and Alzheimer's disease](#)
Ignacio Echegoyen, David López-Sanz, Fernando Maestú et al.
- [Characterization of the dynamic behavior of neural activity in Alzheimer's disease: exploring the non-stationarity and recurrence structure of EEG resting-state activity](#)
Pablo Núñez, Jesús Poza, Carlos Gómez et al.
- [Exploring dynamic functional connectivity alterations in the preclinical stage of Alzheimer's disease: an exploratory study from SILCODE](#)
Fan Yang, Xueyan Jiang, Feng Yue et al.



The Breath Biopsy® Guide
Fourth edition

FREE

DOWNLOAD THE FREE E-BOOK

BREATH BIOPSY

OWLSTONE MEDICAL



PAPER

OPEN ACCESS

RECEIVED
4 August 2022REVISED
27 January 2023ACCEPTED FOR PUBLICATION
6 February 2023PUBLISHED
17 February 2023

Original Content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



An attention-based deep learning approach for the classification of subjective cognitive decline and mild cognitive impairment using resting-state EEG

Elena Sibilano^{1,†,*} , Antonio Brunetti^{1,†,*} , Domenico Buongiorno¹ , Michael Lassi² ,
Antonello Grippo³ , Valentina Bessi⁴ , Silvestro Micera^{2,5} , Alberto Mazzoni²
and Vitoantonio Bevilacqua¹

¹ Department of Electrical and Information Engineering, Polytechnic University of Bari, Via Orabona 4, 70125 Bari, Italy

² The BioRobotics Institute, Scuola Superiore Sant'Anna, 56025 Pisa, Italy

³ IRCCS Fondazione Don Carlo Gnocchi, 50143 Florence, Italy

⁴ Department of Neuroscience, Psychology, Drug Research and Child Health, University of Florence, Azienda Ospedaliera Careggi, Florence, Italy

⁵ Bertarelli Foundation Chair in Translational Neuroengineering, Center for Neuroprosthetics, Institute of Bioengineering, School of Engineering, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland

† These authors contributed equally to this work.

* Authors to whom any correspondence should be addressed.

E-mail: elena.sibilano@poliba.it and antonio.brunetti@poliba.it

Keywords: electroencephalography, subjective cognitive decline, mild cognitive impairment, Alzheimer's disease, transformer, attention

Abstract

Objective. This study aims to design and implement the first deep learning (DL) model to classify subjects in the prodromic states of Alzheimer's disease (AD) based on resting-state electroencephalographic (EEG) signals. **Approach.** EEG recordings of 17 healthy controls (HCs), 56 subjective cognitive decline (SCD) and 45 mild cognitive impairment (MCI) subjects were acquired at resting state. After preprocessing, we selected sections corresponding to eyes-closed condition. Five different datasets were created by extracting delta, theta, alpha, beta and delta-to-theta frequency bands using bandpass filters. To classify SCD vs MCI and HC vs SCD vs MCI, we propose a framework based on the transformer architecture, which uses multi-head attention to focus on the most relevant parts of the input signals. We trained and validated the model on each dataset with a leave-one-subject-out cross-validation approach, splitting the signals into 10 s epochs. Subjects were assigned to the same class as the majority of their epochs. Classification performances of the transformer were assessed for both epochs and subjects and compared with other DL models. **Main results.** Results showed that the delta dataset allowed our model to achieve the best performances for the discrimination of SCD and MCI, reaching an Area Under the ROC Curve (AUC) of 0.807, while the highest results for the HC vs SCD vs MCI classification were obtained on alpha and theta with a micro-AUC higher than 0.74. **Significance.** We demonstrated that DL approaches can support the adoption of non-invasive and economic techniques as EEG to stratify patients in the clinical population at risk for AD. This result was achieved since the attention mechanism was able to learn temporal dependencies of the signal, focusing on the most discriminative patterns, achieving state-of-the-art results by using a deep model of reduced complexity. Our results were consistent with clinical evidence that changes in brain activity are progressive when considering early stages of AD.

1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia in the elderly population, accounting for up to 80% of cases worldwide [1]. In AD,

the progression of neurodegeneration is well established by the stage of symptomatic disease. This might represent a considerable limitation in developing disease-modifying therapies, since the majority of interventions have been tested in cohorts with

substantial synaptic and neuronal damage [2]. Thus, a further investigation of initial stages of AD is needed not only for prognostic purposes, but also to define populations with still sufficient functional compensation to be targeted in early clinical trials [3].

According to recent neuroimaging, neuropathological and biochemical investigations, the pathophysiological process of AD can begin decades before cognitive impairment [4]. It is now known that biological markers, such as amyloid-beta ($A\beta$) protein accumulation, which is distinctive in AD, may be found in the brain up to 20 years before the stage of dementia [5].

This evidence led to a new biological definition of the disease, which assumes that the cognitive decline in AD occurs over a long period [6] and develops as a continuum rather than as distinct, clinically-defined entities [7]. On this continuum, three broad phases can be distinguished: preclinical AD, mild cognitive impairment (MCI) due to AD and dementia due to AD [8].

While MCI refers to a well-defined, intermediate stage between normal ageing and pathological status [9], many patients experience a subjective cognitive decline (SCD) in memory and other cognitive domains prior to demonstrable impairment. SCD is not linked to a particular disease status itself [10]. However, it has been proved that the subjective decline, even at the stage of normal cognitive performance on mental tests, is associated with an increased risk of positive biomarkers for Alzheimer's and later conversion to dementia [11–14]. In this context, it has been established that SCD can occur at late stages of preclinical AD, before MCI is reached. This phase can be also referred to as pre-MCI or pre-prodromal AD. In particular, since new diagnostic guidelines have been released, SCD individuals with pathological $A\beta$ levels in cerebrospinal fluid could be considered to be in AD continuum [15]. Nonetheless, SCD constitutes a heterogeneous group, as it could be related to conditions such as normal aging, personality traits, psychiatric conditions, neurological and medical disorders, substance use, and medication [16].

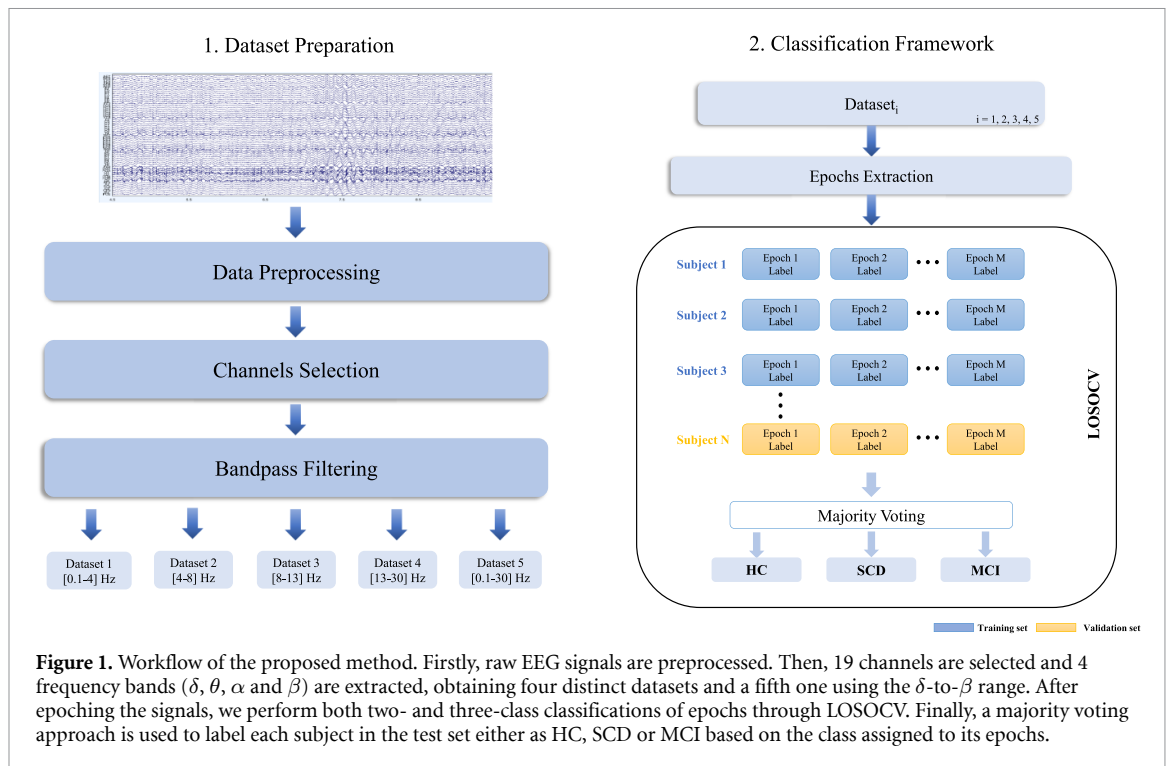
Following these principles, many studies have focused on recognizing biomarkers to characterize and identify SCD at risk of progression to objective cognitive decline. Recently, Viviano and Damoiseaux reviewed several works making discrimination between SCD and healthy controls (HCs) using functional neuroimaging biomarkers, also proposing a model to integrate common features found in subjects affected by SCD [17]. It has also been noted that SCD individuals have a pattern of brain atrophy similar to that measured in AD pathology when compared to HCs without SCD [18]. Moreover, using 18F-fluorodeoxyglucose positron emission tomography (PET), it was found that subjects with SCD show glucose and neuronal hypometabolism

with respect to HCs, which is correlated with decline in memory domain [14]. Altered activation of prefrontal cortex in SCD patients was detected using functional magnetic resonance imaging (MRI), even if there were no changes in verbal episodic memory encoding [14]. Additionally, longitudinal studies results have shown that it is possible to use these markers to predict patients with SCD and MCI who will convert to AD [19, 20]. Since $A\beta$ burden is distinctive in the progression of AD from early stages, Maserejian *et al* developed a statistical framework based on multimodal data, including apolipoprotein E genotype status, to predict $A\beta$ positivity in subjects with SCD and MCI [20]. Results on separate validation datasets indicated an estimated probability of $A\beta$ positivity of up to 0.75 for patients with MCI and of 0.60 for SCD subjects.

Although the task of classifying SCD and MCI subjects from HCs has been addressed in several studies [21, 22], the discrimination between SCD and MCI conditions from a functional point of view is still poorly investigated in literature since anatomical and functional changes in brain between the two classes are subtler, making it a more challenging task to deal with [23]. Nevertheless, the intricacy of brain alterations in the early stages of AD makes it difficult to recognize patterns and develop accurate indicators for diagnosing and monitoring the development of AD on an individual basis [24, 25]. Furthermore, whilst advanced neuroimaging methods like PET and MRI enable to capture relevant modifications in brain processes related to AD, their use is limited in clinical settings due to cost, invasiveness and time consumption [26].

In this respect, electroencephalography (EEG) can represent an alternative technique that is both non-invasive and cost-effective, and much more practical for clinical applications [26, 27]. Since EEG signals reflect functional changes in the cerebral cortex, EEG-based biomarkers can be used to assess neuronal degeneration caused by AD progression long before actual tissue loss or behavioral symptoms appear.

Several studies proposed resting-state EEG (rsEEG) rhythms as candidate biomarkers of AD [28–31]. A more comprehensive review of research in this field can be found in the work by Babiloni *et al* [32]. Cassani *et al* summarized EEG changes related to AD progression into four main categories: slowing, complexity reduction, synchronization decrement and neuromodulatory deficit [33]. At the MCI stage, such EEG abnormalities were found to be intermediate between HCs and dementia patients, and more severe compared to subjects with SCD [34]. Changes in relative and absolute power of theta (θ) frequency band appear to be significant among AD, MCI and HC at individual level [35]. Significantly higher global delta (δ) and theta power, lower global alpha (α) power and a higher global peak frequency



have also been found in patients with SCD that have progressed to MCI and dementia [34]. Hence, measures of EEG-recorded brain activity can represent sensitive, non-invasive markers in the prediction of clinical development of AD. This assumption holds true also when comparing EEG to other neuroimaging methods, both structural and functional [36].

Research into the application of deep learning (DL) models based on EEG signals is growing thanks to the increasing availability of larger EEG datasets. DL enables end-to-end learning from raw inputs, thus overcoming the limitation of processing high-dimensional volumes of data encountered by traditional machine learning (ML) approaches. In the field of EEG data processing, DL has been used to improve and extend existing methods, reducing the need for domain-specific processing and feature extraction pipelines [37]. Compared with convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which are the most extensively used architectures for time-series data classification [38], transformers [39] have shown higher ability to deal with long-range dependencies and recognize patterns in sequences of data [40], as well as employ more interpretable decision-making processes [40, 41]. Although transformers have become the standard models in natural language processing (NLP), recent efforts in exploring their applications on time-series data, such as EEG or electromyography signals, are showing interesting results [42].

In this work, we propose a DL framework based on the transformer model for the binary classification

of resting-state EEG signals of SCD and MCI patients. To the best of our knowledge, this is the first study applying DL to EEG in a well-characterized clinical population at risk for AD, as MCI and SCD patients. The same framework is then employed to perform a multiclass classification among HC, SCD and MCI. Figure 1 shows the implemented workflow. We exploited the power of self-attention mechanism to extract relevant information from the signal in the temporal domain. Specifically, we firstly pre-processed the EEG signals with a well-consolidated, standardized pipeline [43]. Then, we filtered clean EEG signals to extract four main frequency bands, each used to build a new dataset from the original one. An additional dataset was obtained by filtering the signals in the full delta-to-beta range [0.1–30] Hz. Employing a leave-one-subject-out cross-validation (LOSOCV) approach, we trained and tested a model on each dataset to label the subjects based on their brain activity. Then, we compared the obtained results with three CNN-based models, both for the binary and multiclass classification tasks.

The rest of this paper is organized as follows: section 2 introduces the state-of-the-art about studies dealing with the discrimination between SCD and MCI based on biomarkers; section 3 describes the EEG acquisition protocol, the dataset and the pre-processing pipeline; section 4 shows the classification methodology, detailing the transformer model and its application to rsEEG signals; section 5 reports the classification performances which are then discussed in section 6. Lastly, section 7 shows the conclusions

of this work, highlighting the limitations and proposing improvements that could be made in future studies.

2. Related works

Despite longitudinal studies have assessed the increased risk for both SCD and MCI patients to develop Alzheimer's dementia, to the best of our knowledge a limited number of works have investigated changes of distinctive biomarkers to differentiate early AD stages.

Yue *et al* evaluated the extent of asymmetry of hippocampus and amygdala volumes from MRI scans in HC, SCD and MCI subjects [44]. They found significant differences between the latter two groups only when considering asymmetry of hippocampus, indicating that this marker could help the diagnosis of early AD stages. On the other hand, they found significant differences between HC and SCD in the volume of the right hippocampus, right amygdala and asymmetry of amygdala, and those differences were reflected in the comparison of HC and MCI. In a recent study by Li *et al*, an approach based on ML models was exploited on features extracted from MRI data to predict the scores of cognitive tests, i.e. Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment, of HC, SCD and MCI subjects, respectively. Results showed that imaging volumetric features of the brain were more correlated with the scores of cognitive tests than individual features extracted from brain subregions, such as the hippocampal area [45]. Such neuroimaging-based studies, although allow to characterize SCD and MCI effectively, still require time-consuming and expensive techniques to acquire data and thus are not easily replicable.

A study by Scheijbeler *et al* [46] used magnetoencephalography (MEG) data to compute brain network interactions in SCD and MCI patients by means of a permutation index, called inverted joint permutation entropy, which was used to train a logistic regression model. The area under the ROC curve (AUC) value obtained with this index (0.784 for SCD-MCI classification), was higher when compared to other MEG markers. However, a limited number of 18 SCD and 18 MCI subjects was employed and thus a replication of their method on larger samples is needed.

Even fewer works have focused on the role of EEG-derived biomarkers in the classification of SCD and MCI, although a lot of work has been done in discriminating AD subjects from both MCI and HC [31, 47] also employing DL models [48–50].

Recently, quantitative EEG was used by Engedal *et al* to predict the conversion to dementia from a large dataset composed of 200 HC, SCD and MCI subjects for whom follow-up information was available [51]. Spectral features were extracted

from the signal to calculate a dementia index, and a statistical pattern recognition method was employed to evaluate the predictive power of the index, reaching an accuracy of 69% in discriminating converters from non-converters. However, Engedal *et al* predicted conversion to dementia from EEG data of subjects already diagnosed. Lazarou *et al* [25] investigated the power of graph metrics derived from high-density EEG (HD-EEG) to discriminate among HC, SCD, MCI and AD individuals. They expected to find differences in brain connectivity in terms of correlation matrices constructed from the EEG activity. The statistical analyses showed that SCD individuals present network values intermediate to HC and MCI, underlying a common disconnection pattern of the brain connectome in SCD but not to the same extent as in MCI. Nonetheless, in the SCD vs MCI comparison, classification performances of both local and global network measures, evaluated with AUC values, were lower than 60%. Similarly, Abazid *et al* investigated connectivity links in the brain networks derived from rsEEG of SCD, MCI and AD patients by exploiting measures of statistical entropy and a support vector machine to discriminate the classes of patients. They demonstrated the effectiveness of the entropy measure to identify different stages of cognitive dysfunction when considering different graph parameters, reaching high accuracy levels, over 90% [52]. However, these results depend on several stages of signal manipulation (e.g. feature extraction, thresholding and selection) which can highly affect the classification performance.

Indeed, none of the above cited studies addressed the SCD vs MCI classification task by using DL approaches. Thus, we adapted an end-to-end model mainly employed in NLP, the transformer, and the self-attention mechanism, to classify resting-state EEG signals in a dataset of HC, SCD and MCI subjects by focusing on the global patterns of the brain oscillatory activity.

3. Materials

Resting-state EEG recordings of 17 HC, 56 SCD and 45 MCI subjects were collected at *IRCCS Don Carlo Gnocchi* in Florence, Italy. Table 1 reports clinical-demographic information of the study population. Patients with SCD and MCI who self-referred to the Regional Reference Center for Alzheimer's Disease and Cognitive Disorders of Careggi Hospital, Florence were enrolled in the 'Predicting the Evolution of Subjective Cognitive Decline to Alzheimer's Disease With machine learning (PREVIEW)' project, an ongoing prospective cohort study started in October 2020.

Patients were classified as SCD according to the terminology proposed by the SCD initiative

Table 1. Clinical-demographic characteristics of the study population. HC: healthy controls; SCD: subjective cognitive decline; MCI: mild cognitive impairment; MMSE: mini-mental state examination; TIB: Italian brief intelligence Test; SD: standard deviation.

Characteristics	HC ($n = 17$)	SCD ($n = 56$)	MCI ($n = 45$)
Age (mean \pm SD)	64.29 \pm 4.77	66.26 \pm 8.72	74.26 \pm 8.20
Females (%)	41.2	78.3	54.3
Age onset (mean \pm SD)	—	55.15 \pm 8.04	62.09 \pm 9.97
Years of education (mean \pm SD)	15.50 \pm 3.78	12.58 \pm 3.47	10.18 \pm 4.17
MMSE (mean \pm SD)	28.92 \pm 1.19	27.48 \pm 2.28	27.52 \pm 2.13
TIB (mean \pm SD)	—	107.22 \pm 20.48	111.00 \pm 6.01

working group [10], which requires the subject to self-experience a persistent decline in cognitive capacity in comparison with a previously normal status and unrelated to an acute event, as well as normal age-, gender-, and education-adjusted performances on standardized cognitive tests. Patients were classified as MCI according to the National Institute on Aging-Alzheimer's Association workgroups criteria for the diagnosis of MCI [9], specifically requiring: cognitive concern reflecting a change in cognition reported by the clinician or the patient, objective evidence of impairment in one or more cognitive domains (all patients underwent an extensive neuropsychological investigation, with estimation of premorbid intelligence, and assessment of depression), preservation of independence in functional abilities and no signs of dementia. The study was approved by a local ethics committee and individual informed consent was obtained. Experimental procedures were conformed to the Declaration of Helsinki and national guidelines.

Data were acquired using EBNeuro's GalNt system (EBNeuro, Florence, Italy) with 64 channels digitized at a sampling rate of 512 Hz. Among the 64 electrodes, 61 electrodes covered the whole scalp to record EEG while the remaining ones recorded electrooculographic and electrocardiographic activity, and thus were not considered for further analysis. The electrodes were placed according to the 10–10 montage system and electrode-skin impedance was set below 5 k Ω . Subjects were sat in a reclined chair for approximately 20 min.

The acquisition protocol was structured to include both closed and open eyes conditions. Specifically, each subject was asked to open the eyes at irregular intervals and when the signal registered drowsiness. However, since it is known that the two conditions show very different signal properties (e.g. higher alpha band power in the eyes-closed (EC) condition) and according to previous studies, such as Lazarou *et al* [25], we extracted and employed only the EC epochs of the original signal for all the subjects (mean length = 15.03 \pm 1.41 min), which represent the largest part of the protocol.

Raw data were preprocessed offline using Matlab R2019b (The Mathworks, Natick, MA, USA)

and EEGLAB toolbox v.2021.0. Even though it is still not clear whether heavy signal preprocessing is needed when employing DL methods [53], a systematic review on EEG classification carried out by Roy *et al* pointed out how most published papers in this field still preprocess the EEG data before feeding it to deep models [37]. In this work, a standardized pipeline, the PREP pipeline [43], was adapted and employed as a first step to clean the signal. This pipeline uses a robust re-referencing algorithm to interpolate noisy channels and leverages routines from the *cleanline* method to remove line noise components [43]. Although the biggest advantage of this approach is that it removes only deterministic line components, while preserving substantial spectral energy, it can present some drawbacks due to the assumption of signal stationarity [43]. To overcome these limitations, a 50 Hz notch filter was further applied to ensure line noise cleaning. This method can be safely applied on our data since high frequencies of the signal, which could be distorted, were not analyzed [43].

The EEG data recorded from scalp electrodes can be considered summations of real EEG signals and artifacts, which are independent of each other. Independent component analysis has been widely used to remove EEG artifacts, such as eye blinks and muscle activity [54]. Thus, a semi-automatic method employing EEGLAB's ICLabel [55] and manual choice of independent components to retain has then been applied to the signals. Lastly, epochs with excessive noise or artifacts were visually inspected and removed. Figure 2 shows an example of the EEG signal of the first subject before and after applying the preprocessing pipeline.

A cluster of 19 channels, namely Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2, was then selected. Since these channels evenly cover the scalp area, this EEG pattern is the most employed in the literature for similar studies [56] and has been proven to ensure sufficient quality along with possible comparison with previous rsEEG findings of other projects [30]. Subsequently, the signals were bandpass filtered between 0.1 Hz and 45 Hz.

Four main frequency bands, namely delta (δ) [0.1–4] Hz, theta (θ) [4–8] Hz, alpha (α) [8–13] Hz

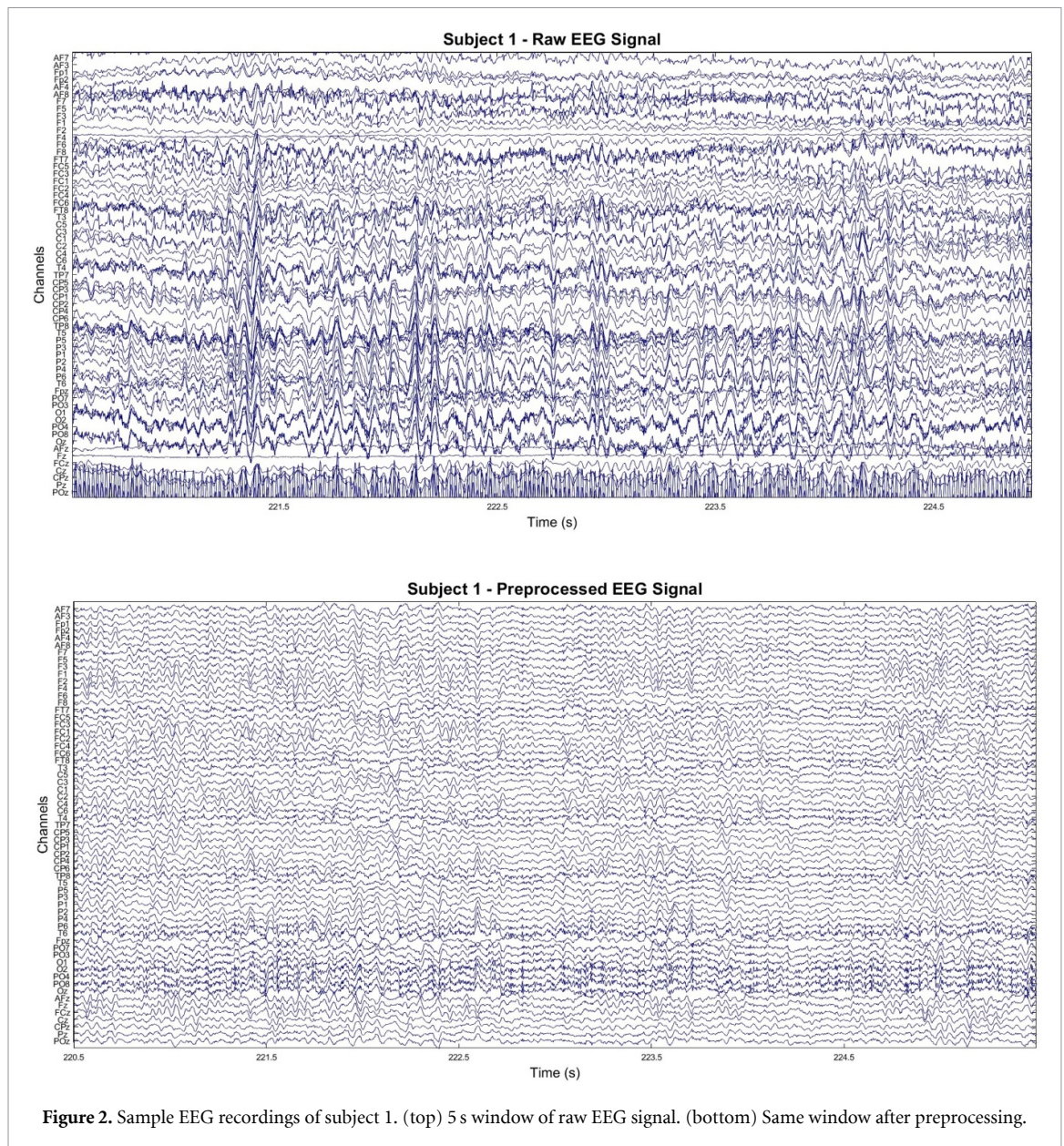


Figure 2. Sample EEG recordings of subject 1. (top) 5 s window of raw EEG signal. (bottom) Same window after preprocessing.

and beta (β) [13–30] Hz were extracted from each EEG signal using designed bandpass filters, and each related dataset was created. Furthermore, in order to assess which frequency band was the most distinctive in the classification of HC, SCD and MCI, we also filtered the signals in the entire range [0.1–30] Hz, and an additional dataset (all-band) was generated. Gamma (γ) band [30–70] Hz was excluded from the analysis since the EEG signal in this band can be significantly contaminated with muscle artifacts [57]. To design filters, we used the *pop_eegfiltnew* function from EEGLAB, which has a heuristic for automatically determining the filter length and order. This function employs a zero-phase Hamming windowed sinc finite impulse response filter [58].

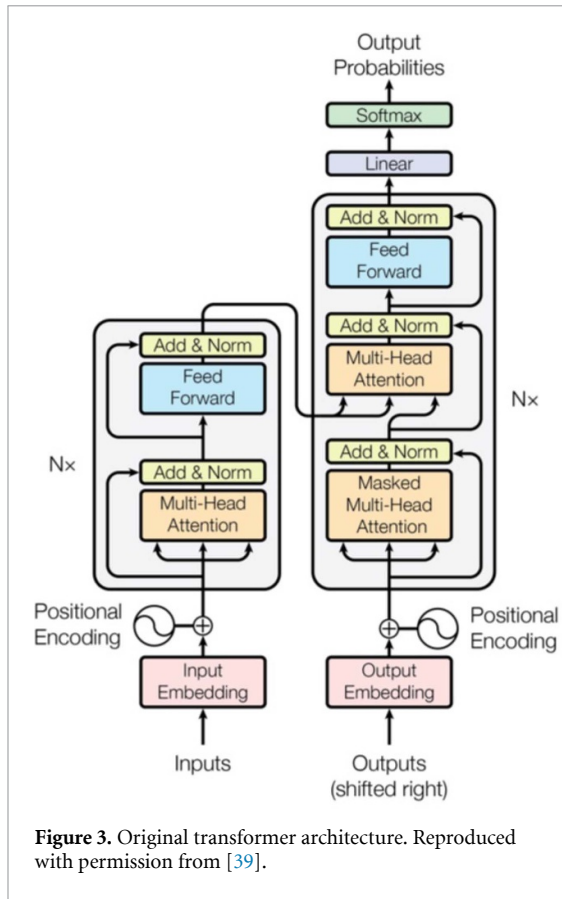
Hence, five different datasets were constructed from the original one, respectively corresponding to specific frequency intervals.

4. Methods

To address the clinical problem of discriminating early stages of AD, we propose an EEG classification method based on the self-attention mechanism and the transformer architecture [39].

Although firstly employed in NLP, transformers have also been proven effective in computer vision tasks [59–62], offering a valid alternative to CNNs and RNNs. In this context, a model called vision transformer (ViT), proposed by Dosovitskiy *et al* in 2020 [63], has yielded interesting results on multiple image recognition benchmarks when compared to state-of-the-art models [64].

Following this path, recent studies have also partly introduced the attention mechanism to EEG decoding [65–69]. In the work by Wei *et al*, the integration of an attention module downstream of a CNN



improved the classification accuracy of MCI and HC [70]. However, these approaches work on hybrid architectures that still heavily rely on CNNs and RNNs to learn discriminative information from EEG, thus not exploiting the computational advantages of transformers at its fullest. Song *et al* [71] implemented a variant of the ViT called spatial-temporal tiny transformer (S3T) to convert the input EEG signal into a discernible representation for motor imagery EEG (MI-EEG) classification purposes. In their model, both spatial and temporal features were captured by applying attention firstly on the EEG channels and then on the EEG time series. The output was a new representation of data that could be classified using a fully-connected layer. As ViT, S3T almost completely disengages from using convolution layers or recurrent layers and relies on the attention mechanism to learn informative features from raw EEG signals. One single convolution operation is kept to learn global positional dependencies of signal segments. Compared to baseline DL models, the authors reached state-of-the-art results using models with a smaller amount of parameters, thus alleviating computational burden and improving scalability.

4.1. Transformer

As shown in figure 3, the core of a transformer consists of an encoder and a decoder with several blocks of the same type. The encoder generates encodings of inputs, while the decoder generates

the output sequence from the encodings. Each transformer block is composed of an attention layer, a feed-forward neural network, shortcut connection and layer normalization. The attention layer is based on the concept of self-attention, which computes an attention function of the inputs to retrieve the dependencies of each element to the others.

Specifically, the input vector is first transformed 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 different inputs are then merged together into three different matrices, namely Q , K and V . Subsequently, the attention function between different input vectors is calculated according to equation (1)

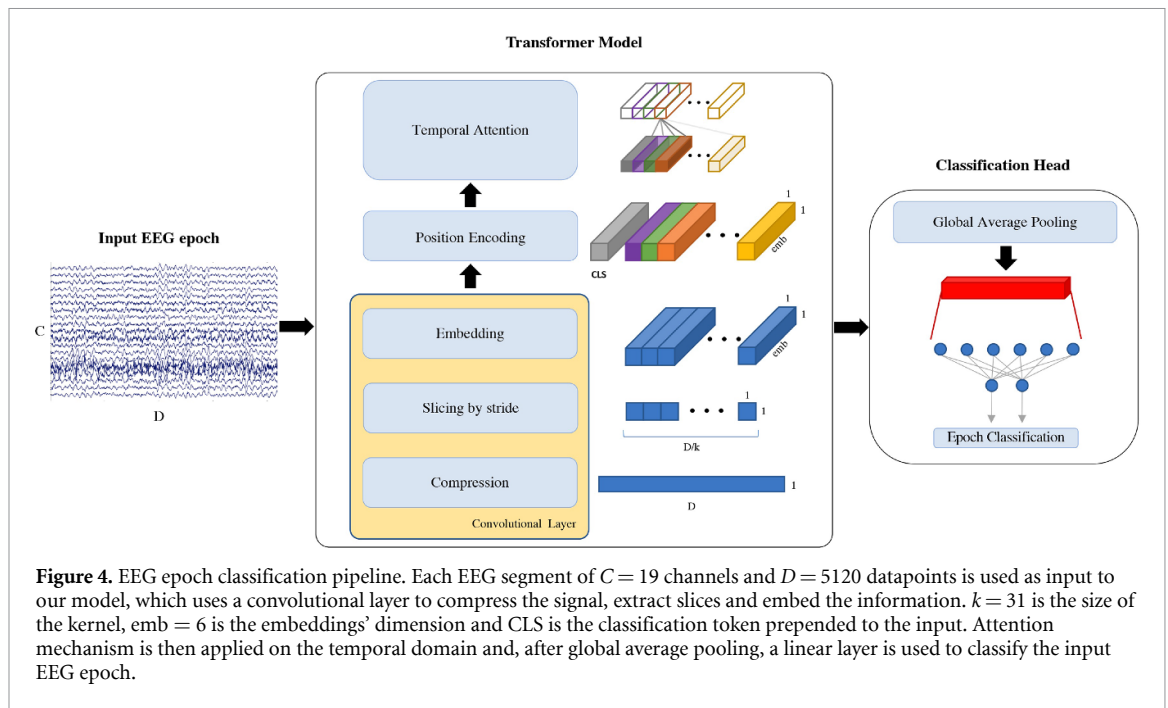
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{dk}}\right) \cdot V. \quad (1)$$

The function computes scores between each pair of inputs, and these values impact how much attention we give to other inputs when encoding the current input. These scores are normalized for gradient stability and then translated into probabilities using the softmax function. Finally, each value vector is multiplied by the sum of the probabilities. The subsequent layer focuses on vectors with higher probability.

The original transformer employs layers of multi-head attention (MHA), which generalize the concept of attention by computing different representation subspaces using H randomly initialized query, key and value matrices, where H is the chosen number of heads. These representations are then concatenated to feed the classification layer. This method allows the model to focus on one or more specific input positions without influencing the attention on other equally important positions at the same time.

ViT directly applies the MHA mechanism to sequences of image patches for image classification tasks [63]. Few modifications are implemented to the original architecture, even though only the transformer encoder module is kept. In such model, sequences of image patches are treated as sequences of words in NLP. 2D images are reshaped into a series of patches of dimension $xp \in \mathbb{R}^{N \times (P^2 \cdot C)}$ where C is the number of image channels, (P, P) is the resolution of each image patch, and N is the total number of resulting patches. A similar approach was proposed by Cordonnier *et al*, but images were divided into patches of dimension 2×2 pixels, thus limiting its use only to small-resolution images [72].

The sequence of patches is then flattened and linearly projected to have a sequence of patch embeddings, which are added to extra learnable embeddings and positional embeddings before being fed to the encoder stack. Since MHA is permutation-equivariant with respect to its inputs, the latter are used to retain spatial information on the position of each patch in the original image.



Lastly, a multilayer perceptron head performs the classification of the resulting encoded representation.

4.2. Proposed model

Following the work by Song *et al* [71], we implemented a pipeline to classify EEG epochs, as shown in figure 4, by designing and training a modified version of the S3T on EC rsEEG signals of SCD and MCI subjects. The same pipeline was followed for the classification of HC, SCD and MCI. For this second task, the last fully connected layer was composed of three output units.

The major difference between the two architectures concerns the way attention is applied to the signals. The proposed model dismisses the spatial attention module, which is used to weight the information encoded by each EEG channel, and prioritizes the temporal domain of the signal. This difference is due to the fact that the objective of this work is to classify resting-state signals, instead of MI signals as in Song *et al* [71]. In fact, while different MI processes activate different areas of the cerebral cortex, and thus spatial channel information was revealed to be of fundamental importance when engaging in a MI classification task [73, 74], resting-states reflect the spontaneous brain activity, thus there is not an established spatial correlation also when investigating cognitive decline associated with AD [75].

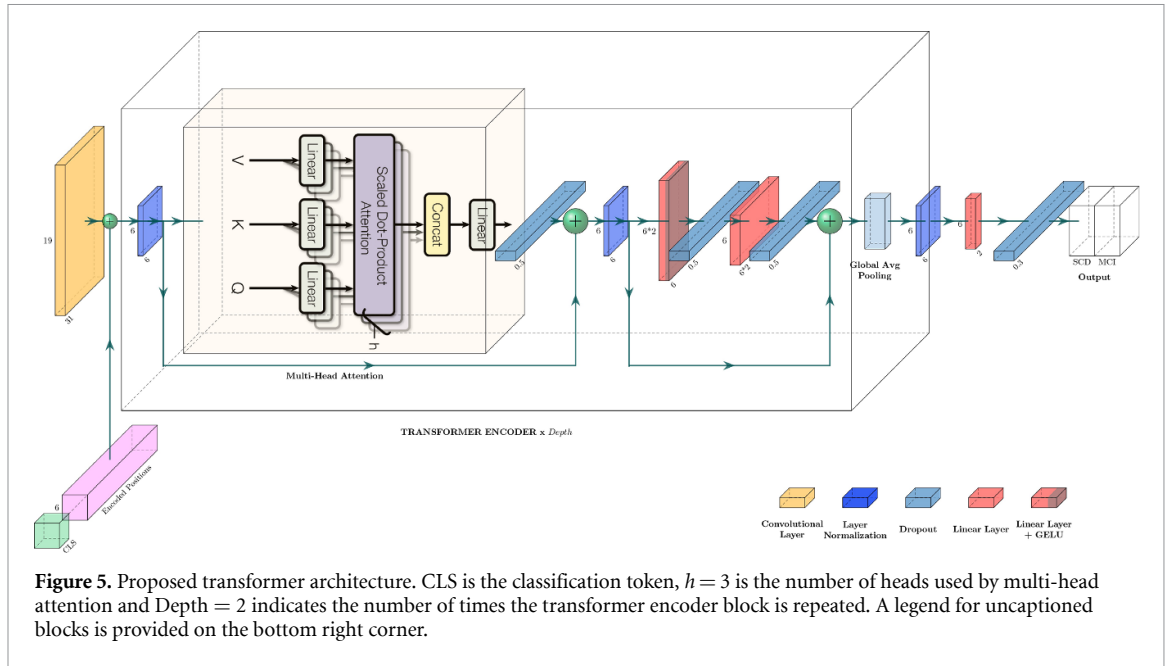
Consequently, our model aims to exploit MHA to understand if temporal dependencies of the EEG sequences can highlight discriminative patterns among HC, SCD and MCI subjects. The MHA layer is included in an encoder block, which combines it with a feed-forward module, a normalization layer and dropout. The encoder block is replicated a number of times specified by the *depth* parameter, which

was set to 2, whereas the number of heads was set to 3. It is worth noting that this configuration is of low complexity and reduced computational cost since it requires fewer parameters than traditional CNNs and RNNs. A graphical representation of the implemented transformer model is shown in figure 5 with reference to SCD vs MCI classification.

Similarly to the original transformer architecture, the proposed model also needs some information on the position of inputs in the time series. This is achieved by Song *et al* by using a convolutional layer on the time dimension before compression, rather than positional encodings as in the original model [71]. Instead, we use a convolutional layer to embed channels' information, compressing it to a single channel representation, and to extract slices from EEG sequences as shown in figure 4. Then, we encode the positions of all slices in the sequence, and the vector of positions is linearly added to the input. Furthermore, we prepend an extra-learnable classification token to each input sequence, which is used to predict the final class after being updated by attention, as in the ViT [63]. Compared to the original S3T model, this position encoding method requires fewer parameters and avoids the use of an additional convolutional layer, which increases the complexity of the model. After the global average pooling, a classification head composed of a fully-connected layer, after layer normalization, is then used to classify the new representation of the input.

4.3. Experimental details

After preprocessing, on each dataset, namely the all-band dataset and delta, theta, alpha and beta datasets, LOSOCV was performed, meaning in each fold all of the subjects except one were used to train the



model, and the remaining subject was used to test it. This cross-validation strategy is the most used across studies that employ rsEEG for AD diagnosis and progression analysis [33].

We split the EEG signal of each subject into epochs of 10 s, meaning our models were trained on windows of $N = 10 \text{ s} \cdot 512 \text{ Hz} = 5120$ data points. Each epoch was associated to the label of the corresponding subject. Since the duration of the recording was different for each subject, the number of epochs generated per subject was variable. However, in order to improve the learning capabilities of the model, the number of EEG epochs of the majority classes, i.e. the SCD in the two-way and both SCD and MCI in the multiclass classification, has been reduced by random sampling for being equal to the number of epochs of the minority class in the training set.

Furthermore, all the epochs were normalized using z-score normalization, which was revealed to be an optimal normalization technique for giving models the ability to make classification across an inter-subject population [76], and effective in classification problems, as well [77]. For each subject, per-channel mean and standard deviation were computed and used to normalize the signal.

The Adam optimizer was used in the training process ($\text{lr} = 10e - 4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\text{eps} = 1e - 08$). The batch size was set to 8, whereas the total number of training iterations was set to 250. We selected cross-entropy as loss function. We also implemented an early-stop mechanism in order to prevent the overfitting of the model during the training. Specifically, the training of the model was stopped if the loss on the validation set did not decrease for 15 consecutive epochs.

First, we evaluated the ability of each model to classify EEG epochs in each class. Then, we implemented a hard voting mechanism to predict the label for each subject, according to equation (2), where L_{pi} is the predicted label for a subject made by the i th classifier and l_{ni} is the predicted label for the n th epoch of the same subject made by the i th classifier

$$L_{pi} = \text{mode}(l_{1i}, l_{2i}, \dots, l_{ni}). \quad (2)$$

In other words, in the two-class problem, given a subject, if the number of epochs labeled as class c was higher than 50% of the total number of epochs of that subject, then the subject was assigned to c . If exactly one half of the epochs were labeled as SCD and the other half as MCI, the subject was classified as MCI in order to reduce the number of false negatives. In the multiclass problem, if the number of epochs labeled as class c was higher than the number of epochs labeled as the other two classes, then the subject was assigned to c . As in the first case, if, for a given subject, different classes had the same number of assigned epochs, the subject was assigned to the class which corresponded to the highest level of progression of impairment.

5. Results

5.1. SCD vs MCI classification

Different metrics were employed to evaluate the performances of the models. To compute all the metrics, we considered the MCI class as positive and the SCD class as negative.

To estimate the epoch-level classification performances, we gathered the predictions and the true

Table 2. Confusion Matrix.

		Predicted Class	
		SCD	MCI
True Class	SCD	TN	FP
	MCI	FN	TP

labels from all the cross-validation test sets, and computed the percentage of correct predictions per class. We evaluated the results in terms of accuracy (equation (3)), sensitivity (equation (4)), specificity (equation (5)) and F1-score (equation (6)), where TP, TN, FP and FN are considered according to the confusion matrix reported in table 2. We also reported the AUC scores, whereas the corresponding ROC plots are shown in figure 6. Due to dataset imbalance, the reported F1-score has been weighted by the number of samples in each class

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)$$

$$\text{F1 - score} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FN} + \text{FP}}. \quad (6)$$

The classification results are reported in table 3 for all the datasets. The best performances have been reached by the transformer model on delta and theta bands. Specifically, an accuracy of 67.4% and a F1-score of 67.3% were obtained for delta, whereas a value of 65.0% was obtained for both metrics on theta. AUC scores on delta as well as on theta were higher than 0.8, revealing that both the classifiers have an overall excellent diagnostic accuracy in discriminating SCD and MCI [78].

Subsequently, we evaluated the capabilities of the models on the classification of patients, which is the main objective of this study, as previously described in section 4. We report the classification performances for all the datasets in terms of accuracy (equation (3)), sensitivity (equation (4)), specificity (equation (5)) and F1-score (equation (6)). The results are detailed in table 4. On the delta band, the model reached the highest value for all the computed metrics, with an accuracy and F1-score of 76.2%, a sensitivity of 73.3% and a specificity of 78.6%. It is worth noting that, when considering the epochs' classification task, both single-band delta and Theta datasets perform better than the all-band dataset, upholding the idea that changes in particular EEG rhythms are more discriminative of SCD and MCI conditions

and easier to be detected by our model. On patient-level classification, delta outperforms all the other datasets.

Lastly, in order to demonstrate the efficacy of the entire workflow, we selected the best-performing frequency bands (i.e. delta and theta) and constructed two supplementary HD-EEG datasets, following the same pipeline as in figure 1, but skipping the channel selection step. Specifically, the new EEG segments used as input to the transformer model had dimensions $C = 61$ channels and $D = 5120$ datapoints. We used the same LOSOCV approach and computed all the metrics in order to compare the results with the previous datasets. On the HD-EEG delta dataset, we obtained an accuracy of 62.8% and F1-score of 62.7% on epochs' classification, while 67.3% and 67.4% were obtained for accuracy and F1-score on patients' classification. On the HD-EEG theta dataset, we obtained 59.8% and 59.7% for accuracy and F1-score on epochs' classification, respectively. On patients' classification, accuracy reached a value of 61.5%, whereas we obtained 61.2% for F1-score. Although, even in this case, the delta band shows the best results, all the metrics are lower when compared to the 19-channel datasets, meaning the information added by using more EEG channels is not useful for our model to perform the classification of SCD and MCI subjects.

5.2. HC vs SCD vs MCI classification

Then, we assessed the performances of our model on the classification of HC, SCD and MCI subjects. As for the SCD vs MCI classification, we reported the results for both epochs and patients. In particular, table 5 reports the performances on epochs in terms of accuracy (equation (3)), F1-score (equation (6)) and AUC, and figure 7 shows the corresponding ROC curves. Specifically, the micro-average ROC curve is reported aggregating, for each dataset, the contribution of all classes.

The best performances have been reached by the transformer model on alpha and theta bands. Specifically, an accuracy of 48.8% and a F1-score of 49.4% were obtained for alpha, whereas values of 48.6% and 49.8% were obtained on the theta band for the same metrics, respectively. AUC scores on both bands were higher than 0.7, revealing that the classifiers have an overall acceptable diagnostic accuracy in discriminating HC, SCD and MCI [78].

Also in this case, we evaluated the capabilities of the models to classify individual subjects. Table 6 reports the performances in terms of accuracy (equation (3)) and F1-score (equation (6)), showing that the theta band has the highest discriminatory power with an accuracy of 54.2% and a F1-score of 54.9%.

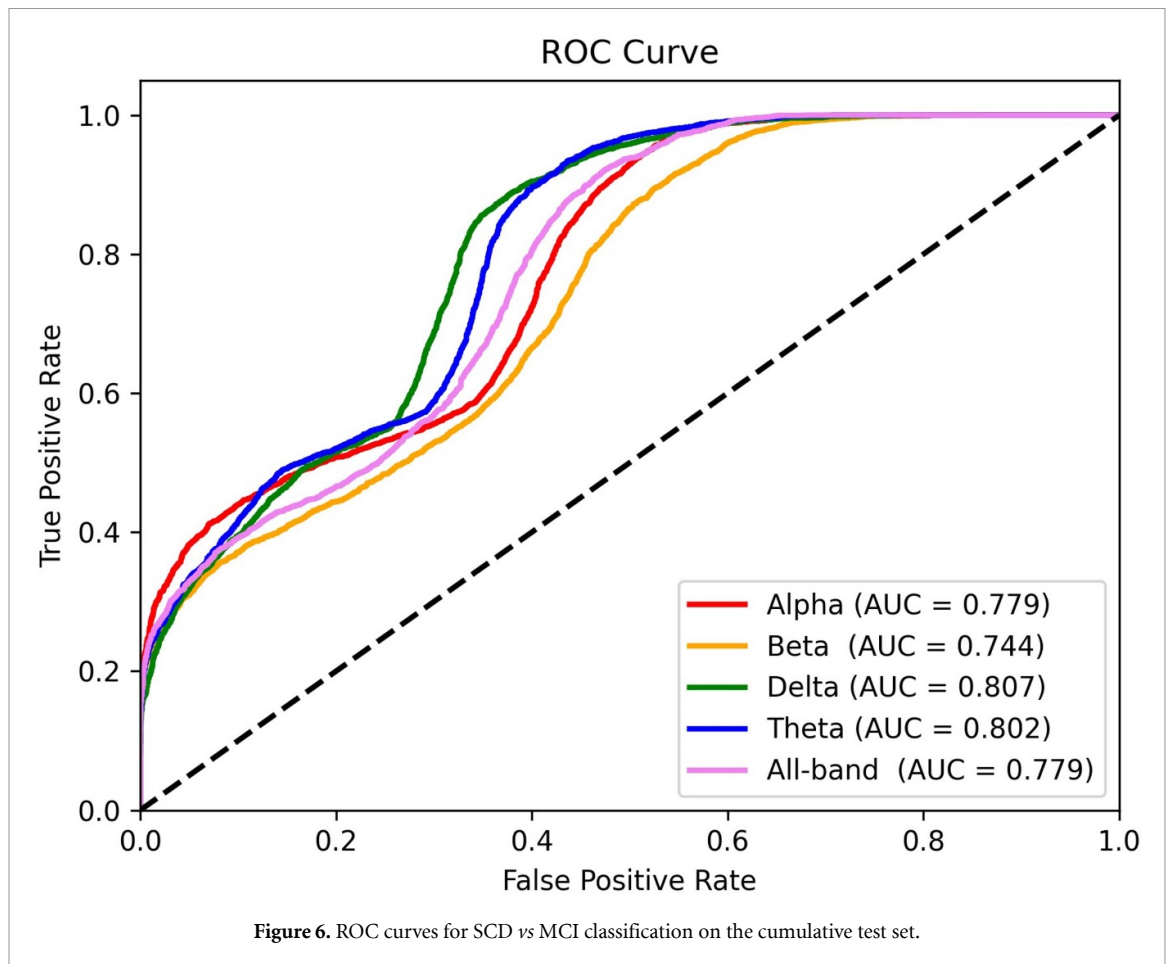


Table 3. Per epoch classification performances. Metrics are computed on the cumulative test confusion matrix. No information rate (NIR) = 0.553. The 95% confidence interval (CI) was calculated for each set with the Clopper–Pearson method for a binomial distribution (accuracy > NIR, *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$). F1-score is weighted for the number of samples per class.

Dataset	Accuracy	CI	AUC	Sensitivity	Specificity	F1-score
Alpha	0.628	[0.618, 0.638]***	0.779	0.602	0.650	0.629
Beta	0.619	[0.608, 0.628]***	0.744	0.598	0.635	0.619
Delta	0.674	[0.664, 0.683]***	0.807	0.620	0.717	0.673
Theta	0.650	[0.640, 0.660]***	0.802	0.591	0.698	0.650
All-band	0.642	[0.632, 0.652]***	0.779	0.561	0.707	0.640

Table 4. Per patient classification performances. Metrics are computed on the cumulative test confusion matrix. No information rate (NIR) = 0.554. The 95% confidence interval (CI) was calculated for each set with the Clopper–Pearson method for a binomial distribution (accuracy > NIR, *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$). F1-score is weighted for the number of samples per class.

Dataset	Accuracy	CI	Sensitivity	Specificity	F1-score
Alpha	0.653	[0.552, 0.745]*	0.644	0.661	0.654
Beta	0.624	[0.552, 0.718]	0.600	0.643	0.624
Delta	0.762	[0.667, 0.841]***	0.733	0.786	0.762
Theta	0.673	[0.573, 0.763]**	0.600	0.732	0.672
All-band	0.673	[0.573, 0.763]**	0.578	0.750	0.671

As in the SCD vs MCI classification task, we obtained two 61-channel datasets corresponding to the bands with the best performances, namely theta and alpha, and compared the results with the corresponding 19-channel datasets. For alpha, we obtained an accuracy of 49.6% and a F1-score of 50.1%

on epochs' classification and 55.1% and 55.3% on patients' classification, respectively. For theta we obtained an accuracy of 45.9% and a F1-score of 47.1% on epochs' classification and 48.3% and 49.8% on patients' classification. These results show that, even in the HC vs SCD vs MCI classification task,

Table 5. Per epoch HC vs SCD vs MCI classification performances. Metrics are computed on the cumulative test confusion matrix. No information rate (NIR) = 0.473. The 95% confidence interval (CI) was calculated for each set with the Clopper–Pearson method for a binomial distribution (accuracy > NIR, *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$). F1-score is weighted for the number of samples per class.

Dataset	Accuracy	CI	AUC	F1-score
Alpha	0.488	[0.479, 0.498]***	0.750	0.494
Beta	0.446	[0.436, 0.455]	0.693	0.448
Delta	0.449	[0.440, 0.458]	0.662	0.470
Theta	0.486	[0.476, 0.495]***	0.745	0.498
All-band	0.443	[0.434, 0.452]	0.681	0.455

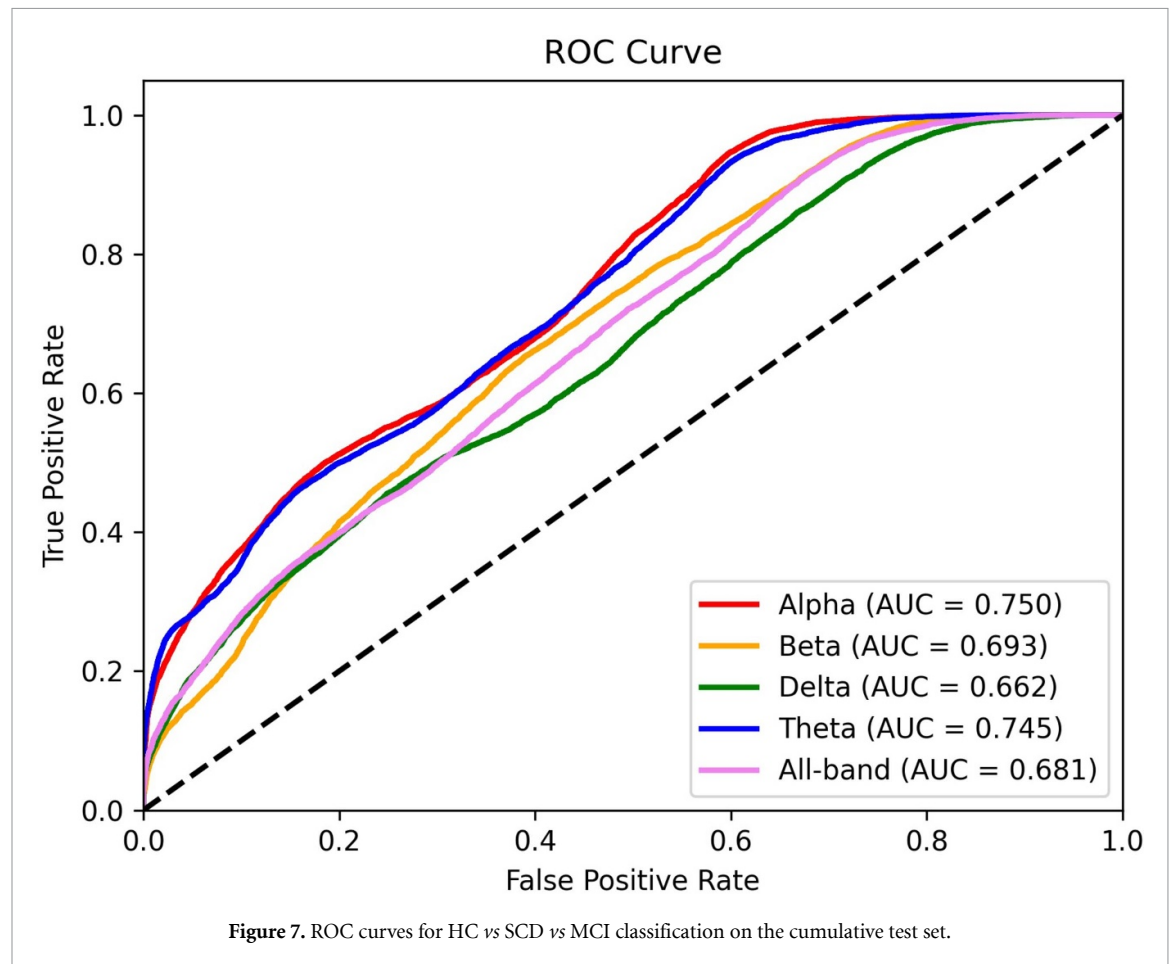


Table 6. Per patient HC vs SCD vs MCI classification performances. Metrics are computed on the cumulative test confusion matrix. No information rate (NIR) = 0.475. The 95% confidence interval (CI) was calculated for each set with the Clopper–Pearson method for a binomial distribution (accuracy > NIR, *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$). F1-score is weighted for the number of samples per class.

Dataset	Accuracy	CI	F1-score
Alpha	0.500	[0.407, 0.593]	0.503
Beta	0.500	[0.407, 0.593]	0.503
Delta	0.500	[0.407, 0.593]	0.527
Theta	0.542	[0.448, 0.634]	0.549
All-band	0.517	[0.423, 0.610]	0.532

using an higher number of EEG channels does not have a significant impact on the performances of our model. In fact, although there was a small increase

in the results on alpha, on theta, which is the best-performing band on subject-wise classification, our transformer continues to give the highest results considering the dataset with 19 channels.

5.3. Performance comparison with CNN-based models

In order to compare our model with state-of-the-art EEG classification models, we conducted experiments with some recent CNN-based models, namely Deep-ConvNet [79], EEGNet [80] and EEG-TCNet [81]. These architectures were mainly developed for MI-based EEG signals decoding, as well as for the classification and interpretation of EEG-based BCIs. Park *et al* employed them in the field of the identification of preclinical AD from EEG to overcome the limitation

Table 7. SCD vs MCI classification performance comparison in terms of overall accuracy on the cumulative test set of the DL models. No information rate (NIR) for epochs classification = 0.553; NIR for patients classification = 0.554. The 95% confidence interval (CI) was calculated for each set with the Clopper–Pearson method for a binomial distribution (accuracy > NIR, *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$).

Model	Dataset	Epochs		Patients	
		Accuracy	CI	Accuracy	CI
Transformer	Delta	0.674	[0.664, 0.683]***	0.762	[0.667, 0.841]***
	Theta	0.650	[0.640, 0.660]***	0.673	[0.573, 0.763]**
EEGNet	Delta	0.726	[0.716, 0.735]***	0.733	[0.635, 0.816]***
	Theta	0.669	[0.659, 0.678]***	0.683	[0.583, 0.772]**
DeepConvNet	Delta	0.590	[0.580, 0.600]***	0.653	[0.552, 0.745]*
	Theta	0.589	[0.579, 0.599]***	0.594	[0.492, 0.691]
EEG-TCNet	Delta	0.673	[0.664, 0.683]***	0.703	[0.604, 0.790]**
	Theta	0.693	[0.683, 0.702]***	0.683	[0.583, 0.772]**

Table 8. HC vs SCD vs MCI classification performance comparison in terms of overall accuracy on the cumulative test set of the DL models. No information rate (NIR) for epochs classification = 0.473; NIR for patients classification = 0.475. The 95% confidence interval (CI) was calculated for each set with the Clopper–Pearson method for a binomial distribution (accuracy > NIR, *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$).

Model	Dataset	Epochs		Patients	
		Accuracy	CI	Accuracy	CI
Transformer	Alpha	0.488	[0.479, 0.498]***	0.500	[0.407, 0.593]
	Theta	0.486	[0.476, 0.495]***	0.542	[0.448, 0.634]
EEGNet	Alpha	0.472	[0.462, 0.481]	0.483	[0.390, 0.577]
	Theta	0.451	[0.442, 0.461]	0.424	[0.333, 0.518]
DeepConvNet	Alpha	0.479	[0.469, 0.488]	0.492	[0.398, 0.585]
	Theta	0.476	[0.467, 0.486]	0.500	[0.407, 0.593]
EEG-TCNet	Alpha	0.467	[0.458, 0.477]	0.500	[0.407, 0.593]
	Theta	0.495	[0.486, 0.505]***	0.508	[0.415, 0.602]

of high inter-subject variability, which affects the possibility of extracting robust handcrafted features [82]. However, to our knowledge, they have never been used for the specific task of discriminating SCD from MCI. It is worth noting that these models are characterized by a higher number of parameters than our transformer. Indeed, while the transformer contains a total of 5.2 k parameters, DeepConvNet, EEGNet and TCNet have 298.6 k, 9.8 k and 14.1 k parameters, respectively.

The models' parameters were adjusted to take EEG epochs of dimension $C \times D$ in input as our transformer model. The comparison was performed on the best-performing datasets for both classification tasks, i.e. delta and theta for SCD vs MCI and alpha and theta for HC vs SCD vs MCI. Results are reported in tables 7 and 8, respectively. For SCD vs MCI classification, all the models reached comparable performances in terms of accuracy, which was always significantly higher than no-information rate for epoch classification. For the delta band, the classification accuracy of patients was 76.2% for the transformer, while the same metric has values of 73.3% for EEGNet, 65.3% for DeepConvNet and 70.3% for EEG-TCNet. In all the cases, except for DeepConvNet, the accuracy was always significantly higher than no-information rate ($p \leq 0.001$ for transformer and EEGNet, $p \leq 0.01$ for EEG-TCNet).

Concerning the HC vs SCD vs MCI classification, the epochs' classification accuracy was significantly higher than no-information rate for the transformer, for both alpha and theta bands, and EEG-TCNet, for the theta band only ($p \leq 0.001$ for all the cases). However, patients' classification accuracy was not significantly higher than the no-information rate, except for the transformer which reached a near significance ($p = 0.08$), with a value of 54.2% against 48.3% for EEGNet, 49.2% for DeepConvNet and 50.0% for EEG-TCNet.

6. Discussion

In this work we propose the first DL framework based on the attention mechanism to discriminate between SCD and MCI using rsEEG signals. Previous studies that have addressed the task of discriminating SCD and MCI patients in the AD continuum, with statistical or traditional ML approaches, have highlighted that this problem is much more challenging than other classification tasks in the same field. This evidence can be deduced both from works that employ MRI data [23, 44] and EEG data [25]. It is also supported by other works in literature [31, 47–50], some of which show in general better classification performances than those obtained in this work but

considering different classes of subjects, e.g. HC *vs* SCD, HC *vs* MCI or MCI *vs* AD.

For the SCD *vs* MCI classification task, by comparing the results on all the test sets gathered from a LOSOCV approach, we found that delta and theta bands had the best performances with AUC values of 0.807 and 0.802, respectively. Furthermore, the other classification metrics, i.e. accuracy, sensitivity, specificity and F1-score, were the highest on delta, reporting a value of 67.4% for accuracy and 67.3% for F1-score on epoch-wise classification and a value of 72.6% for both accuracy and F1-score on patient-wise classification. On the same band, the model reached good sensitivity and specificity values, respectively of 73.3% and 78.6%, showing it is capable of discriminating SCD and MCI subjects when they have that specific condition. For both delta and theta bands, the classification accuracy was significantly higher than the no-information rate ($p \leq 0.001$ and $p \leq 0.01$ for epoch-wise and patient-wise classifications, respectively), assessing that the classifier model performed better than one could do by always predicting the most common class. Indeed, changes in relative power in the lower frequencies (δ and θ) indicate a diffuse slowing of brain oscillations, which is a hallmark feature in the progression of AD [33]. In this context, EEG spectral analysis revealed that higher delta and theta powers are associated with clinical progression of SCD patients towards MCI and dementia, mainly when considering EC resting-state activity, as it has been done in this study [34].

Our results uphold this evidence, showing that changes in delta and theta are particularly useful in characterizing the brain activity of subjects affected by SCD or MCI, both when compared to other common EEG rhythms and to the all-band dataset, which includes the signals filtered in the range [0.1–30] Hz. Furthermore, the MHA mechanism well captures temporal dependencies of rsEEG, highlighting their importance in the discrimination between SCD and MCI. This is supported also by a recent work by Wei *et al*, who employed the attention mechanism to classify MCI and HC using EEG signals recorded during cognitive tasks [70]. In fact, this approach allowed to improve the performances of a traditional CNN by almost 10%, suggesting that the use of this technique should be further investigated.

On the other hand, we found that adding more spatial details by using all available 61 EEG channels, instead of a cluster of 19 channels, not only did not improve the performances of the model, but all the metrics reported lower values for both epochs' and patients' classification performed on delta and theta datasets. Hence, we showed that more spatial information increases the complexity and redundancy of the signal pattern produced by the selected 19 channels, which already holds enough information for the model to distinguish between the two classes.

In order to further assess the performance of our model, we added a control group of 17 healthy subjects and conducted a multiclass classification to discriminate HC, SCD and MCI simultaneously. We found that, in this case, the best-performing frequency bands were alpha and theta both on epochs' and patients' classification tasks. Specifically, on alpha the AUC was 0.750, and slightly lower for theta as shown by the ROC curves in figure 7. However, theta reported the best performances in terms of accuracy and F1-score when classifying subjects. In addition, alpha and theta bands were the only ones that reached classification accuracies significantly higher than no-information rate ($p \leq 0.001$ for both bands). These results are in line with evidence reported in literature that both SCD and MCI subjects are characterized by lower amplitude of posterior alpha rhythms in rsEEG in relation to cognitive functions when compared to controls [83] and that this feature, along with higher amplitude of δ - θ rhythms, is related to worsening of impairment over time [30].

Also in relation to this task, single-band datasets performed better than the all-band dataset, showing that specific EEG rhythms can be strong prognostic biomarkers for cognitive impairment in the context of AD. Furthermore, we conducted experiments using HD-EEG alpha and theta datasets and showed that, as in the previous case, increasing the number of channels does not significantly improve the capabilities of our model in discriminating among the three classes of subjects, since marginally higher results were obtained on alpha but not on theta.

In terms of classification performances, we compared the transformer with three DL models based on CNNs for both binary and multiclass classification tasks. The results reported in tables 7 and 8 show that all the models achieve overall good performances. In particular, focusing on the patients' binary classification, all the classifiers, except DeepConvNet on the theta band, reach good accuracy levels ($>70\%$), significantly higher than the no-information rate ($p \leq 0.001$ for transformer and EEGNet; $p \leq 0.01$ for EEG-TCNet). The classification accuracy of patients in the multiclass approach, instead, was not significantly higher than the no-information rate in any case. Nevertheless, accuracy higher than 50% was achieved only by the transformer and EEG-TCNet on the theta band; in these cases, the performance on epochs' classification was significantly higher than no-information rate ($p \leq 0.001$), meaning that both models uncovered a pattern underlying EEG data which allows the discrimination of HC, SCD and MCI subjects.

We also performed statistical analysis in order to assess the significance of our results on the cumulative test set. One-way ANOVA was carried out for each group of data (i.e. SCD *vs* MCI on delta and theta bands, and HC *vs* SCD *vs* MCI on alpha and theta

bands), considering the model as factor. The analysis did not reach the statistical significance ($p < 0.05$) in all the cases, except for the SCD vs MCI classification on delta band ($p = 0.023$).

This result should be interpreted considering that we conducted the study implementing a LOSOCV approach which, in any case, allows an estimation of the generalization capabilities of the implemented models on the data of unseen subjects [84].

Despite the performances of the transformer for the specific classification tasks are not outperforming when compared to the results obtained by CNN-based models, the use of this model still brings advantages that are worth considering. In fact, as already reported in the previous section, the transformer model is less complex, with 5.2 k of trainable parameters, when compared to DeepConvNet, EEGNet and EEG-TCNet, which have 298.6 k, 9.8 k and 14.1 k parameters, respectively. As evidenced by a recent survey by Hu *et al*, reducing the complexity of DL models while guaranteeing, at the same time, a sufficient level of expressive capacity by the model itself for a given task, is an open research problem [85]. In this perspective, the transformer model already demonstrated classification capabilities comparable with more complex models.

In addition, the attention mechanism implemented by the transformer, which is perfectly suited for the classification of temporal signals, may allow the exploration of its interpretability and explainability capabilities by analyzing temporal dependencies in the EEG signals exploiting the attention weights [86, 87].

The results obtained in this work are very promising, mostly considering that the works that discriminate between SCD and MCI subjects have addressed the problem by employing data whose acquisition is still time- and cost-consuming or by relying on techniques that are highly dependent on signal manipulation and feature extraction.

7. Conclusion

In this study we presented the first DL framework that employs the attention mechanism implemented by the transformer model to classify patients affected by early-stage conditions of AD at individual level, using resting-state EEG signal. SCD might represent a pre-clinical manifestation of AD, but several of its aspects remain unclear. In particular, the boundary between SCD and MCI, which is the first clinically diagnosed stage of the disease, is very blurry [4]. Although studies based on neuroimaging methods have investigated the possibility of identifying SCD and MCI in the progression of AD, reporting interesting results, they still require expensive and time-consuming diagnostic tools [21, 45].

In this context, the prognostic power of resting-state EEG has been investigated to address the need

for non-invasive and cost-aware markers to stratify patients in the Alzheimer's continuum [30, 88]. Since spectral analyses conducted by previous studies revealed a correlation between clinical progression of the disease and signal alterations in specific frequency bands [34, 35, 89], e.g. power spectrum shifts from high-frequency components (α , β , and γ) towards low-frequency components (δ and θ) [33], we hypothesized that our model would perform better when trained on signals corresponding to these ranges. The implemented workflow and the obtained results confirmed our hypothesis, also revealing that EEG signals should be further taken into consideration as a clinical biomarker for the characterization of SCD and MCI subjects at risk for AD. Our conclusion is that DL might indeed be the key to move from using EEG for a mere characterization of prodromic states of AD to use it as a predictive tool easy to acquire and process. Early identifying the preclinical stages of AD has become fundamental; as evidenced by Rabin *et al*, risky subjects might represent a target population for disease-modifying therapies in the future [90]. In addition, populations in the initial stages of AD might be targeted in early clinical trials, in order to drive the research towards novel pharmacological approaches to slow the progression to dementia [91].

Nonetheless, a further improvement of the proposed model, which could include the application of a spatial attention module on the EEG channels, is worth exploring. Moreover, a deeper interpretation of the results of this study could be provided by leveraging the attention weights in order to better understand the model's decision processes. Future works may also focus on novel approaches for signal analysis and processing for classification pipelines, in order to remove all the irrelevant information from the input signals. Finally, the availability of a larger cohort of subjects would allow to further assess the effectiveness of our method and improve its generalization capabilities.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

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.

Ethical statement

The local ethics committee of IRCCS Don Carlo Gnocchi in Florence—Italy approved the protocol of the study. All participants gave written informed consent. All procedures involving experiments on human subjects were done in accordance with the ethical standards of the Committee on Human Experimentation of the institution in which the experiments were done or in accordance with the Helsinki Declaration of 1975. Specific national laws have been observed.

ORCID iDs

Elena Sibilano  <https://orcid.org/0000-0002-0878-7675>

Antonio Brunetti  <https://orcid.org/0000-0002-1934-0983>

Domenico Buongiorno  <https://orcid.org/0000-0002-2024-5369>

Michael Lassi  <https://orcid.org/0000-0002-8805-0293>

Antonello Grippo  <https://orcid.org/0000-0002-9997-8564>

Valentina Bessi  <https://orcid.org/0000-0002-6176-3584>

Silvestro Micera  <https://orcid.org/0000-0003-4396-8217>

Alberto Mazzoni  <https://orcid.org/0000-0002-9632-1831>

Vitoantonio Bevilacqua  <https://orcid.org/0000-0002-3088-0788>

References

- [1] Alzheimer's Association 2021 Alzheimer's disease facts and figures *Alzheimer's Dement.* **17** 327–406
- [2] Sperling R, Mormino E and Johnson K 2014 The evolution of preclinical Alzheimer's disease: implications for prevention trials *Neuron* **84** 608–22
- [3] Perrin R J, Fagan A M and Holtzman D M 2009 Multimodal techniques for diagnosis and prognosis of Alzheimer's disease *Nature* **461** 916–22
- [4] Cheng Y-W, Chen T-F and Chiu M-J 2017 From mild cognitive impairment to subjective cognitive decline: conceptual and methodological evolution *Neuropsychiatr. Dis. Treat.* **13** 491
- [5] Bateman R J et al 2012 Clinical and biomarker changes in dominantly inherited Alzheimer's disease *New Engl. J. Med.* **367** 795–804
- [6] Resnick S et al 2010 Longitudinal cognitive decline is associated with fibrillar amyloid-beta measured by [11C] PiB *Neurology* **74** 807–15
- [7] Jack J C R et al 2018 NIA-AA research framework: toward a biological definition of Alzheimer's disease *Alzheimer's Dement.* **14** 535–62
- [8] Prince M, Comas-Herrera A, Knapp M, Guerchet M and Karagiannidou M 2016 Improving healthcare for people living with dementia: coverage, quality and costs now and in the future *World Alzheimer Report 2016*
- [9] Albert M S et al 2011 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 Dement.* **7** 270–9
- [10] Jessen F et al 2014 A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease *Alzheimer's Dement.* **10** 844–52
- [11] Dufouil C, Fuhrer R and Alperovitch A 2005 Subjective cognitive complaints and cognitive decline: consequence or predictor? The epidemiology of vascular aging study *J. Am. Geriatr. Soc.* **53** 616–21
- [12] Rami L, Fortea J, Bosch B, Solé-Padullés C, Lladó A, Iranzo A, Sánchez-Valle R and Molinuevo J L 2011 Cerebrospinal fluid biomarkers and memory present distinct associations along the continuum from healthy subjects to AD patients *J. Alzheimer's Dis.* **23** 319–26
- [13] Amariglio R E et al 2012 Subjective cognitive complaints and amyloid burden in cognitively normal older individuals *Neuropsychologia* **50** 2880–6
- [14] Sun Y, Yang F-C, Lin C-P and Han Y 2015 Biochemical and neuroimaging studies in subjective cognitive decline: progress and perspectives *CNS Neurosci. Ther.* **21** 768–75
- [15] Dubois B et al 2016 Preclinical Alzheimer's disease: definition, natural history and diagnostic criteria *Alzheimer's Dement.* **12** 292–323
- [16] Margolis S A, Kelly D A, Daiello L A, Davis J, Tremont G, Pillemer S, Denby C and Ott B R 2021 Anticholinergic/sedative drug burden and subjective cognitive decline in older adults at risk of Alzheimer's disease *J. Gerontol. A* **76** 1037–43
- [17] Viviano R P and Damoiseaux J S 2020 Functional neuroimaging in subjective cognitive decline: current status and a research path forward *Alzheimer's Res. Ther.* **12** 1–18
- [18] Perrotin A et al 2015 Hippocampal subfield volumetry and 3D surface mapping in subjective cognitive decline *J. Alzheimer's Dis.* **48** S141–50
- [19] Liu Y, Yue L, Xiao S, Yang W, Shen D and Liu M 2022 Assessing clinical progression from subjective cognitive decline to mild cognitive impairment with incomplete multi-modal neuroimages *Med. Image Anal.* **75** 102266
- [20] Maserejian N, Bian S, Wang W, Jaeger J, Syrjanen J A and Aakre J, Jack C R Jr., Mielke M M and Gao F 2019 Practical algorithms for amyloid β probability in subjective or mild cognitive impairment *Alzheimer's Dement.* **11** 710–20
- [21] Chen H, Li W, Sheng X, Ye Q, Zhao H, Xu Y and Bai F 2022 Machine learning based on the multimodal connectome can predict the preclinical stage of Alzheimer's disease: a preliminary study *Eur. Radiol.* **32** 448–59
- [22] Liu T, Wang Y, Yan T, Liu Y, Xu R, Li J and Xie Y 2018 Preclinical stages of Alzheimer's disease classification by a Rs-fMRI study 2018 11th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (IEEE) pp 1–6
- [23] Huang W, Li X, Li X, Kang G, Han Y and Shu N 2021 Combined support vector machine classifier and brain structural network features for the individual classification of amnesic mild cognitive impairment and subjective cognitive decline patients *Front. Aging Neurosci.* **13** 687927
- [24] Rathore S, Habes M, Ifikhar M A, Shacklett A and Davatzikos C 2017 A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages *NeuroImage* **155** 530–48
- [25] Lazarou I, Georgiadis K, Nikolopoulos S, Oikonomou V P, Tsolaki A, Kompatsiaris I, Tsolaki M and Kugiumtzis D 2020 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 Sci.* **10** 392
- [26] Rossini P M et al 2020 Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal

- analysis. Report from the IFCN-sponsored panel of experts *Clin. Neurophysiol.* **131** 1287–310
- [27] Lizio R, Vecchio F, Frisoni G B, Ferri R, Rodriguez G and Babiloni C 2011 Electroencephalographic rhythms in Alzheimer's disease *Int. J. Alzheimer's Dis.* **2011** 927573
- [28] Ding Y, Chu Y, Liu M, Ling Z, Wang S, Li X and Li Y 2022 Fully automated discrimination of Alzheimer's disease using resting-state electroencephalography signals *Quant. Imaging Med. Surg.* **12** 1063
- [29] Torabinikeh M, Asayesh V, Dehghani M, Kouchakzadeh A, Marhamati H and Gharibzadeh S 2022 Correlations of frontal resting-state EEG markers with MMSE scores in patients with Alzheimer's disease *Egypt. J. Neurol. Psychiatry Neurosurg.* **58** 1–7
- [30] Babiloni C et al 2020 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 *Neurobiol. Aging* **90** 43–59
- [31] Triggiani A I et al 2017 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 *Front. Neurosci.* **10** 604
- [32] Babiloni C et al 2021 Measures of resting state EEG rhythms for clinical trials in Alzheimer's disease: recommendations of an expert panel *Alzheimer's Dement.* **17** 1528–53
- [33] Cassani R, Estarellas M, San-Martin R, Fraga F J and Falk T H 2018 Systematic review on resting-state EEG for Alzheimer's disease diagnosis and progression assessment *Dis. Markers* **2018** 5174815
- [34] Gouw A A, Alsema A M, Tijms B M, Borta A, Scheltens P, Stam C J and van der Flier W M 2017 EEG spectral analysis as a putative early prognostic biomarker in nondemented, amyloid positive subjects *Neurobiol. Aging* **57** 133–42
- [35] Musaeus C S et al 2018 EEG theta power is an early marker of cognitive decline in dementia due to Alzheimer's disease *J. Alzheimer's Dis.* **64** 1359–71
- [36] Parker A F, Ohlhauser L, Scarapicchia V, Smart C M, Szoek C and Gawryluk J R 2022 A systematic review of neuroimaging studies comparing individuals with subjective cognitive decline to healthy controls *J. Alzheimer's Dis.* **86** 1545–67
- [37] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk T H and Faubert J 2019 Deep learning-based electroencephalography analysis: a systematic review *J. Neural Eng.* **16** 051001
- [38] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L and Muller P-A 2019 Deep learning for time series classification: a review *Data Min. Knowl. Discov.* **33** 917–63
- [39] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* p 30
- [40] Sun J, Xie J and Zhou H 2021 EEG classification with transformer-based models 2021 *IEEE 3rd Global Conf. on Life Sciences and Technologies (LifeTech)* (IEEE) pp 92–93
- [41] Gilpin L H, Bau D, Yuan B Z, Bajwa A, Specter M and Kagal L 2018 Explaining explanations: an overview of interpretability of machine learning 2018 *IEEE 5th Int. Conf. on Data Science and Advanced Analytics (DSAA)* (IEEE) pp 80–89
- [42] Montazerin M, Zabihi S, Rahimian E, Mohammadi A and Naderkhani F 2022 ViT-HGR: vision transformer-based hand gesture recognition from high density surface EMG signals (arXiv:220110060)
- [43] Bigdely-Shamlo N, Mullen T, Kothe C, Su K-M and Robbins K A 2015 The PREP pipeline: standardized preprocessing for large-scale EEG analysis *Front. Neuroinform.* **9** 16
- [44] Yue L, Wang T, Wang J, Li G, Wang J, Li X, Li W, Hu M and Xiao S 2018 Asymmetry of hippocampus and amygdala defect in subjective cognitive decline among the community dwelling Chinese *Front. Psychiatry* **9** 226
- [45] Li A, Yue L, Xiao S and Liu M 2022 Cognitive function assessment and prediction for subjective cognitive decline and mild cognitive impairment *Brain Imaging Behav.* **16** 645–58
- [46] Scheijbeler E P, van Nifterick A M, Stam C J, Hillebrand A, Gouw A A and de Haan W 2022 Network-level permutation entropy of resting-state MEG recordings: a novel biomarker for early-stage Alzheimer's disease? *Netw. Neurosci.* **6** 382–400
- [47] Gunes S, Aizawa Y, Sugashi T, Sugimoto M and Rodrigues P P 2022 Biomarkers for Alzheimer's disease in the current state: a narrative review *Int. J. Mol. Sci.* **23** 4962
- [48] Perez-Valero E, Lopez-Gordo MA, Gutiérrez C M, Carrera-Muñoz I and Vilchez-Carrillo R M 2022 A self-driven approach for multi-class discrimination in Alzheimer's disease based on wearable EEG *Comput. Methods Programs Biomed.* **220** 106841
- [49] Alvi A M, Siuly S, Wang H, Wang K and Whittaker F 2022 A deep learning based framework for diagnosis of mild cognitive impairment *Knowl.-Based Syst.* **248** 108815
- [50] Fouladi S, Safaei A A, Mammone N, Ghaderi F and Ebadi M 2022 Efficient deep neural networks for classification of Alzheimer's disease and mild cognitive impairment from scalp EEG recordings *Cogn. Comput.* **14** 1247–68
- [51] Engedal K, Barca M L, Høgh P, Andersen B B, Dombernowsky N W, Naik M, Gudmundsson T, Øksengaard A-R, Wahlund L-O and Snaedal J 2020 The power of EEG to predict conversion from mild cognitive impairment and subjective cognitive decline to dementia *Dement. Geriatr. Cogn. Disorders* **49** 38–47
- [52] Abazid M, Houmani N, Boudy J, Dorizzi B, Mariani J and Kinugawa K 2021 A comparative study of functional connectivity measures for brain network analysis in the context of AD detection with EEG *Entropy* **23** 1553
- [53] Craik A, He Y and Contreras-Vidal J L 2019 Deep learning for electroencephalogram (EEG) classification tasks: a review *J. Neural Eng.* **16** 031001
- [54] Peng W 2019 EEG preprocessing and denoising *EEG Signal Processing and Feature Extraction* (Berlin: Springer) pp 71–87
- [55] Pion-Tonachini L, Kreutz-Delgado K and Makeig S 2019 ICLabel: an automated electroencephalographic independent component classifier, dataset and website *NeuroImage* **198** 181–97
- [56] Ferri R et al 2021 Stacked autoencoders as new models for an accurate Alzheimer's disease classification support using resting-state EEG and MRI measurements *Clin. Neurophysiol.* **132** 232–45
- [57] Whitham E M et al 2007 Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG *Clin. Neurophysiol.* **118** 1877–88
- [58] Widmann A, Schröger E and Maess B 2015 Digital filter design for electrophysiological data—a practical approach *J. Neurosci. Methods* **250** 34–46
- [59] Zheng S et al 2021 Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 6881–90
- [60] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S 2020 End-to-end object detection with transformers *European Conf. on Computer Vision* (Springer) pp 213–29
- [61] Zhu X, Su W, Lu L, Li B, Wang X, Dai J 2020 Deformable DETR: deformable transformers for end-to-end object detection (arXiv:201004159)
- [62] Zhou L, Zhou Y, Corso J J, Socher R and Xiong C 2018 End-to-end dense video captioning with masked transformer *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 8739–48
- [63] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T et al 2020 An image is worth 16 × 16 words: transformers for image recognition at scale (arXiv:201011929)

- [64] Han K et al 2022 A survey on vision transformer *IEEE Trans. Pattern Anal. Mach. Intell.* **45** 87–110
- [65] Xiao G, Shi M, Ye M, Xu B, Chen Z and Ren Q 2022 4D attention-based neural network for EEG emotion recognition *Cogn. Neurodyn.* **16** 805–18
- [66] Eldele E, Chen Z, Liu C, Wu M, Kwok C-K, Li X and Guan C 2021 An attention-based deep learning approach for sleep stage classification with single-channel EEG *IEEE Trans. Neural Syst. Rehabil. Eng.* **29** 809–18
- [67] Zheng X and Chen W 2021 An attention-based bi-LSTM method for visual object classification via EEG *Biomed. Signal Process. Control* **63** 102174
- [68] Tao W, Li C, Song R, Cheng J, Liu Y, Wan F and Chen X 2020 EEG-based emotion recognition via channel-wise attention and self attention *IEEE Trans. Affective Comput.* **1**–1
- [69] Zhang D, Chen K, Jian D and Yao L 2020 Motor imagery classification via temporal attention cues of graph embedded EEG signals *IEEE J. Biomed. Health Inform.* **24** 2570–9
- [70] Wei J, Xiao W, Zhang S and Wang P 2020 Mild cognitive impairment classification convolutional neural network with attention mechanism 2020 *IEEE 16th Int. Conf. on Control and Automation (ICCA)* (IEEE) pp 1074–8
- [71] Song Y, Jia X, Yang L, Xie L 2021 Transformer-based spatial-temporal feature learning for EEG decoding (arXiv:210611170)
- [72] Cordonnier J B, Loukas A, Jaggi M 2019 On the relationship between self-attention and convolutional layers (arXiv:191103584)
- [73] Rahman M, Uddin M S and Ahmad M 2019 Modeling and classification of voluntary and imagery movements for brain–computer interface from fNIR and EEG signals through convolutional neural network *Health Inf. Sci. Syst.* **7** 1–22
- [74] Rahman M and Joadder M A M 2020 A space-frequency localized approach of spatial filtering for motor imagery classification *Health Inf. Sci. Syst.* **8** 1–8
- [75] Meghdadi A H, Stevanović Karić M, McConnell M, Rupp G, Richard C, Hamilton J, Salat D and Berka C 2021 Resting state EEG biomarkers of cognitive decline associated with Alzheimer's disease and mild cognitive impairment *PLoS One* **16** e0244180
- [76] Arevalillo-Herráez M, Cobos M, Roger S and García-Pineda M 2019 Combining inter-subject modeling with a subject-based data transformation to improve affect recognition from EEG signals *Sensors* **19** 2999
- [77] Singh D and Singh B 2020 Investigating the impact of data normalization on classification performance *Appl. Soft Comput.* **97** 105524
- [78] Mandrekar J N 2010 Receiver operating characteristic curve in diagnostic test assessment *J. Thoracic Oncol.* **5** 1315–6
- [79] Schirrmester R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggersperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
- [80] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces *J. Neural Eng.* **15** 056013
- [81] Ingolfsson T M, Hersche M, Wang X, Kobayashi N, Cavigelli L and Benini L 2020 EEG-TCNet: an accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces 2020 *IEEE Int. Conf. on Systems, Man and Cybernetics (SMC)* (IEEE) pp 2958–65
- [82] Park J, Jang S, Gwak J, Kim B C, Lee J J, Choi K Y, Lee K H, Jun S C, Jang G-J and Ahn S 2022 Individualized diagnosis of preclinical Alzheimer's disease using deep neural networks *Expert Syst. Appl.* **210** 118511
- [83] Babiloni C et al 2010 Cortical sources of resting EEG rhythms in mild cognitive impairment and subjective memory complaint *Neurobiol. Aging* **31** 1787–98
- [84] Gholamiangonabadi D, Kiselov N and Grolinger K 2020 Deep neural networks for human activity recognition with wearable sensors: leave-one-subject-out cross-validation for model selection *IEEE Access* **8** 133982–94
- [85] Hu X, Chu L, Pei J, Liu W and Bian J 2021 Model complexity of deep learning: a survey *Knowl. Inf. Syst.* **63** 2585–619
- [86] Pradeepkumar J, Anandakumar M, Kugathasan V, Suntharalingham D, Kappel S L, De Silva A C and Edussooriya C U S 2022 Towards interpretable sleep stage classification using cross-modal transformers (arXiv:220806991)
- [87] Xie J, Zhang J, Sun J, Ma Z, Qin L, Li G, Zhou H and Zhan Y 2022 A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification *IEEE Trans. Neural Syst. Rehabil. Eng.* **30** 2126–36
- [88] Kim D and Kim K 2018 Detection of early stage Alzheimer's disease using EEG relative power with deep neural network 2018 *40th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE) pp 352–5
- [89] Jeong H T, Youn Y C, Sung H-H and Kim S Y 2021 Power spectral changes of quantitative EEG in the subjective cognitive decline: comparison of community normal control groups *Neuropsychiatr. Dis. Treat.* **17** 2783
- [90] Rabin L A, Smart C M and Amariglio R E 2017 Subjective cognitive decline in preclinical Alzheimer's disease *Annu. Rev. Clin. Psychol.* **13** 369–96
- [91] Kverno K 2022 New treatment aimed at preventing Alzheimer's dementia *J. Psychosoc. Nurs. Ment. Health Serv.* **60** 11–14