IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 May 2023, accepted 8 July 2023, date of publication 12 July 2023, date of current version 19 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294618

RESEARCH ARTICLE

DICE-Net: A Novel Convolution-Transformer Architecture for Alzheimer Detection in EEG Signals

ANDREAS MILTIADOUS[1], EMMANOUIL GIONANIDIS[2], KATERINA D. TZIMOURTA[1], NIKOLAOS GIANNAKEAS[1], AND ALEXANDROS T. TZALLAS[1]

[1]Department of Informatics and Telecommunications, University of Ioannina, Kostakioi, 47150 Arta, Greece
[2]Independent Researcher, Atlanta, GA 30318, USA

Corresponding author: Alexandros T. Tzallas (tzallas@uoi.gr)

This work was supported in part by the Project "Immersive Virtual, Augmented and Mixed Reality Center of Epirus," Which is Implemented under the Action "Reinforcement of the Research and Innovation Infrastructure" under Grant MIS:5047221; and in part by the Operational Program "Competitiveness, Entrepreneurship and Innovation," Co-Financed by Greece and the European Union (European Regional Development Fund) under Grant NSRF 2014–2020.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Scientific and Ethics Committee of AHEPA University Hospital, Aristotle University of Thessaloniki, under Approval No. 142/12- 04-2023, and performed in line with the Declaration of Helsinki.

ABSTRACT Objective: Alzheimer's disease (AD) is a progressive neurodegenerative disorder that affects a significant percentage of the elderly. EEG has emerged as a promising tool for the timely diagnosis and classification of AD or other dementia types. This paper proposes a novel approach to AD EEG classification using a Dual-Input Convolution Encoder Network (DICE-net). Approach: Recordings of 36 AD, 23 Frontotemporal dementia (FTD), and 29 age-matched healthy individuals (CN) were used. After denoising, Band power and Coherence features were extracted and fed to DICE-net, which consists of Convolution, Transformer Encoder, and Feed-Forward layers. Main results: Our results show that DICE-net achieved an accuracy of 83.28% in the AD-CN problem using Leave-One-Subject-Out validation, outperforming several baseline models, and achieving good generalization performance. Significance: Our findings suggest that a convolution transformer network can effectively capture the complex features of EEG signals for the classification of AD patients versus control subjects and may be expanded to other types of dementia, such as FTD. This approach could improve the accuracy of early diagnosis and lead to the development of more effective interventions for AD.

INDEX TERMS Alzheimer's disease, deep learning, detection, EEG, Frontotemporal dementia, transformers.

I. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and one of the most frequently diagnosed dementia types among the elderly [1]. It is characterized by cognitive decline and behavioral changes, and its prevalence, along with the prevalence of other dementia types, is expected to rise as the population ages [2]. According to research, AD is the sixth leading cause of death in the United States and the only one among the top 10 causes still significantly increasing [3]. Over 50 million cases of dementia were reported in 2020, and it is estimated that the number of AD patients will reach 75 million by 2030 and 131 million by 2050 [4]. The AD prevalence ratio is the same among women and men and is at 1.4% for individuals aged 65-70 and 24% for individuals over 85 [5]. Regarding its symptoms, the disease's initial sign is hardness in recalling events of short-time memory. It progresses to problems that may include speech and orientation difficulties, mood swings, lack of self-care, and behavioral alterations [6]. Ultimately, the functions of the body systems

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Wang.

deteriorate, and the patient is finally led to death. Currently, there is no cure for AD, and available treatments only provide limited symptomatic relief.

To diagnose probable AD, the patient must meet specific clinical criteria such as postmortem confirmation of specific neuropathological changes (accumulation of neuritic plaques and neurofibrillary tangles containing hyperphosphorylated tau proteins) [7]. Nonetheless, emphasis has been given to early diagnosis and intervention as the number of individuals with dementia is increasing. To conform with the need for timely AD diagnosis, reliable biomarkers through structural Magnetic Resonance Imaging (MRI), molecular Positron Emission Tomography (PET) neuroimaging, and cerebrospinal fluid analyses have been employed in clinical practice for AD diagnosis [8]. However, these imaging tools' costly and time-consuming nature often leads to patients being diagnosed after having already shown significant neurodegeneration. Thus, the need for accurate prediction of AD (or other dementia types) future onset is great since it may accelerate the identification of high-risk individuals and support planning the overall treatment. Other faster and cheaper biomarker alternatives should be explored.

Brain activity alterations and network disruptions are key findings in neurodegenerative disorders such as AD or Frontotemporal dementia (FTD) [9]. Although there are various methods for measuring brain activity, they differ in spatiotemporal resolution and applicability. Techniques such as single-unit recordings provide high spatiotemporal precision but lack relevance due to being invasive. Methods such as functional MRI and Electroencephalogram (EEG) permit the assessment of brain activity in a non-invasive manner. However, EEG was not a widely employed tool since it provides low spatial but high temporal resolution and is prone to noise. Nonetheless, modern computational techniques such as Low-Resolution Electromagnetic Tomography (LORETA) provide estimation capabilities of the location of the underlying brain generators, thus promising increased spatial resolution [10] and techniques such as Independent Component Analysis (ICA) [11] and Artifact Subspace Reconstruction (ASR) [12], that perform external interference deduction (such as eye artifacts) or signal correction, have become computationally available during the last decade, making the EEG a promising tool in neurodegenerative disease diagnosis.

EEG is an affordable and widely accessible diagnostic tool that records the electrical activity alterations of the cerebral cortex by measuring the electrical postsynaptic potentials produced by brain neurons through scalp (or intracranial) electrodes [13]. In recent years, quantitative EEG has been established as a reliable clinical tool for the detection and assessment of brain diseases such as epilepsy [14] and Parkinson's disease [15] and has been tested on the evaluation of neurodevelopmental disorders and emotional conditions such as dyslexia [16] and stress [17]. Also, there has been growing interest in using EEG to detect and discriminate dementia variants, especially AD.

Due to the complex, non-stationary, and non-linear nature of the EEG signals, various efficient methodologies for feature extraction have been proposed for the different EEG problems. One of the most common ways to analyze the EEG signal is to decompose it into other frequency bands, such as delta, theta, alpha, beta, and gamma. Each frequency band represents a diverse range of electrical activity in the brain and is thought to be associated with different cognitive and physiological processes [18]. For example, the delta band (0.5-4 Hz) is often associated with deep sleep and the maintenance of bodily functions, while the alpha band (8-13 Hz) is thought to be related to attentional processes and relaxation [19]. Thus, the EEG signals are usually transformed to the frequency domain using a Fourier methodology such as Fast Fourier Transform (FFT) [20] or the Welch Power Spectral Density (PSD) [21] analysis or transformed to the time-frequency domain using decomposition such as the Discrete Wavelet Transform (DWT) [22] or the Empirical Mode Decomposition (EMD) [23]. Another way of analyzing the EEG signal that is becoming increasingly popular is coherence analysis and graph theory methods, as they provide powerful tools for investigating the functional connectivity and organization of the brain [24]. Coherence analysis is an approach that measures the degree of synchronization between different brain regions at specific frequency bands, providing information on the strength and patterns of functional connectivity. Graph theory methods are used to construct a network representation of the brain based on the coherence values, where nodes represent brain regions, and edges represent the strength of the coherence between them. By analyzing the topology and properties of the network, researchers can gain insights into the organization and dynamics of the brain, as well as its ability to process information [25]. Following the transformation or analysis of EEG data with one of the abovementioned methodologies, band power, entropy, fractal dimension, or statistical features are usually extracted to be fed to a Machine Learning framework for automatic detection, prediction, severity assessment, or evaluation of the given EEG task.

A wide variety of machine learning algorithms is used in the published literature of EEG classification studies regarding dementia detection. Traditional, well-established methodologies such as Support Vector Machines (SVM) [18], [19], k-Nearest Neighbors (kNN) [20], logistic regression [21] or Random Forests [22] still hold relevance in AD (or other types of dementia) classification. However, Deep Learning methodologies have become increasingly popular in classifying EEG signals in AD or further dementia research. Learning methodologies may extract and learn features from the raw data without the need for hand-crafted features or prior knowledge of the signal [23] (a concept known as Representation Learning [24]), or they can utilize the same feature extraction techniques as conventional machine learning does, as described in the previous paragraph [25]. Examples of deep learning models that have been used

for EEG classification in AD research include convolutional neural networks (CNN) [26], recurrent neural networks (RNN) [27], and autoencoders [5]. These models have shown promising results in accurately classifying EEG signals and identifying biomarkers for AD, providing insights into the disease pathology and potential targets for intervention.

However, the latest advancements in Natural Language Processing (NLP) based Deep Learning, namely the Transformers Neural Networks [28], have sparked a surge of interest in various subject areas beyond their original domain and have demonstrated superior performance to their counterparts in a wide range of fields such as image classification, speech recognition, biology, finance, and social media analysis. Their potential lies in their ability to process variable-length sequences of data and their performance scalability with big datasets. Recently, there has been growing interest in exploring the potential of transformers in other domains, including biomedical signal processing. In the EEG emotion recognition area, Guo et al. [29] proposed a Transformer methodology for the classification of emotion state EEG data and achieved 83.03% accuracy (ACC) at a three-class problem and outperformed most of the published methods in the same database. In another study related to the classification of motor-imagery EEG, a Transformer approach on unprocessed signal proposed by Xie et al. [30] was reported to achieve 83.31%, 74.44%, 64.22% ACC on two, three, and four class problems, respectively, outperforming in most cases other methodologies on the same dataset. Studies with such findings prove the effectiveness of the transformer encoder in EEG tasks and lead to the necessity of exploring their application in neurodegenerative EEG classification of AD and other dementia types.

The Transformer network architecture relies on the self-attention mechanism, which enables the model to attend to different parts of the input sequence and modify the output accordingly by computing a weighted sum of the input, where each weight depends on the similarity between each element in each sequence. The main idea behind the self-attention mechanism is that it allows the model to give more attention to the most relevant parts of the sequence (in the case of NLP, sentence). The Transformer architecture consists of an encoder block and a decoder block, each composed of multiple self-attention and Feed-Forward layers, residual connections, and layer normalizations [28]. Later advancements to the transformer methodology made it able to perform text or image classification tasks, with architectures such as Bidirectional Encoder Representations from Transformers (BERT) [31] (published in 2018) and Vision Transformer (ViT) [32] (published in 2021). These methodologies use the Transformer Encoder and the Class Token embedding, an extra sequence embedded in the input sequence that acts as a sequence-level representation of the classification task and aims to capture a contextualized representation of the entire sequence. The output of the encoder, or the CLS token alone, is fed to a Neural Network architecture for classification. Modifications to these methodologies have led to the widespread use of Transformers for classification problems in various domains beyond NLP and Computer vision, such as speech recognition [33], protein classification [34] and time-series analysis [35].

Various automatic methodologies that employ Machine Learning architectures have been proposed during the latest years to address the AD detection topic but are limited to the generalizability of their findings due to small sample sizes or no published dataset [27] or lack of proper validation methodology suitable for epoched datasets (for example reporting of extremely high-performance results because of biased testing due to the inclusion of same-subject data on training and test set by using k-fold validation on epoched and overlapping data) [5], [27], [36], [37]. Furthermore, there is a lack of studies focusing on incorporating the latest advancements in Deep Learning (namely the Transformer architecture) in EEG-based dementia detection studies [38]. That being said, there is a need for more accurate and efficient deep-learning diagnostic tools that can leverage the wealth of information provided by EEG recordings. Such diagnostic tools should have performance results that are properly validated and be reproducible (by promoting the availability of the datasets used) to ensure their reliability and efficacy in clinical practice.

In this study, we propose a novel methodology for classifying EEG signals from AD patients, combining a Convolutional Network architecture with a Transformer encoder on a dual feature/input scheme, namely Dual-Input Convolutional Encoder Network (DICE-net). Specifically, we extract two of the most promising biomarkers for AD detection, namely Relative Band Power (RBP) (literature has shown an increase in Theta/Alpha ratio in AD patients [39]) and Spectral Coherence Connectivity (SCC) (literature has shown decreased synchronization likelihood in AD patients [40]), and we express them in image-like representations (3d matrixes) which were fed in 2 parallel Convolution blocks. The Convolution blocks reduced the dimensions of these features extracting relevant information. The outputs were fed to 2 parallel Transformer Encoder blocks, along with randomly initialized CLS tokens used to conceptualize the sequence content. Finally, a Feed-Forward Neural Network (FFN) was trained to classify the instances as AD or healthy. The model was also evaluated using a group of FTD patients to explore its generalizability potential to other dementia types. It should be noted that this methodology is proposed considering AD classification performance optimization. In order to propose a scheme that best classifies FTD cases, more experimentation should be conducted.

II. MATERIALS AND METHODS

The analysis of the proposed methodology for the automatic classification of AD EEG signals versus Control EEG signals consists of four stages: data acquisition, signal denoising, feature extraction, and classification. These steps will be

analyzed individually in the following sections. Moreover, the well-established machine learning algorithms that were used to benchmark our proposed methodology's performance will be briefly presented.

A. DATABASE DESCRIPTION AND DATA ACQUISITION

To evaluate the proposed methodology, recordings from 88 participants were acquired from the 2nd Department of Neurology of AHEPA General University Hospital of Thessaloniki. 36 (13 males) of them were diagnosed with Alzheimer's disease (AD group), 23 (14 males) were diagnosed with Frontotemporal Dementia (FTD group), and 29 (11 males) were healthy subjects (CN group). The cognitive and neuropsychological state was evaluated by the international Mini-Mental State Examination (MMSE). MMSE score ranges from 0 to 30, with lower MMSE indicating a more severe cognitive decline. The duration of the disease was measured in months, and the median value was 25, with IQR range (Q1-Q3) being 24 - 28.5 months. Concerning the AD groups, no dementia-related comorbidities have been reported. The average MMSE for the AD group was 17.75 (sd=4.5), for the FTD group was 22.17 (sd=8.22,) and for the CN group was 30. The mean age of the AD group was 66.4 (sd=7.9), for the FTD group was 63.6 (sd=8.2), and for the CN group was 67.9 (sd=5.4).

The study was approved by the Scientific and Ethics Committee of AHEPA University Hospital, Aristotle University of Thessaloniki, under protocol number 142/12-04-2023. The investigations were carried out following the rules of the Declaration of Helsinki of 1975 (http://www.wma.net/en/30publications/10policies/b3/), revised in 2008. Informed consent was obtained from all subjects involved in the study.

For the recording of the EEG signals, a Nihon Kohden EEG 2100 clinical device was used, with 19 scalp electrodes (Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2) according to the 10-20 international system and two reference electrodes (A1 and A2) placed on the mastoids. Each recording was performed according to the clinical protocol, with participants being in a sitting position having their eyes closed. Before the initialization of each recording, the skin impedance value was ensured to be below 5kΩ. The sampling rate was 500 Hz with 10 μV/mm resolution. Each recording lasted approximately 13.5 minutes for AD group (min=5.1, max=21.3), 12 minutes for FTD group (min=7.9, max=16.9) and 13.8 for CN group (min=12.5, max=16.5). The dataset used in this study was made redistributable and publicly available at openneuro.org [41], [42].

B. SIGNAL PROCESSING AND DENOISING

The preprocessing pipeline of the EEG signals is as follows. First, a Butterworth band-pass filter 0.5-45 Hz was applied, and the signals were re-referenced to A1-A2. Then, the ASR routine [12] which is an automatic artifact reject method that can remove transient or large-amplitude artifacts [43], was applied to the signals, removing lousy data periods which exceeded the max acceptable 0.5-second window standard deviation of 17 (which is considered a conservative window). Next, the ICA method (RunICA algorithm) was performed, transforming the 19 EEG signals into 19 ICA components. ICA components that were classified as "eye artifacts" or "jaw artifacts" by the automatic classification routine "ICLabel" in the EEGLAB platform [44] were automatically rejected. It should be noted that, even though the recording was performed in a resting state, eyes-closed condition, eye artifacts of eye movement were still found in some EEG recordings.

TABLE 1. Demographic characteristics/Database description.

	Gender	Age	MMSE	CDR	Disease Duration in Months
AD	13/23	66.4 (7.9)	17.75 (4.5)	1 (0.54)	25 (9.88)
CN	11/18	67.9 (5.4)	30		
FTD	14/9	63.6 (8.2)	22.17 (8.22)	0.75 (0.26)	23 (9.35)

C. FEATURE EXTRACTION

Various EEG biomarkers have been extracted and employed in Machine Learning studies for automatic dementia diagnosis, automatic dementia progression assessment, or differentiation diagnosis between types of dementia, such as FTD versus AD. These may be time-domain features (statistical metrics) [45], spectral features such as relative brain band power ratios or absolute band power [5], time-frequency domain characteristics extracted from methodologies such as Discrete Wavelet Transform [37], complexity features such as permutation entropy or spectral entropy [21], coherence analysis features such as spectral coherence [46] and more. In this study, RBP and SCC have been extracted as features, as analyzed in the following paragraphs.

First, each recording was divided into 30-second time windows with 15 seconds overlap to create the pool of EEG signals that will be used for the classification task. Next, the following two features (described in sections II-C1. and 2.3.2.) have been extracted for T = 30 one-second periods, for each channel of the EEG signal (C=19), for each of the B=5 frequency bands, which describe the five brain rhythms of interest of the EEG signal. So, in total, two 3-dimensional arrays of dimensions [T,B,C] were generated for 30-second time-window.

The five frequency bands of B were defined as:
Delta: 0.5 – 4 Hz
Theta: 4 – 8 Hz
Alpha: 8 – 13 Hz
Beta: 13-25 Hz
Gamma: 25-45 Hz

1) RELATIVE BAND POWER

According to the literature, AD patients may exhibit changes in the EEG signal, such as reduced alpha power and increased

theta power [39]. A widely used approach for obtaining the Power Spectral Density (PSD) of a signal, such as an EEG signal, is the Welch method, sometimes referred to as the periodogram method [47]. The technique entails splitting the signal into overlapping segments and calculating each segment's squared magnitude of the discrete Fourier transform. A final estimate of the PSD is created by averaging the obtained values.

The mth windowed segment from a signal x is computed as:

$$x_m(n) \triangleq \mathrm{w}(n)\,\mathrm{x}(n+mR), \quad n = 0, 1, \ldots, \mathrm{M}-1, \mathrm{m} = 0, 1, \ldots, \mathrm{K}-1, \tag{1}$$

where R is the window hop size, K the number of available windows and $w(n)$ the Hamming window. The periodogram of the mth segment is calculated as:

$$P_{x_m,M}(\omega_k) = \frac{1}{M}\left|FFT_{N,k}(x_m)\right|^2 \frac{1}{M} \triangleq \left|\sum_{n=0}^{N-1} x_m(n)\, e^{-j2\pi nk/N}\right|^2 \tag{2}$$

Thus, the estimation of the PSD is calculated as the average of K segments:

$$\widehat{S_x^W} \triangleq (\omega_k)\frac{1}{K}\sum_{m=0}^{K-1} P_{x_m,M}(\omega_k) \tag{3}$$

where $FFT_{N,k}(x_m)$ is a Fast Fourier Transform (FFT), N the length of the FFT and is set to 256.

$\forall$ t $\in$ T, $\forall$ channel $\in$ C, the relative ratio of PSD of each band $\in$ B was calculated, resulting in a 3-dimensional matrix of [T,B,C], which constitutes the RBP feature.

2) SPECTRAL COHERENCE CONNECTIVITY

SCC (eq. 4) is used to quantify the synchronization of brain signals. It involves calculating the spectral coherence between each pair of signals, which measures the similarity of the frequency content between the two signals, and then averaging these values for each electrode.

$$SCC_x = \frac{1}{C}\sum\nolimits_{y=1}^{C} \frac{|S_{xy}|}{\sqrt{S_{xx} * S_{yy}}} \tag{4}$$

S_{xx} is the PSD of x(t) and S_{yy} is the PSD of y(t), S_{xy} is the Cross Spectral Density of signals x(t) and y(t), and $S_{xy}(f) = \lim_{T\to\infty}\frac{1}{T}\left[\hat{x}_T^*(f)\,\hat{y_T}(f)\right]$ exploiting Parseval's theorem. To calculate the PSD's of each signal, the signals were transformed in the Time-Frequency domain using a Morlet Wavelet Transform where:

$$w(\omega, t) = \left(\pi^{\left(-\frac{1}{4}\right)}\right) \times \left(e^{(i*\omega*t)} - e^{\left(-\frac{1}{2}*\omega 0^2\right)}\right) \times e^{\left(-\frac{t^2}{2}\right)}, \omega \in 2, 6, 10, 18, 35\} \tag{5}$$

And the wavelet transform is calculated as the convolution of x(t) with w(ω,t) as:

$$C(\omega, \tau) = \langle x, \mathrm{w}_{\omega,\mathrm{t}}\rangle = \int_{\mathrm{R}} x(t)\,\psi^*_{\omega,\tau}(t)\,dt \tag{6}$$

$\forall$ t $\in$ T, $\forall$ channel $\in$ C, the SCC of each band $\in$ B was calculated, resulting in a 3-dimensional matrix of [T,B,C].

D. CLASSIFICATION

This section describes the proposed DICE-net model, the algorithms employed to benchmark its performance, and the validation method used.

1) MODEL

The DICE-net model is structured as described. First, there are two parallel blocks, each receiving input $X_i \in \mathbb{R}^{Ba\times T\times B\times C}$, where Ba denotes the batch size of the Neural Network, and [T,B,C] the dimensions of the RBP and SCC features (one for each block).Each parallel block is consisted of a depthwise convolution layer, a positional embedding layer a class token embedding and a transformer encoder layer. Then a concatenation layer is applied, followed by a Feed-Forward Network (FFN) which determines the class of the input. Fig. 1 represents a flowchart of the proposed methodology and Table 2 represents the detailed architecture of the model.

Early stopping is performed to determine the best number of epochs for the model (the number of epochs represents the number of times each train sample will be fed to the model for training). Train, validation, and test sets are created, and after each epoch the performance of the validation set is evaluated. At the nth epoch, if the performance in terms of accuracy has not improved for 20 epochs, the training is stopped and the best model so far, is returned. The validation set is created by iteratively leaving out 6 subjects (randomly). The rest of the subjects are train-test splitted using Leave-One-Subject-Out (LOSO) validation. The best performing number for epochs is then selected.

Every hyperparameter optimization activity or ablation experiment has taken place regarding the AD-CN problem, and not considering the FTD dataset. The FTD-CN performance optimization was not the goal, but rather a comparison tool on how this methodology performs on other types of dementia.

2) CONVOLUTION LAYER

Given the input dimensions are [T,B,C], the total number of values in an input matrix is prohibitive for effectively training the neural network. In DICE-net architecture, a depthwise convolution layer, which is a convolution layer that allows the independence of the data at a given dimension of the input layer is employed to reduce the dimensions of the input array and extract spectro-temporal relationship information from the input matrix. Moreover, the convolution layer can capture frequency band associated relationships that extend further than 1 time-point, since the size of the kernel in

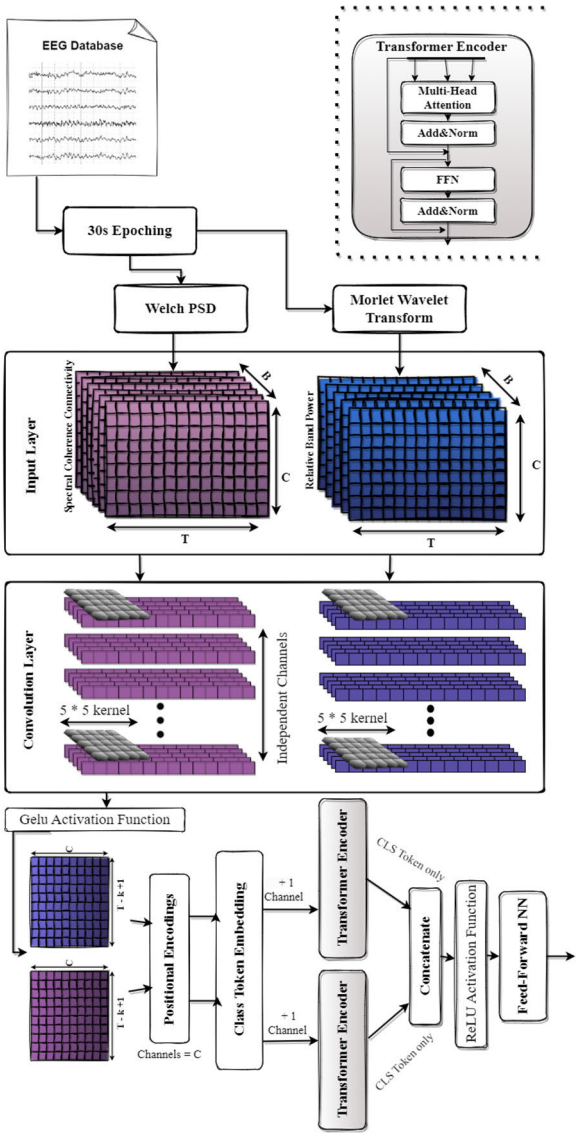

FIGURE 1. Flowchart of the proposed DICE-net methodology.

TABLE 2. The Architecture of the DICE-net Model.

Layer	Type	Input	Parameters	Output
A1, A2	Input			[B,30,5,19]
C1	Conv2d	A1	kernel=[5,5], stride=[1,1], groups=19	[B,26,19]
	Gelu	C1		
P1	PositionalEncoding1D	C1	channels=19	
CLS1	Parameter(Randn)	__		[1,26,1]
	torch.expand	CLS1	(expand to batch size)	[B,26,1]
TR1	torch.concat	CLS1 P1	dim=2	[B,26,20]
	TransformerEncoderLayer	TR1	num_layers=1, dmodel=2, nhead=2	
	drop channels	TR1	[:,:,0] (only CLS1)	[B,26]
C2	Conv2d	A2	kernel=[5,5], stride=[1,1], groups=19	[B,26,19]
	Gelu	C2		
P2	PositionalEncoding1D	CG2	channels=19	
CLS2	Parameter (Randn)	__		[1,26,1]
	torch.expand	CLS2	(expand to batch size)	[B,26,1]
TR2	torch.concat	CLS2 P2	dim=2	[B,26,20]
	TransformerEncoderLayer	TR2	num_layers=1, dmodel=2, nhead=2	
	drop channels	TR2	[:,:,0] (keep only CLS2)	[B,26]
FFN	torch.concat	TR1 TR2	dim=1	[B,52]
	LayerNorm	FFN	normalized_shape=52	
	Dropout		prob=0.2	
	Linear		in_features=52, out_features=24	[B,24]
	BatchNorm1d			
	ReLU			
	Dropout		prob=0.2	
	Linear			[B,1]
	Sigmoid			
Loss	BCEWithLogitsLoss			

the first dimension is higher than 1. So, in total, utilizing the convolution layer makes the data less computationally expensive for the next layers and reveal hidden frequency-band associations.

Specifically, the depthwise convolution layer consists of C convolutional kernels (contrary to the canonical convolution layer that consists of one kernel). Each kernel performs striding in a tensor of size [T,B] and assume [k,k] the dimensions of the kernel with stride = 1. No padding is added to the convolution input (zero-padding). The spatial dimensions of a convolution layer can be calculated as $Out(x, y) = (W\_in(x, y) - k(x, y) + 2P)/stride + 1$, where W_in is the dimensions of the input and k(x,y) the kernel size. Thus, the output of each kernel is [T-k+1,B-k+1]. The kernel size used was [5,5], so each kernel output was [26,1], and the output of the convolution layer was [26,1,C] = [26,1,19], flattened to [26,19]. The kernel weights for each kernel are trained with backpropagation using a Gaussian Error Linear Units (GELU) function. A GELU function can be thought of as a smoother ReLU function. In pyTorch, the GELU is calculated as:

$$GELU(x) = 0.5 \times x \times (1 + Tanh(\sqrt{\frac{2}{\pi}} \times (x + 0.044715 \times x^3)) \quad (7)$$

By using depthwise convolution layer, the output of each channel is processed independently, thereby eliminating interference between channels. If, instead, a canonical convolution would be employed, a 3-dimensional kernel would be required. Assume k the 3rd dimension of the kernel, the output values of a channel c_i would be affected by the values of channels $c_j, j \in [i-k, i+k]$. However, positional relationship

of the order of the channels does not exist, thus this would be wrong. So, a depthwise seperable convolution is preferred.

3) POSITIONAL ENCODING LAYER

Contrary to CNNs, or RNNs, a Transformer Encoder is unaware of the positional information of the input data. To model such positional relationships, a positional encoding layer is employed. Spatial information about the data's absolute or relative position is provided by the positional encoding layer, as first described in "Attention Is All You Need" [28]. Usually, in most transformer related architectures for natural language processing (NLP) or computer vision a positional encoding layer precedes an encoder.

Suppose $X \in \mathbb{R}^{T\times C}$, sequentially ordered data across the T axis. To express the positional relationships as data, the Positional Encodings (PE) are calculated as:

$$p_{k,2i} = \sin\left(\frac{k}{10000^{2i/d}}\right), p_{k,2i+1} = \cos\left(\frac{k}{10000^{2i/d}}\right) \quad (8)$$

where $k \in \{0, 1, \ldots, C-1\}$ and $i \in \{0, 1, .., T/2\}$. According to the study it was first proposed [28], positional encodings allow the model to learn relative positions, since any fixed offset P_{k+off} can be represented as a linear function of Pk. Positional Encoding Layer has no trainable parameters, meaning it does not require gradient computation during back-propagation, thus it does not get weight-modified during training.

4) CLASS TOKEN EMBEDDING

In vision transformers, the CLS token is a special token that is incorporated to the input sequence to capture the overall meaning of a sequence. It is typically used as a representation for the entire sequence for tasks such as image classification or object detection. During the training of the transformer encoder, the CLS token attends to important information from anywhere in the sequence and make use of the entire context of the image, or in this case the EEG feature representations.

In this implementation, an extra column named CLS token of size [T,1] was then appended to each of the two tensors, resulting in 2 [T,C+1] tensors. The values of the CLS token were initialized randomly from a canonical distribution.

5) TRANSFORMER ENCODER LAYER

The Transformer is a relatively new deep learning architecture that was first introduced in the Natural Language Processing domain and has applications in text and image classification. The encoder $f_\theta : \mathbb{R}^{n\times d} \rightarrow \mathbb{R}^{n\times d}$ is a block of the transformer model which reconstructs a collection of n objects to another collection of n objects and encodes the relational structure of the input as data in the reconstructed input. These objects are sequences; however the encoder is oblivious of the sequential positioning of the values of the objects. This is the reason a positional encoding layer is previously employed. A Transformer Encoder Layer may be comprised of several stacked Transformer Encoders (TE), and each TE output serves as the input for the next TE. Each TE is consisted of a Multi-Head Self-Attention (MSA) Layer with residual connection around it, followed by a FFN with residual connection.

A MSA layer is consisted of several Self-Attention heads. A Self-Attention head calculates the relationships between different parts of an input sequence in sentence, or in this case the relationships of the C input channels, representing each individual's importance in relation to the others. First, the input sequence is transformed to three linear projections namely query (Q), key (K) and value (V).

$$Q^{(h)}(x_i) = W^T_{h,q}x_i, K^{(h)}(x_i) = W^T_{h,k}x_i, V^{(h)}(x_i) = W^T_{h,v}x_i \quad (9)$$

A score matrix that determines the attention of each channel is calculated as:

$$\alpha^{(h)}_{i,j} = softmax_j\left(\frac{\langle Q^{(h)}(x_i), K^{(h)}(x_j)\rangle}{\sqrt{k}}\right), \quad \text{k the dimension of Q and K} \quad (10)$$

Finally, the MSA is calculated as:

$$MSA(\boldsymbol{x_i}) = \sum_{h=1}^{H} W_0 \sum_{i=1}^{C} \alpha^{(h)}_{i,j} V^{(h)}(\boldsymbol{x_i}) \quad (11)$$

where C is the input channels, H the number of Self-Attention heads and W_0 is trainable weights for each head. The output of the MSA is fed to an FFN, followed by a dropout layer with dropout probability of 0.1. The result of the FFN is of dimensions is of equal dimensions as the input of the TE. The activation function of the FFN is a ReLU function.

6) FEED FORWARD NETWORK

To perform the classification of the inputs, a FFN is utilized. Assume [T,C+1] the output dimensions of the TE layers, and 2 parallel TE layers were employed, each for one of the inputs. All channels except the CLS token(s) channel are discarded, and the remaining T values of each array are concatenated in a 2*T array, normalized, and then fed into the FFN, which is consisted of 1 input layer (52 neurons), 1 hidden layer of 24 neurons and the output layer. A Dropout layer with a dropout probability of 0.2 is added before each Linear layer in the architecture. After each Linear layer there is a Batch Normalization layer. The activation function of the hidden layer is a ReLU function.

A sigmoid cross entropy loss function was used as the loss function. Batch size was set to 32, learning rate was 0.001 and L2 regularization weight decay was set to 0.01.

7) ABLATION EXPERIMENTS

1) NO-TRANS: Removed TE and PE. The results of CNN were directly fed to FFN.
2) E-DICE: Early concatenation. The concatenation of the inputs happened before exactly after the CNN layers. Only one CLS token was generated.
3) 2-DICE: Two stacked encoder layers.

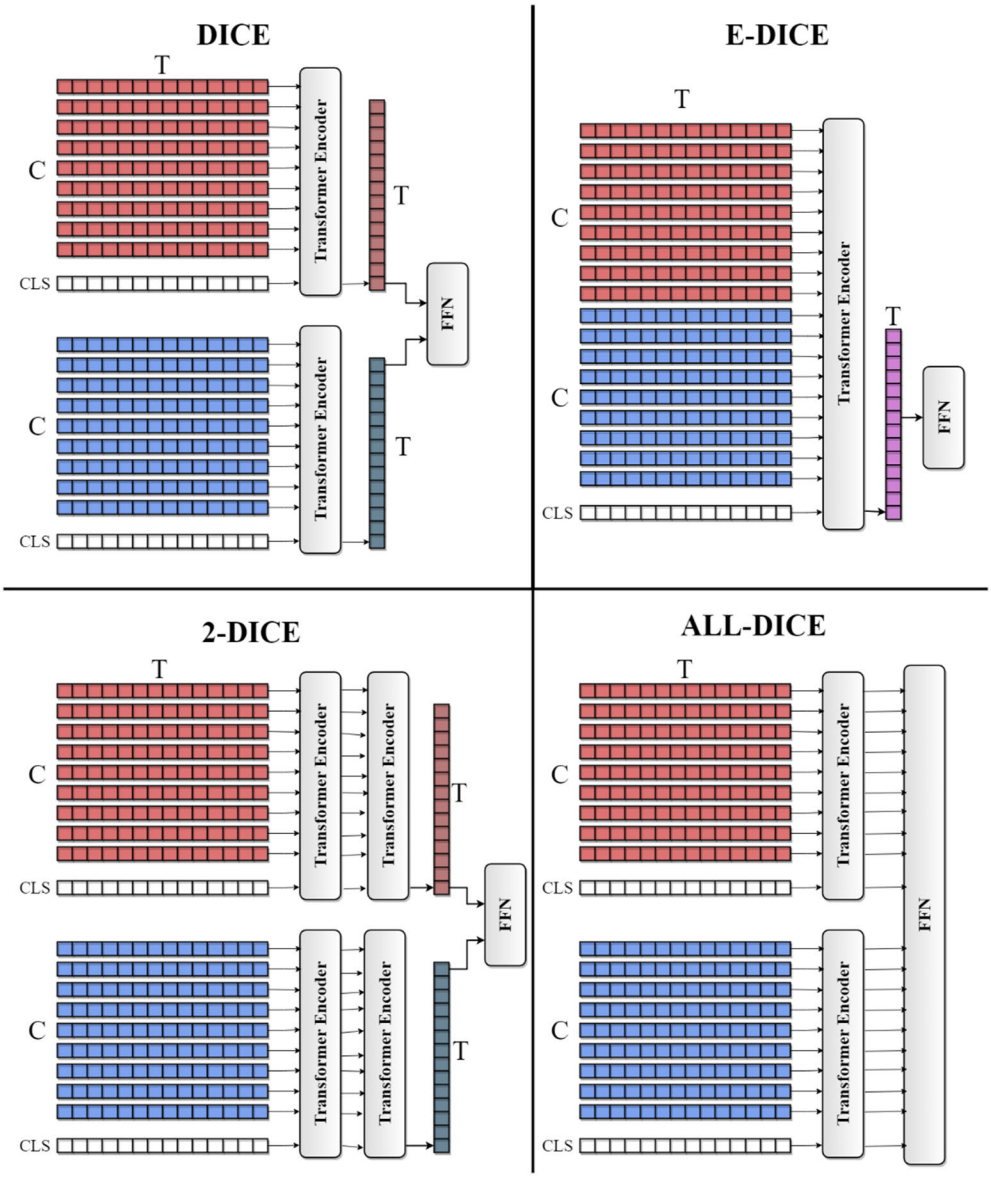

FIGURE 2. Different ablation configurations.

4) M-CLS: The CLS token is not randomly initialized but rather initialized by the mean values of each row.
5) ALL-DICE: No channels are removed prior to the FFN layer. Instead, the values of all channels are fed.
6) ALL-E-DICE: Early concatenation, no channels removed prior to FFN.

8) COMPARISON ALGORITHMS

To validate the robustness of the proposed methodology, the following benchmarking algorithms have been employed, and their performance metrics are reported in the Results section: 1) k-Nearest Neighbors with Principal Component Analysis (PCA-kNN), 2) XGBoost, 3) LightGBM, 4) CatBoost, 5) Support Vector Machines with PCA (PCA-SVM), 6) Multilayer Perceptron (MLP). All gradient boosting algorithms have been hyperparameter optimized with Hyperopt [48], which is a python package for machine learning hyperparameter optimization. MLP had 1 hidden layer of 96 neurons (layer structure: 190-96-1). For k-NN, all k values from 1 to 12 were tested and 5 was found to achieve the best ACC. To train and test these algorithms, the same feature extraction techniques were applied. However, these algorithms do not support 3d matrix input, so the training and test set is required to be in a conventional format where each row represents a subject or an observation. So, a time-window division of 15 seconds was applied, and 190 features were extracted (RBP [5 bands * 19 channels] + SCC [5 bands * 19 channels]).

Furthermore, state of the art deep learning architectures designed to classify raw EEG signal were employed.

These architectures were EEGNet [49], EEGNetSSVEP [50], DeepConvNet, ShallowConvNet [51].

9) VALIDATION METHODOLOGY

The Leave-One-Subject-Out (LOSO) validation method has been employed for the performance evaluation of the model. In this method, all the feature matrixes regarding one subject are left out as test set and all the other subjects form the training set. This is repeated one time for each subject, and then the weighted average performance results are presented. Thus, for a given problem, the EEG recordings of all subjects except one are used as the training set and the left out subject EEG recordings are used for testing. This procedure is repeated iteratively for all subjects and a total confusion matrix is created. The performance metrics are then calculated from this confusion matrix.

E. EXPERIMENTAL SETUP

The recording step of the experiment was described in the Database Description and Data Acquisition section. The preprocessing step of the experiment was implemented in EEGLAB Matlab (2021a) environment [44]. The time-frequency transforms and the feature extraction steps were implemented in Python 3.10 using the MNE library. The Deep Learning model was implemented, trained, and evaluated in Python 3.8 using the PyTorch library [52] and the implementation and evaluation of the comparison algorithms was implemented using the Scikit-Learn library. The models were trained on a RTX 3060 Ti GPU with CUDA 11.7 version. The computational complexity of the DICE-net algorithm was 137 GFlops. The trainable parameters and computational complexity of the DICE-net algorithm and each ablation experiment are presented in Table 3.

TABLE 3. Computational complexity of DICE-net and ablation models.

Model	N_params(M)	FLOPs(G)
DICE-net	170.5	137.4
ALL-DICE	368.6	140.6
NO-TRANS	18.8	1.42
2-DICE	338.7	274
E-DICE	163.5	133.9
M-CLS	170.5	137.4
ALL-E-DICE	357.6	130

III. RESULTS

The importance of the selection of the specific features should be first evaluated by visualizing each feature across the different groups. The significance of RBP as a feature can be observed in Fig 3(a), which presents the PSD of a healthy subject (1st), a subject with AD in the early stages with MMSE=16/30 (2nd), and a subject with severe AD with MMSE=4/30 (3rd). Reduction in alpha power (8-13 Hz) can be observed, as the severity of the AD increases.

Moreover, a scalp heatmap representation of the PSD of the different frequency bands for each group is presented in Fig 3(b). Each column represents a group (AD, CN, FTD) and each row represents a frequency band. The min and max value of each colormap may differ, but the range of all colormaps is same and equal to 7 uV^2/Hz. Considerable differences can be observed between AD and CN heatmaps across all frequency bands. On the other hand, FTD-CN discrimination appears harder, based on visual inspection of the heatmaps.

Regarding the SCC feature, Fig. 4(a) represents the spectral connectivity calculated, averaged across all subjects for each group. Each row represents a group (AD, CN, FTD) and each column represents a frequency band. Each graph is a rectangle heatmap (upper and lower triangular matrixes are symmetric) that every cell (X,Y) expresses the spectral connectivity of the electrode X with the electrode Y. Fig. 4(b) represents the spectral connectivity calculated, averaged across all subjects and across each electrode for each group, which technically is the averaged SCC feature. It can be visually observed that AD group has lower delta connectivity than CN group in multiple brain locations. This finding is supported by the literature [53] and indicates the importance of spectral connectivity as a feature. Reduced delta connectivity is also observed for the FTD group.

The size of the dataset should be noted here. In total, the AD group consisted of 953 sets of 3-dimensional matrixes (PSD + SCC matrix), the FTD group consisted of 541 sets and the CN group consisted of 788 sets.

Multiple ablation experiments were conducted and hyperparameters have been evaluated, to present the methodology that has achieved the best results. The comparison of the performance of the different ablation experiments was performed in regards with the LOSO accuracy. Moreover, to find the optimal number of epochs for each ablation experiment, an evaluation-train-test split has been employed on top of the LOSO validation method. Specifically, a P groups split was iteratively performed, P being the integer number of groups that round to the 1/6th of the dataset. The P groups were left as validation set, and the rest 5/6th of the dataset was evaluated with LOSO. Early stopping was employed, meaning that if the accuracy of the model is not employed for 20 consecutive epochs, the training is stopped to avoid overfitting, and the best achieved accuracy so far is kept. For the statistical evaluation of the performance metrics, the training of each model is repeated 10 times, and the difference in the performance of the proposed DICE-net is found to be statistically important (independent samples t-test, p-value $<$ 0.05), in comparison to all the other methods at almost all the metrics. On the Tables 4-6, the star symbol (*) indicates statistically important difference (independent samples t-test, p-value $<$ 0.05) in the particular metric in regards with the DICE-net.

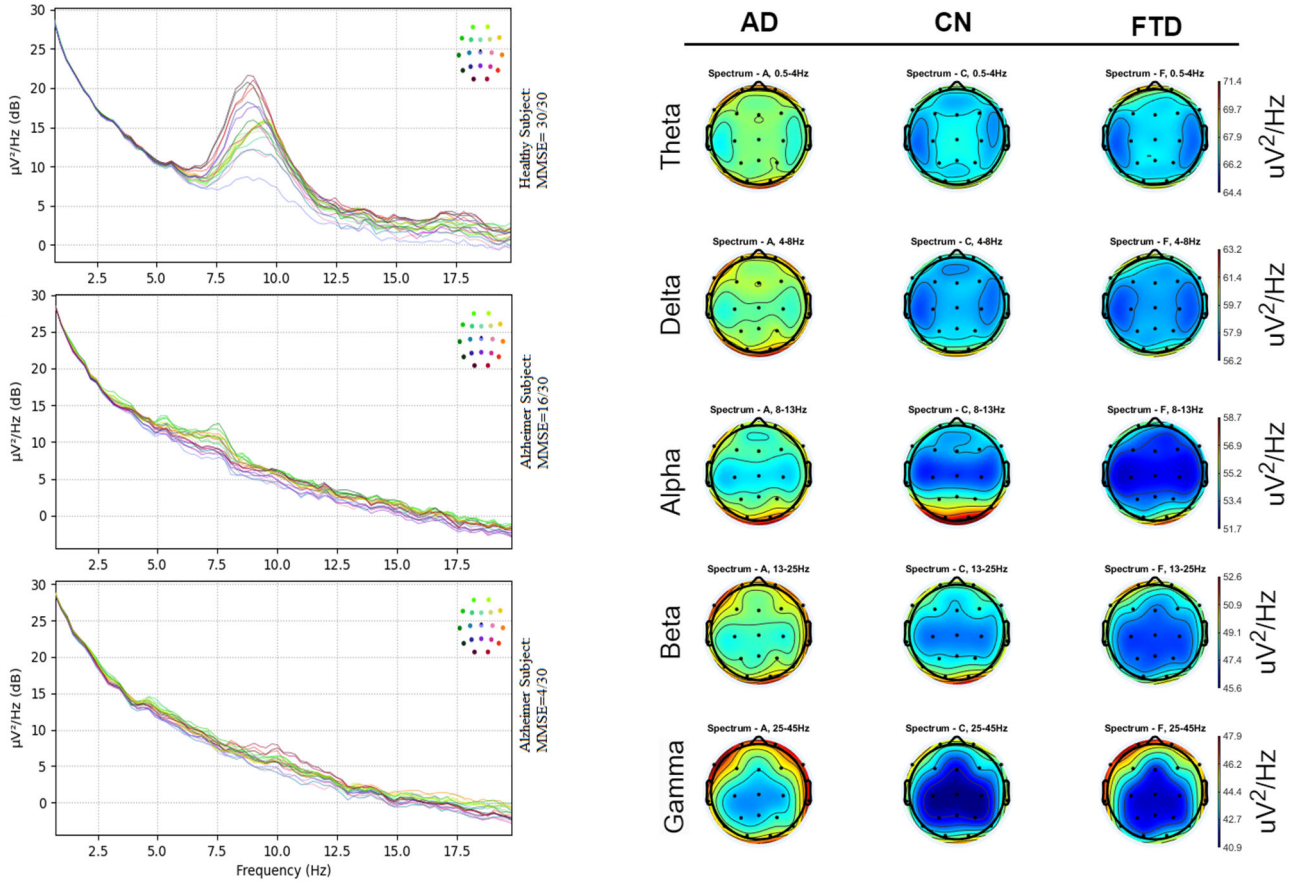

FIGURE 3. a) (left) PSD of a severe AD case (bottom), a mild AD case (middle) and a healthy subject (top). b) (right) scalp heatmaps of PSD across 5 frequency bands, averaged across groups AD, CN, FTD.

Table 4 presents the performance metrics of the different ablation experiments, as well as the proposed methodology in terms of accuracy (ACC), sensitivity (SENS), specificity (SPEC), precision (PREC) and F1 score for the AD-CN problem. The ACC of DICE-net reached 83,28%, followed by E-DICE, M-CLS and 2-DICE with ACC of 80,75%, 80,7% and 80,41% respectively and not statistically important differences between them. The NO-TRANS model, that does not utilize a transformer layer achieved ACC of 79,12%, followed by the transformer models that do not drop the channels prior to FFN, ALL-E-DICE and ALL-DICE with ACC 78,84% and 78% respectively. The training epochs that need to be utilized for each model to achieve its best performance may vary. The DICE-net model achieves its best performance at around 80 epochs.

To evaluate the effectiveness of the proposed methodology, other state of the art and/or well-established machine learning algorithms have been employed, as presented in Comparison Algorithms section. Table 5 presents the performance results of these algorithms for AD-CN classification. Fig. 5 (left) presents the ROC curves and the Area under ROC of each of the comparison algorithms along with DICE-net for the AD-CN problem. Furthermore, the classification capabilities of DICE-net along with the comparison algorithms for the FTD-CN classification problem has been examined and presented in Table 6. Fig. 5 (right) presents the ROC curves and the Area under ROC of each of the comparison algorithms along with DICE-net for the FTD-CN problem.

TABLE 4. Performance metrics of DICE-net and ablation models in AD-CN problem.

AD/CN	ACC	SENS	SPEC	PREC	F1
NO-TRANS	79.12% *	79.87%	78.29% *	80.46% *	80.17% *
E-DICE	80.75% *	76.49% *	85.91%	86.78% *	81.31% *
2-DICE	80.41% *	74.39% *	87.69%	87.35% *	80.61% *
M-CLS	80.70% *	82.58%	78.42% *	82.23% *	82.40% *
ALL-DICE	78.00% *	79.32%	76.39% *	80.25% *	79.78% *
ALL-E-DICE	78.84% *	80.14%	77.25% *	81.01% *	80.22% *
DICE-net	**83.28%**	**79.81%**	**87.94%**	**88.94%**	**84.12%**

Fig. 6 is utilized to visualize the individual predictions on each participant. Each graph represents the classification performance of the DICE-net algorithm and the comparison algorithms. Each dot represents the classification accuracy

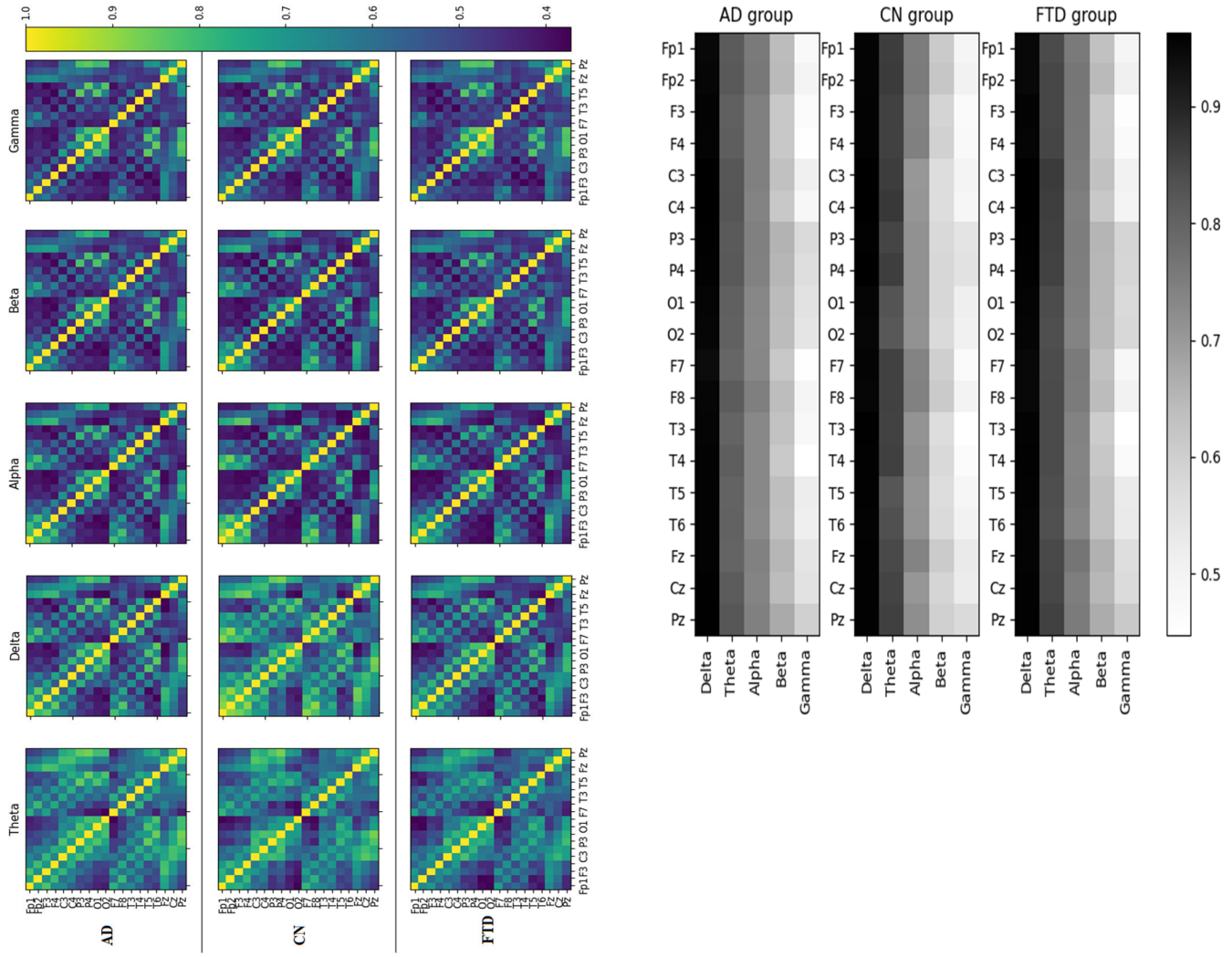

FIGURE 4. a) (left) Spectral Coherence Correlation (SCC) heatmaps for each group (AD, CN, FTD), for each frequency band. Each cell (X,Y) represents the spectral correlation of electrode X with electrode Y, averaged across each group. b) (right) SCC averaged across electrodes, so that each cell represents the average SCC of each electrode with all other electrodes.

TABLE 5. Performance metrics of DICE-net and comparison algorithms in AD-CN problem.

AD/CN	ACC	SENS	SPEC	PREC	F1
LightGBM	76.28% *	76.08% *	76.52% *	79.67% *	77.83% *
XGBoost	75.53% *	76.08% *	74.87% *	78.55% *	77.29% *
CatBoost	75.39% *	75.50% *	75.25% *	76.68% *	77.05% *
SVM+PCA	73.75% *	71.51% *	76.46% *	78.60% *	74.89% *
PCA-kNN	72.52% *	70.30% *	75.19% *	77.41% *	73.69% *
MLP	73.69% *	72.98% *	74.81% *	77.80% *	75.31% *
DICE-net	**83.28%**	**79.81%**	**87.94%**	**88.94%**	**84.12%**

TABLE 6. Performance metrics of DICE-net and comparison algorithms in FTD-CN problem.

FTD/CN	ACC	SENS	SPEC	PREC	F1
LightGBM	69.13% *	51.57% *	81.54%	65.72%	57.79% *
XGBoost	69.22% *	52.02% *	81.73%	65.71%	57.44% *
CatBoost	68.66% *	47.41% *	83.25%	66.02%	55.19% *
SVM+PCA	70.93% *	45.85% *	86.21%	75.26%	56.98% *
PCA-kNN	67.80% *	41.50% *	85.85%	66.82%	51.20% *
MLP	69.98% *	53.60% *	81.22%	66.21%	59.24% *
DICE-net	**74.96%**	**60.62%**	**78.63%**	**64.01%**	**62.27%**

for a certain subject and the color of the dot represents the class of the subject. It can be observed that the area of the DICE-net algorithm on the upper side of the diagram is larger, and the misclassified subjects are far less than any of the other classifiers compared to other indicating the superiority of our methodology.

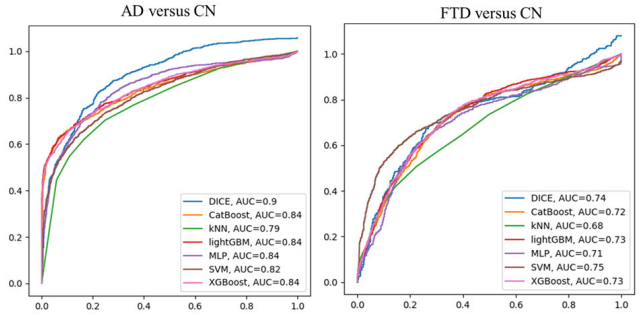

FIGURE 5. ROC curves of DICE-net and comparison algorithms for AD-CN and FTD-CN classification.

A variety of state-of-the-art deep learning architectures designed for EEG signal classification were also examined, in order to compare the effectiveness of this algorithm. These architectures were: EEGNet, EEGNet_SSVEP, DeepConvNet and ShallowConvNet. All these models have in common that they take raw EEG signal as input and not a feature vector. Epochs of 4 second with 2 second overlap were used with 128Hz sampling rate (also 250 and 500 were tested), from the same dataset. Smaller time windows, specifically 4 seconds in duration, were chosen for these algorithms due to their input being raw signal, resulting in a considerably larger size.For training, over 200 epochs were used, to make sure that training set accuracy was over 95%. However, none of these algorithms managed to classify the instances correctly, neither in the AD-CN nor in the FTD-CN problem. Table 7 contains the performance results of these algorithms using LOSO validation in terms of ACC, SENS, SPEC, PREC, F1.

TABLE 7. Performance results of state-of-the-art methodologies that use raw EEG signal as input, for the AD-CN and FTD-CN classification problem with LOSO validation.

AD/CN	ACC	SENS	SPEC	PREC	F1
EEGNet	41%	47.20 %	37.67 %	37.89 %	42.04 %
EEGNetSSVEP	51.46 %	56.78 %	45.39 %	47.65 %	51.82 %
DeepConvNet	54.21 %	45.43 %	57.59 %	48.71 %	47.01 %
ShallowConvNet	42.18 %	46.50 %	41.11 %	49.74 %	48.07 %
FTD/CN	**ACC**	**SENS**	**SPEC**	**PREC**	**F1**
EEGNet	46%	42.20 %	57.46 %	45.21 %	43.65 %
EEGNetSSVEP	61.46 %	53.51 %	75.00 %	51.40 %	52.43 %
DeepConvNet	64.21 %	62.41 %	37.05 %	58.14 %	60.20 %
ShallowConvNet	46.38 %	42.58 %	53.21 %	42.37 %	42.47 %

To explore which channels, and therefore which brain areas were most significant for the discrimination of AD-CN and for FTD-CN, the magnitude of the absolute value of the convolution layer weights was examined. Theoretically, larger absolute kernel weights indicate higher importance in the classification, due to back-propagation. Fig. 7 represents 2-d heatmap representations of the scalp, where hotter (red) colors mean higher absolute magnitude of weights and higher importance in the classification. The results have been normalized in 0-1, thus the bluest is the less significant area (although this does not indicate lack of significance) and the most red is the most significant areas. Fig. 7 represents the average values obtained after a complete LOSO iteration. Higher importance of the RBP feature in comparison with the SCC feature can be observed in both classification problems. Also, for the AD discrimination, the electrodes T5, O1, O2, T4, F8, mainly on the temporal and occipital lobe had been given the greater attention from the DICE-net model. Respectively, for the FTD discrimination, the frontal Fp1 and Fp2 and the temporal T3 and T4 electrodes have been given the greater attention from the model, as expected.

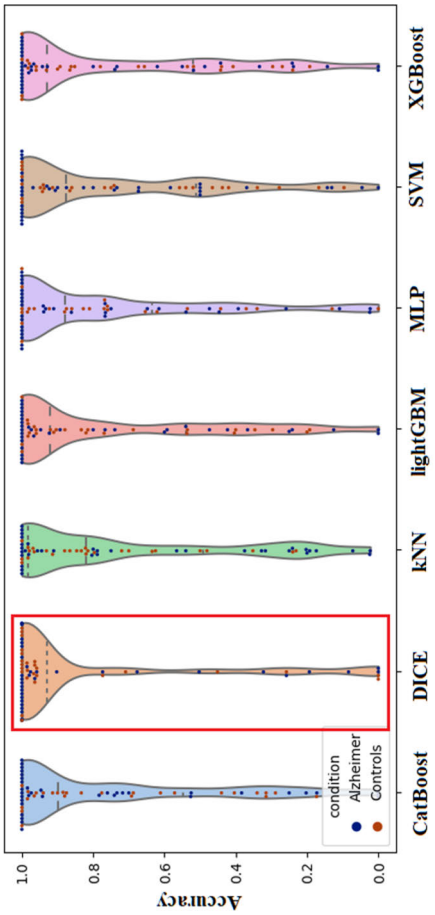

FIGURE 6. Violin plots of the distribution of accuracies for each subject prediction. The width of the violin indicates the density of scores at each value. Individual dots represent a single subject classification accuracy.

IV. DISCUSSION

This work proposed a novel convolution-transformer-based Deep Learning architecture to discriminate clinical dementia EEG. Specifically, the methodology is proposed and optimized for AD detection. In order to examine the

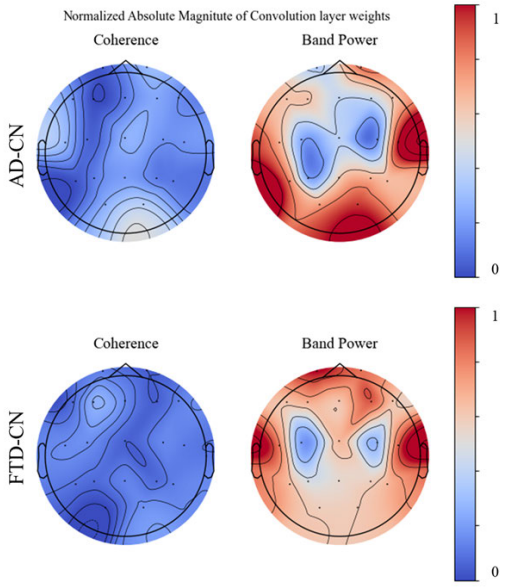

FIGURE 7. Normalized Absolute Magnitude of Convolution layer weights for DICE-net in AD-CN (top) and FTD-CN (bottom) classification.

generalizability of this methodology to other dementias, the methodology has also been tested with a FTD dataset. The methodology consists of 3 parts. In the first part, the raw EEG signals that are acquired following a strict protocol from a neurology department were preprocessed using the ASR routine to ensure that no corrupted data may be used for the training of the model, and the ICA algorithm to ensure the elimination of brain and jaw artifacts. In the second step, the recordings were divided into 30s time-windows. Two of the most well-established methodologies were employed for the frequency domain transformation of the signal and the feature extraction. Specifically, the Welch method, which is a frequency-domain transformation of the EEG signal that uses sliding windows to calculate the Fast Fourier Transform and then averages them to achieve a smoother frequency curve than FFT on all the signals, was employed to extract the Relative Band Power of each frequency band of the signals and a Wavelet Transform using the Morlet Wavelet was employed to extract the average spectral coherence of each channel for each frequency band. In the third step, a Deep Neural Network consisting of two parallel Convolution-Transformer blocks leading to a Feed-Forward NN was trained. Its discriminative capabilities were evaluated with LOSO cross-validation.

Multiple studies have addressed the problem of the detection of various types of dementia in EEG signals using machine learning methodologies [54]. The most advanced methodologies usually propose a Deep Neural Network scheme that first processes the time-domain signal through a time-frequency transform such as a Wavelet Transform [55] and then utilizes the capabilities of a Convolution Layer for information extraction and dimensionality reduction and/or Neural Network architectures such as autoencoders [5] or FFNs. However, Transformers Networks perform exceptionally well in dealing with long-range dependencies and in recognizing patterns in sequences of data. This is a significant advantage in the context of EEG analysis over traditional convolutional neural networks (CNNs), because EEG signals are often highly correlated over long time intervals, and capturing these correlations is critical for accurate classification of Alzheimer's disease. Still, they have yet to be employed in the AD EEG detection problem. To the best of our knowledge, only one study has used a transformer encoder on a Raw-EEG framework for Mild Cognitive Impairment (which is the prodromic state of AD) detection [38]. The combination of Convolution-Transformer layers in a classification task has been evaluated in other EEG areas, such as emotion recognition [29] and the results were promising. The novelty of our methodology is that is the first to introduce a Convolution-Transformer combination in EEG AD detection, which significantly outperforms other state-of-the-art algorithms performed on the same dataset.

From a medical perspective, the proposition of a novel transformer architecture for classifying AD in EEG signals is highly significant. Automated early detection of AD with minimal medical attendance is essential for prompt treatment and management and EEG signals have been widely used in medical research for neurological disorder diagnosis. Although the most used imaging tools for the detection of AD are MRI and PET, EEG does provide a faster, cheaper, and more portable alternative. One of the most notable EEG changes in AD that can be adequately captured by the proposed architecture is the reduction of alpha and beta waves, the reduction in amplitude which is believed to be related to decreased cortical activity in the brain, the increase in theta waves and the decreased synchronization among brain regions that may reflect the progressive loss of neuronal connections in the brain. Thus, should these changes be detectable in the early stages of the disease through a machine learning architecture, the EEG would have the potential to be used as a biomarker for the disease.

From a technical perspective, transformers have several advantages over traditional deep learning architectures such as Recurrent Neural Networks. The main advantage is the attention mechanism which allows the model to dynamically focus on the most relevant features in the input data. This is particularly useful in EEG signals, where different frequency bands and electrodes may contain different information relevant to the classification task. Another advantage is the ability to scale to large datasets. Transformer architectures have been shown to outperform the, until recently state-of-the-art, convolutional architectures in the classification of images when a large enough dataset was provided. Hence, in the medical domain where accurate predictions are crucial, transformer architectures may be the solution since they can take advantage of huge EEG datasets. Furthermore, the motivation of deciding to employ a transformer architecture for the EEG Alzheimer detection problem lies on two key

factors: firstly, the absence of previous methodologies that have utilized transformers for EEG detection, making it an unexplored area with significant potential for innovation; and secondly, the inherent ability of transformers to effectively capture long-range dependencies, which aligns well with the complex temporal relationships present within EEG signals. By leveraging this synergy, we aim to enhance the accuracy and efficacy of Alzheimer's disease detection using EEG data. In conclusion, the importance of this research is prominent from a both medical and technical perspective.

In order to use the transformer encoder we had to find a way to take advantage of its capabilities of detecting dependencies of different sequences (words) in a sentence that are widely used in Natural Language Processing. However, modifications to the original transformer that deal with image classification have been already proposed and are called vision transformer (ViT) [32]. The main idea behind a vision transformer is to split the image into patches, where each patch represents a word. Then add positional information to the patches, with a positional encoding layer and finally add another patch or word namely the CLS token that will learn the semantics of all the other words representing a sentence, or in the ViT case, an image. Thus, to create image-like input for the transformer, the feature extraction procedure changed and instead of the conventional ''1 row – 1 sample'', each sample was constructed as a 3d matrix. Finally, the capabilities of the convolution layer were exploited to reduce the dimensionality of the 3d matrix and acquire pattern information.

The performance of the DICE-net methodology for the AD-CN problem was compared to other, state-of-the-art ensemble classifiers such as CatBoost, XGBoost, and LightGBM and was found to be significantly better (7% higher accuracy, 6,29% higher F1 score, p=0.05, from the second best, LightGBM). Moreover, the performance for the FTD-CN problem was supplementarily evaluated, compared with the same algorithms, and found to be statistically better than the second-best SVM in terms of accuracy (4% higher) and the second-best in terms of F1 score, MLP (3% higher). Last, state-of-the-art deep learning architectures specifically designed to get raw EEG signal as input were tested, such as EEGNet, DeepConvNet, ShallowConvNet. However, they did not achieve to classify correctly neither the AD-CN problem, nor the FTD-CN problem. One possible explanation for the poor performance is that these methodologies do not perform feature extraction on the raw EEG signals, and therefore may not be able to effectively capture the relevant information in the data. This can lead to issues with overfitting, as well as reduced classification accuracy, especially with such a small dataset. It is possible that these methodologies could perform better if the size of the training dataset was significantly larger. Thus, no conclusions can be made regarding the comparison of the performance of these raw EEG input methodologies and our proposed methodology.

The performances of various ablation experiments that were conducted to evaluate the best model to propose were also reported. The ablation studies demonstrate the importance of the encoder layer for enhancing the predicting capabilities of the model since the increase in the performance between the NO-TRANS model and the DICE-net model is over 4% (statistically important difference p=0.05). Furthermore, the importance of the Class Token Embedding is also established, since it can learn to gather and attend to the important information of all the other channels. Comparing the DICE-net model with its no channel dropped counterpart ALL-DICE, over 5% increase in ACC is observed, meaning that the ability of the CLS token to keep important information allows us to drop 19/20 of the information that would be fed to the FFN, thus significantly reducing the size of the input layer and achieving better performance with less overfitting risk.

To evaluate which features and which brain areas were most important for the classification task we utilized the absolute magnitude of the convolution kernels as a marker of attention to each channel. The RBP feature was proven to be more important than the SCC feature. This might be the case because the decreased synchronization in the brain is evident in the late stages of AD, and alpha-theta wave alterations are easier to detect. According to the literature, AD primarily affects the hippocampus, amygdala, and neocortex regions [56]. The model mainly focused on the electrodes located onto the occipital, temporal, and frontal regions of the brain. Nonetheless, the exact location of the affected brain activity is difficult to be located without further information from other EEG source localization or phase synchronization techniques. Further analysis of the EEG signals using such techniques and statistical comparison with healthy signals may indicate the source localization specifics. However important this information could be, it would probably not be useful as a channel elimination indicator for this EEG DICE-net methodology, since the convolution and transformer layers automatically focus on important channels.

The classification performance of the DICE-net methodology was also evaluated on the FTD database, and the absolute magnitude of the convolution kernels was also reported. FTD is a group of progressive neurodegenerative disorders that primarily affect the frontal and temporal lobes of the brain and is characterized by progressive focal frontal and temporal lobe atrophy [57]. DICE-net focused specifically on the frontal and temporal regions of the brain, as can be noticed from Fig. 7 validating the ability of the methodology to focus on useful information. Similar to the AD-CN case, the algorithm exhibited less interest in the spectral coherence features.

Studying previous works, not many studies have been published in recent years that propose an EEG machine learning architecture for the detection of AD or Mild Cognitive Impairment (MCI) that reports its classification accuracy using LOSO validation. In the following Table 8 recent studies that address the same problem have been reported. Most methodologies perform acquisition or use published databases of resting state close eyes recordings [20], [57],

TABLE 8. Related studies comparison.

Study	Year	Cohorts	Stimuli	Methodology	Performance
Safi et al. [20]	2021	30 AD 35 CN	N/A	Entropy, Hjorth Parameters, SVM	ACC=81%, SENS=69.8%, SPEC=83.5%
Khatun et al. [61]	2019	8 MCI 15 CN	Auditory	ERP, SVM	ACC=87.9%, SENS=84.8%, SPEC=95%
Dogan et al. [60]	2022	12 AD 11 CN	Resting State	Graph-Based Feature extraction, Tunable Q-Wavelet Transform, kNN	ACC=92.01%, SENS=97.75%, SPEC= 84.03%
Miltiadous et al.[45]	2021	10 AD 8 CN	Resting State	Spectral & Temporal & Nonlinear Features, Random Forests	ACC=78.85%, SENS=82.4%, SPEC=74%
Ruiz-Gomez et al. [59]	2018	74 (AD + MCI) 37 CN	Resting State	Spectral & Nonlinear features, MLP	ACC=78.43%, SENS=82.35%, SPEC=70.59%
Araujo et al. [58]	2022	11 AD 8 MCI 11 CN	Resting State	Nonlinear features, SVM	AD-CN ACC=81%, MCI-CN ACC=79%
Lopes et al. [26]	2023	34 AD 20 CN	Resting State	Modulation Spectrum, CNN, SVM	ACC=87.3% F1=84.6%
This work	2023	36 AD 29 CN	Resting State	RBP, SCC, Dual-Input-Convolutional-Encoder	ACC= 83.28%, SENS=78.81, SPEC=87.94%, F1=84.12%

[58], [59], [60]. However, some published methodologies examine Event Related Potentials (ERP) on stimuli-based setups such as the work presented by Khatun et al. [61] that achieved ACC=87.9%. Methodologies on resting state recordings have proposed a variety of classification algorithms such as kNN [60], Random Forests [45], SVM [58] or Neural Networks [59]. The reported ACC of other methodologies ranges from 70% to 85%, however various studies have reported LOSO ACC over 98%.

Regarding FTD, even fewer studies have been published that propose a machine-learning framework for the classification of EEG signals. By performing a search in Scopus (date of search: 23 February 2023) with keywords ''EEG and Frontotemporal AND (detection or classification)'' for the years 2019-2023, 23 studies were found and only one of them [45], which was performed by our research team as this study was about FTD classification relying only on EEG signals (and not a biomarker combination such as EEG+MRI). Thus, no comparison can be made regarding the results of the FTD-CN problem obtained in this study. While there have been many studies on the use of EEG in the diagnosis and classification of other forms of dementia, there is a noticeable lack of research in this area for FTD. This is concerning, as EEG has the potential to provide valuable information about the underlying neural mechanisms of the disease. More research should be done to explore the use of EEG in FTD classification and diagnosis, and to identify potential biomarkers that could aid in early detection and treatment.

Regarding the limitations of this research the following issues should be addressed. First, the size of the dataset, although decent, is not enough to take advantage of the full potential of the transformer encoder's abilities. It is known that using multiple stacked transformer encoder can enable the deep learning model to learn more complex representations of the input signals by building a hierarchy of representations. Each encoder can learn to capture different levels of abstraction, with higher-level encoders processing the output of lower-level encoders to build a more abstract representation of the input signals. However, increasing the parameter number would require having a larger and more diverse training set to improve generalization and robustness of the model and avoid overfitting. Moreover, issues regarding the importance of the SCC feature should be addressed. Although the selection of the feature is supported by the literature that states that in more advanced stages of AD, EEG recordings may also show decreased synchronization among different brain regions [40], the convolution layer kernel weight magnitude evaluation showed that is far less considered than the relative band feature. Although this is not inherently negative, further investigation regarding other types of connectivity measures that better capture AD characteristics should be performed. Moreover, a limitation to be discussed is the fact that all recordings were obtained from a single medical center. While our study focuses on algorithmic development, we acknowledge that for our model to be truly applicable and useful in medical practice, it should undergo broader and more rigorous validation. In accordance with the reviewer's suggestion, we recognize the importance of adhering to the CLAIM criteria (Credibility, Legality, Affordability, Interpretability, Maintainability) outlined in Radiology: Artificial Intelligence [62]. These criteria emphasize the need for external validation using larger datasets from multiple medical centers. Future research endeavors should aim to include a diverse range of patients from various healthcare settings to ensure the generalizability and robustness of our model for real-world clinical applications.

Further elaborating the limitations, it is important to discuss the potential for transfer learning and the limitations of the current methodology (on using transfer learning). Both convolutional neural networks (CNNs) and transformers offer advantages for transfer learning tasks. While our current methodology may not directly support transfer learning, we believe that future propositions could incorporate this

feature effectively. Additionally, a noteworthy limitation of our approach lies in its dependency on a fixed number of electrodes, which restricts its applicability to different EEG setups. However, by developing a more elaborate scheme that is not bound to a specific number of electrodes, we could potentially alleviate this issue and enable the utilization of transfer learning techniques. The combination of transfer learning with a flexible electrode scheme has the potential to enhance the performance and generalization capabilities of our proposed convolution-transformer architecture for Alzheimer's disease (or other dementia) classification.

Regarding the selection of 30 second as the duration of the time-windows and how this differs from the usual time-window size that is most commonly used in such EEG classification methodologies [63], [64] (that being less than 5 seconds), the following should be noted. Usually, small durations of time windows are considered because of the classifiers inability to capture long-range temporal dependencies in the data and due to the limited size of the dataset that is used, which necessitates the generation of a lot of training samples from a small duration of EEG recordings. Nonetheless, the present study effectively addresses these challenges in two distinct ways. Firstly, by harnessing the inherent capability of Transformers to capture long-range dependencies and incorporating them into a convolution scheme that reduces input dimensionality, this methodology enables the exploitation of larger time-windows. Secondly, the dataset employed in this study proves to be substantial, with 485.5 minutes of AD recordings, 276.5 minutes of FTD recordings, and 402 minutes of CN recordings. As a result, there are no limitations arising from a scarcity of training samples, particularly when utilizing a larger window size. Additionally, regarding the 15-second overlap (50%) that is utilized, it serves the purpose of augmenting the training sample count. Although this approach could pose a challenge when employing k-fold validation due to potential overlap between the training and test sets, such concern is effectively mitigated in our case, as we utilize Leave-One-Subject-Out validation.

Regarding the EEG dataset that was utilized in this study, it was structured and made publicly available by our team. As such, this methodology is the first to explore this dataset for AD detection using a convolutional transformer deep neural network. Given the promising results obtained by this methodology, we encourage other researchers to utilize the same dataset and employ this research as a benchmark for further studies in the field. By adopting this approach, future research can directly compare their methodology to ours, and allow for a more objective assessment of their model's performance. Furthermore, this approach can facilitate the development of standardized evaluation metrics for EEG-based AD detection, ultimately leading to the development of more robust and reliable diagnostic tools. Overall, we believe that the publication of this dataset and the development of this methodology have the potential to make a significant contribution to the field of EEG-based dementia detection, and we look forward to future studies building upon this work.

The convolutional transformer deep neural network proposed for EEG-based AD detection has shown great potential for the accurate classification of EEG signals. However, future work could focus on several areas to improve the model's performance and generalizability. Firstly, expanding the dataset used to train and test the model is crucial to enhance its robustness and applicability. Secondly, the methodology should be refined to take advantage of transfer learning to better leverage the power of transformers for EEG signal analysis. Thirdly, the model's generalizability to other EEG electrode setups should be explored, as this could facilitate the adoption of the methodology in clinical settings. Fourthly, graph theory options could be investigated to take advantage of the spatial information of the channels and enhance the model's ability to capture complex inter-channel relationships. Finally, expanding the methodology to other dementia types, such as FTD (that has already been examined in this study) or Lewy body dementia, could help to assess its effectiveness as a diagnostic tool for a broader range of neurodegenerative diseases. Overall, this research presents exciting opportunities for the development of advanced deep-learning techniques for EEG-based dementia detection, with a potential impact on clinical practice and patient outcomes.

V. CONCLUSION

In this study, we investigated the potential of using a novel convolution transformer deep neural network fed with spectral and coherence characteristics extracted from EEG signals for the automatic detection of AD, that being one of the first studies to introduce the transformers' capabilities of capturing relational and semantic information between words (or channels in our instance) for AD EEG detection. We evaluated the performance of the proposed model on a clinical dataset recorded at AHEPA General Hospital of Thessaloniki, Greece. We demonstrated that it achieved state-of-the-art classification accuracy, outperforming several baseline models in the same dataset. We made the dataset publicly available, allowing other researchers to evaluate different models and use this research as a benchmark. The performance results of ACC=83.28% and F1=84.12% suggest that the DICE-net model can effectively capture EEG-derived feature vectors' spectral and spatial patterns and extract meaningful dependencies for classification. Furthermore, our findings contribute to the growing body of literature on using machine learning techniques for EEG-based diagnosis of AD, which has the potential to assist clinicians in the early detection and monitoring of the disease.

DECLARATION OF COMPETING INTEREST

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

INFORMED CONSENT STATEMENT

Informed consent was obtained from all subjects involved in the study.

DATA AVAILABILITY STATEMENT

Datasets related to this article can be found at https://openneuro.org/datasets/ds004504, hosted at OpenNeuro [41].

ACKNOWLEDGMENT

The publication of the article in OA mode was financially supported by HEAL-Link.

REFERENCES

[1] D. V. Puri, S. L. Nalbalwar, A. B. Nandgaonkar, J. P. Gawande, and A. Wagh, "Automatic detection of Alzheimer's disease from EEG signals using low-complexity orthogonal wavelet filter banks," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104439, doi: 10.1016/j.bspc.2022.104439.

[2] M. Fernández, A. L. Gobartt, and M. Balañá, "Behavioural symptoms in patients with Alzheimer's disease and their association with cognitive impairment," *BMC Neurol.*, vol. 10, no. 1, p. 87, Dec. 2010, doi: 10.1186/1471-2377-10-87.

[3] A. Atri, "The Alzheimer's disease clinical spectrum," *Med. Clinics North Amer.*, vol. 103, no. 2, pp. 263–293, Mar. 2019, doi: 10.1016/j.mcna.2018.10.009.

[4] Y. Ding, Y. Chu, M. Liu, Z. Ling, S. Wang, X. Li, and Y. Li, "Fully automated discrimination of Alzheimer's disease using resting-state electroencephalography signals," *Quant. Imag. Med. Surgery*, vol. 12, no. 2, pp. 1063–1078, Feb. 2022, doi: 10.21037/qims-21-430.

[5] S. Fouladi, A. A. Safaei, N. Mammone, F. Ghaderi, and M. J. Ebadi, "Efficient deep neural networks for classification of Alzheimer's disease and mild cognitive impairment from scalp EEG recordings," *Cognit. Comput.*, vol. 14, no. 4, pp. 1247–1268, Jul. 2022, doi: 10.1007/s12559-022-10033-3.

[6] I.-S. Shin, M. Carter, D. Masterman, L. Fairbanks, and J. L. Cummings, "Neuropsychiatric symptoms and quality of life in Alzheimer disease," *Amer. J. Geriatric Psychiatry*, vol. 13, no. 6, pp. 469–474, Jun. 2005, doi: 10.1097/00019442-200506000-00005.

[7] B. Dubois, A. Padovani, P. Scheltens, A. Rossi, and G. Dell'Agnello, "Timely diagnosis for Alzheimer's disease: A literature review on benefits and challenges," *J. Alzheimer's Disease*, vol. 49, no. 3, pp. 617–631, Dec. 2015, doi: 10.3233/JAD-150692.

[8] B. Dubois, "Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS-ADRDA criteria," *Lancet. Neurol.*, vol. 6, no. 8, pp. 734–746, 2007, doi: 10.1016/S1474-4422(07)70178-3.

[9] C. Moral-Rubio, P. Balugo, A. Fraile-Pereda, V. Pytel, L. Fernández-Romero, C. Delgado-Alonso, A. Delgado-Álvarez, J. Matias-Guiu, J. A. Matias-Guiu, and J. L. Ayala, "Application of machine learning to electroencephalography for the diagnosis of primary progressive aphasia: A pilot study," *Brain Sci.*, vol. 11, no. 10, p. 1262, Sep. 2021, doi: 10.3390/brainsci11101262.

[10] D. C. Hammond, "Definitions, standard of care and ethical considerations," in *Clinical Neurotherapy*. Amsterdam, The Netherlands: Elsevier, 2014, pp. 1–17, doi: 10.1016/B978-0-12-396988-0.00001-5.

[11] C. J. James and C. W. Hesse, "Independent component analysis for biomedical signals," *Physiolog. Meas.*, vol. 26, no. 1, pp. R15–R39, Feb. 2005, doi: 10.1088/0967-3334/26/1/R02.

[12] P. Anders, H. Müller, N. Skjæret-Maroni, B. Vereijken, and J. Baumeister, "The influence of motor tasks and cut-off parameter selection on artifact subspace reconstruction in EEG recordings," *Med. Biol. Eng. Comput.*, vol. 58, no. 11, pp. 2673–2683, Nov. 2020, doi: 10.1007/s11517-020-02252-3.

[13] M. X. Cohen, "Where does EEG come from and what does it mean?" *Trends Neurosci.*, vol. 40, no. 4, pp. 208–218, Apr. 2017, doi: 10.1016/j.tins.2017.02.004.

[14] A. Miltiadous, K. D. Tzimourta, N. Giannakeas, M. G. Tsipouras, E. Glavas, K. Kalafatakis, and A. T. Tzallas, "Machine learning algorithms for epilepsy detection based on published EEG databases: A systematic review," *IEEE Access*, vol. 11, pp. 564–594, 2023, doi: 10.1109/ACCESS.2022.3232563.

[15] V. J. Geraedts, L. I. Boon, J. Marinus, A. A. Gouw, J. J. van Hilten, C. J. Stam, M. R. Tannemaat, and M. F. Contarino, "Clinical correlates of quantitative EEG in Parkinson disease," *Neurology*, vol. 91, no. 19, pp. 871–883, Nov. 2018, doi: 10.1212/WNL.0000000000006473.

[16] P. Christodoulides, A. Miltiadous, K. D. Tzimourta, D. Peschos, G. Ntritsos, V. Zakopoulou, N. Giannakeas, L. G. Astrakas, M. G. Tsipouras, K. I. Tsamis, E. Glavas, and A. T. Tzallas, "Classification of EEG signals from young adults with dyslexia combining a brain computer interface device and an interactive linguistic software tool," *Biomed. Signal Process. Control*, vol. 76, Jul. 2022, Art. no. 103646, doi: 10.1016/j.bspc.2022.103646.

[17] V. Aspiotis, A. Miltiadous, K. Kalafatakis, K. D. Tzimourta, N. Giannakeas, M. G. Tsipouras, D. Peschos, E. Glavas, and A. T. Tzallas, "Assessing electroencephalography as a stress indicator: A VR high-altitude scenario monitored through EEG and ECG," *Sensors*, vol. 22, no. 15, p. 5792, Aug. 2022, doi: 10.3390/s22155792.

[18] N. N. Kulkarni and V. K. Bairagi, "Extracting salient features for EEG-based diagnosis of Alzheimer's disease using support vector machine classifier," *IETE J. Res.*, vol. 63, no. 1, pp. 11–22, Jan. 2017, doi: 10.1080/03772063.2016.1241164.

[19] N. Sharma, M. H. Kolekar, and K. Jha, "EEG based dementia diagnosis using multi-class support vector machine with motor speed cognitive test," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102102, doi: 10.1016/j.bspc.2020.102102.

[20] M. S. Safi and S. M. M. Safi, "Early detection of Alzheimer's disease from EEG signals using Hjorth parameters," *Biomed. Signal Process. Control*, vol. 65, Mar. 2021, Art. no. 102338, doi: 10.1016/j.bspc.2020.102338.

[21] M. Şeker, Y. Özbek, G. Yener, and M. S. Özerdem, "Complexity of EEG dynamics for early diagnosis of Alzheimer's disease using permutation entropy neuromarker," *Comput. Methods Programs Biomed.*, vol. 206, Jul. 2021, Art. no. 106116, doi: 10.1016/j.cmpb.2021.106116.

[22] K. D. Tzimourta, N. Giannakeas, A. T. Tzallas, L. G. Astrakas, T. Afrantou, P. Ioannidis, N. Grigoriadis, P. Angelidis, D. G. Tsalikakis, and M. G. Tsipouras, "EEG window length evaluation for the detection of Alzheimer's disease over different brain regions," *Brain Sci.*, vol. 9, no. 4, p. 81, Apr. 2019, doi: 10.3390/brainsci9040081.

[23] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *J. Neural Eng.*, vol. 16, no. 5, Oct. 2019, Art. no. 051001, doi: 10.1088/1741-2552/ab260c.

[24] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *IEEE Trans. Big Data*, vol. 6, no. 1, pp. 3–28, Mar. 2020, doi: 10.1109/TBDATA.2018.2850013.

[25] D. M. Praveena, D. A. Sarah, and S. T. George, "Deep learning techniques for EEG signal applications—A review," *IETE J. Res.*, vol. 68, no. 4, pp. 3030–3037, Jul. 2022, doi: 10.1080/03772063.2020.1749143.

[26] M. Lopes, R. Cassani, and T. H. Falk, "Using CNN saliency maps and EEG modulation spectra for improved and more interpretable machine learning-based Alzheimer's disease diagnosis," *Comput. Intell. Neurosci.*, vol. 2023, pp. 1–17, Feb. 2023, doi: 10.1155/2023/3198066.

[27] M. Alessandrini, G. Biagetti, P. Crippa, L. Falaschetti, S. Luzzi, and C. Turchetti, "EEG-based Alzheimer's disease recognition using robust-PCA and LSTM recurrent neural network," *Sensors*, vol. 22, no. 10, p. 3696, May 2022, doi: 10.3390/s22103696.

[28] A. Vaswani, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst., (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[29] 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," *Phys. A, Stat. Mech. Appl.*, vol. 603, Oct. 2022, Art. no. 127700, doi: 10.1016/j.physa.2022.127700.

[30] J. Xie, J. Zhang, J. Sun, Z. Ma, L. Qin, G. Li, H. Zhou, and Y. Zhan, "A transformer-based approach combining deep learning network and spatial–temporal information for raw EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2126–2136, 2022, doi: 10.1109/TNSRE.2022.3194600.

[31] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. Accessed: Apr. 1, 2023. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[32] 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 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[33] Z. Tian, J. Yi, Y. Bai, J. Tao, S. Zhang, and Z. Wen, "Synchronous transformers for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7884–7888, doi: 10.1109/ICASSP40776.2020.9054260.

[34] S. Rahardja, M. Wang, B. P. Nguyen, P. Fränti, and S. Rahardja, "A lightweight classification of adaptor proteins using transformer networks," *BMC Bioinf.*, vol. 23, no. 1, p. 461, Nov. 2022, doi: 10.1186/s12859-022-05000-6.

[35] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers Hum. Neurosci.*, vol. 15, Jun. 2021, Art. no. 653659, doi: 10.3389/fnhum.2021.653659.

[36] D. Klepl, F. He, M. Wu, D. J. Blackburn, and P. Sarrigiannis, "EEG-based graph neural network classification of Alzheimer's disease: An empirical evaluation of functional connectivity methods," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2651–2660, 2022, doi: 10.1109/TNSRE.2022.3204913.

[37] K. AlSharabi, Y. B. Salamah, A. M. Abdurraqeeb, M. Aljalal, and F. A. Alturki, "EEG signal processing for Alzheimer's disorders using discrete wavelet transform and machine learning approaches," *IEEE Access*, vol. 10, pp. 89781–89797, 2022, doi: 10.1109/ACCESS.2022.3198988.

[38] E. Sibilano, "An attention-based deep learning approach for the classification of subjective cognitive decline and mild cognitive impairment using resting-state EEG," *J. Neural Eng.*, vol. 20, no. 1, Feb. 2023, Art. no. 016048, doi: 10.1088/1741-2552/acb96e.

[39] Y. Özbek, E. Fide, and G. G. Yener, "Resting-state EEG alpha/theta power ratio discriminates early-onset Alzheimer's disease from healthy controls," *Clin. Neurophysiol.*, vol. 132, no. 9, pp. 2019–2031, Sep. 2021, doi: 10.1016/j.clinph.2021.05.012.

[40] C. J. Stam, Y. Van Der Made, Y. A. L. Pijnenburg, and P. Scheltens, "EEG synchronization in mild cognitive impairment and Alzheimer's disease," *Acta Neurologica Scandinavica*, vol. 108, no. 2, pp. 90–96, Aug. 2003, doi: 10.1034/j.1600-0404.2003.02067.x.

[41] A. Miltiadous, "A dataset of 88 EEG recordings from: Alzheimer's disease, Frontotemporal dementia and healthy subjects," OpenNeuro, doi: 10.18112/openneuro.ds004504.v1.0.1.

[42] A. Miltiadous, K. D. Tzimourta, T. Afrantou, P. Ioannidis, N. Grigoriadis, D. G. Tsalikakis, P. Angelidis, M. G. Tsipouras, E. Glavas, N. Giannakeas, and A. T. Tzallas, "A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG," *Data*, vol. 8, no. 6, p. 95, May 2023, doi: 10.3390/data8060095.

[43] M. Plechawska-Wójcik, P. Augustynowicz, M. Kaczorowska, E. Zabielska-Mendyk, and D. Zapała, "The influence assessment of artifact subspace reconstruction on the EEG signal characteristics," *Appl. Sci.*, vol. 13, no. 3, p. 1605, Jan. 2023, doi: 10.3390/app13031605.

[44] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004, doi: 10.1016/j.jneumeth.2003.10.009.

[45] A. Miltiadous, K. D. Tzimourta, N. Giannakeas, M. G. Tsipouras, T. Afrantou, P. Ioannidis, and A. T. Tzallas, "Alzheimer's disease and frontotemporal dementia: A robust classification method of EEG signals and a comparison of validation methods," *Diagnostics*, vol. 11, no. 8, p. 1437, Aug. 2021, doi: 10.3390/diagnostics11081437.

[46] B. Oltu, M. F. AkŞahin, and S. Kibaroğlu, "A novel electroencephalography based approach for Alzheimer's disease and mild cognitive impairment detection," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102223, doi: 10.1016/j.bspc.2020.102223.

[47] A. Alkan and M. K. Kiymik, "Comparison of AR and Welch methods in epileptic seizure detection," *J. Med. Syst.*, vol. 30, no. 6, pp. 413–419, Nov. 2006, doi: 10.1007/s10916-005-9001-0.

[48] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, S. Dasgupta and D. McAllester, Eds. Atlanta, GA, USA: PMLR, Feb. 2013, pp. 115–123. [Online]. Available: https://proceedings.mlr.press/v28/bergstra13.html

[49] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013, doi: 10.1088/1741-2552/aace8c.

[50] N. Waytowich, V. J. Lawhern, J. O. Garcia, J. Cummings, J. Faller, P. Sajda, and J. M. Vettel, "Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials," *J. Neural Eng.*, vol. 15, no. 6, Dec. 2018, Art. no. 066031, doi: 10.1088/1741-2552/aae5d8.

[51] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017, doi: 10.1002/hbm.23730.

[52] A. Paszke, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[53] M. Hata, H. Kazui, T. Tanaka, R. Ishii, L. Canuet, R. D. Pascual-Marqui, Y. Aoki, S. Ikeda, H. Kanemoto, K. Yoshiyama, M. Iwase, and M. Takeda, "Functional connectivity assessed by resting state EEG correlates with cognitive decline of Alzheimer's disease—An eLORETA study," *Clin. Neurophysiol.*, vol. 127, no. 2, pp. 1269–1278, Feb. 2016, doi: 10.1016/j.clinph.2015.10.030.

[54] K. D. Tzimourta, V. Christou, A. T. Tzallas, N. Giannakeas, L. G. Astrakas, P. Angelidis, D. Tsalikakis, and M. G. Tsipouras, "Machine learning algorithms and statistical approaches for Alzheimer's disease analysis based on resting-state EEG recordings: A systematic review," *Int. J. Neural Syst.*, vol. 31, no. 5, May 2021, Art. no. 2130002, doi: 10.1142/S0129065721300023.

[55] K. D. Tzimourta, T. Afrantou, P. Ioannidis, M. Karatzikou, A. T. Tzallas, N. Giannakeas, L. G. Astrakas, P. Angelidis, E. Glavas, N. Grigoriadis, D. G. Tsalikakis, and M. G. Tsipouras, "Analysis of electroencephalographic signals complexity regarding Alzheimer's disease," *Comput. Electr. Eng.*, vol. 76, pp. 198–212, Jun. 2019, doi: 10.1016/j.compeleceng.2019.03.018.

[56] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk, "Brain imaging in Alzheimer disease," *Cold Spring Harbor Perspect. Med.*, vol. 2, no. 4, Apr. 2012, Art. no. a006213, doi: 10.1101/cshperspect.a006213.

[57] E. Gordon, J. D. Rohrer, L. G. Kim, R. Omar, M. N. Rossor, N. C. Fox, and J. D. Warren, "Measuring disease progression in frontotemporal lobar degeneration: A clinical and MRI study," *Neurology*, vol. 74, no. 8, pp. 666–673, Feb. 2010.

[58] T. Araújo, J. P. Teixeira, and P. M. Rodrigues, "Smart-data-driven system for Alzheimer disease detection through electroencephalographic signals," *Bioengineering*, vol. 9, no. 4, p. 141, Mar. 2022, doi: 10.3390/bioengineering9040141.

[59] S. Ruiz-Gómez, C. Gómez, J. Poza, G. Gutiérrez-Tobal, M. Tola-Arribas, M. Cano, and R. Hornero, "Automated multiclass classification of spontaneous EEG activity in Alzheimer's disease and mild cognitive impairment," *Entropy*, vol. 20, no. 1, p. 35, Jan. 2018, doi: 10.3390/e20010035.

[60] S. Dogan, M. Baygin, B. Tasci, H. W. Loh, P. D. Barua, T. Tuncer, R.-S. Tan, and U. R. Acharya, "Primate brain pattern-based automated Alzheimer's disease detection model using EEG signals," *Cognit. Neurodyn.*, vol. 17, no. 3, pp. 647–659, Aug. 2022, doi: 10.1007/s11571-022-09859-2.

[61] S. Khatun, B. I. Morshed, and G. M. Bidelman, "A single-channel EEG-based approach to detect mild cognitive impairment via speech-evoked brain responses," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 1063–1070, May 2019, doi: 10.1109/TNSRE.2019.2911970.

[62] J. Mongan, L. Moy, and C. E. Kahn, "Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers," *Radiology, Artif. Intell.*, vol. 2, no. 2, Mar. 2020, Art. no. e200029, doi: 10.1148/ryai.2020200029.

[63] O. Özdenizci, S. Eldeeb, A. Demir, D. Erdoğmuş, and M. Akçakaya, "EEG-based texture roughness classification in active tactile exploration with invariant representation learning networks," *Biomed Signal Process Control*, vol. 67, no. March, pp. 1–7, 2021, doi: 10.1016/j.bspc.2021.102507.

[64] J. Seo, T. H. Laine, G. Oh, and K. A. Sohn, "EEG-based emotion classification for Alzheimer's disease patients using conventional machine learning and recurrent neural network models," *Sensors*, vol. 20, no. 24, pp. 1–27, Dec. 2020, doi: 10.3390/s20247212.

ANDREAS MILTIADOUS was born in Ioannina, Greece, in 1996. He received the B.S. and M.S. degrees in computer science from the Aristotle University of Thessaloniki, Greece, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree in bioinformatics with the University of Ioannina, Greece.

He was with the Greek Military Force, from 2018 to 2019. He was a computer science teacher for high school students, from 2019 to 2021. His research interests include the analysis and interpretation of EEG signal and other biosignals, the study of machine learning and deep learning, the development of automatic systems of recognition, and the detection of neurophysiological conditions or states.

EMMANOUIL GIONANIDIS was born in Volos, Greece, in 1996. He received the B.Sc. and M.Sc. degrees in informatics from the Aristotle University of Thessaloniki, in 2018 and 2020, respectively. Currently, he is a Machine Learning Researcher and an Engineer with DataWise Data Engineering LLC. From 2020 to 2021, he was with the Greek Military Force, Research and Informatics Corps. Prior to that, he was a Research Assistant with the Artificial Intelligence and Information Analysis (AIIA) Laboratory, School of Informatics, Aristotle University of Thessaloniki, from 2019 to 2020. His research interests include natural language processing, particularly in the areas of language models and neural networks, and computer vision. He focuses on the research and development of machine learning and deep learning systems.

KATERINA D. TZIMOURTA was born in Thessaloniki, Greece, in 1991. She received the B.S. and M.S. degrees in informatics and telecommunications engineering from the University of Western Macedonia, Kozani, Greece, in 2015, and the Ph.D. degree in bioinformatics from the University of Ioannina, Greece, in 2020.

She was an IT Engineer with the Technological Educational Institute of Epirus, from 2018 to 2019. She has been an Adjunct Lecturer with the Department of Informatics and Telecommunications, University of Ioannina, since 2020. She has also been a Postdoctoral Researcher with the University of Western Macedonia, since 2020. Her research interests include biosignal processing methods, machine learning algorithms, brain–computer interfaces, and wearable devices for movement and brain disorder's analysis. She is involved with rare genetic disorders and particularly the extreme rare Kleefstra syndrome awareness.

NIKOLAOS GIANNAKEAS received the degree in physics from the University of Ioannina, Greece, in 2003, the degree in computer science from Hellenic Open University, Greece, in 2020, and the Ph.D. (Diploma) degree in bioinformatics from the Medical School, University of Ioannina, in 2011. He has worked for more than 15 years in research projects funded by National and European Programs (third CSF, NSRF (2007–2013), NSRF (2014–2020), FP6, FP7, and Horizon 2020). Since 2019, he has been an Assistant Professor with the Informatics and Telecommunications Department, University of Ioannina. His research interests include signal processing and image analysis, artificial intelligence and machine learning, bioinformatics, and biomedical engineering.

ALEXANDROS T. TZALLAS received the B.Sc. degree in physics and the Ph.D. degree in medical physics from the University of Ioannina, Ioannina, Greece, in 2001 and 2009, respectively. He is currently an Associate Professor in biomedical engineering and specifically in the analysis and processing of biomedical data with the Department of Informatics and Telecommunications, University of Ioannina. He is also affiliated as an academic research fellow with a number of research and technology Institutes in Greece and U.K. His research interests include neuroscience, electroencephalography, wearable devices, biomedical signal and image processing, biomedical engineering, decision support medical expert systems, and biomedical applications.