12

Explainable deep learning to information extraction in diagnostics and electrophysiological multivariate time series☆

Francesco Carlo Morabito, Maurizio Campolo, Cosimo Ieracitano, and Nadia Mammone

AI_LAB-NEUROLAB, DICEAM, UNIVERSITY MEDITERRANEA OF REGGIO CALABRIA, REGGIO CALABRIA, ITALY

Chapter outlines

☆To my loved daughter, Valeria.

Artificial Intelligence in the Age of Neural Networks and Brain Computing. https://doi.org/10.1016/B978-0-323-96104-2.00011-7

1 Introduction

The processing of multivariate time series for identification, classification, and prediction problems or information extraction is a research and application topic spanning many different science domains. The strong impact of digitalization, information, and communication technology as well as the prevalence of sensor networks for the emergence of the Internet of Things (IoT) have strongly motivated a resurgence of interest in machine learning (ML) for multivariate time-series analysis. deep learning (DL) techniques are a variant of ML well founded in classical neural network theory. They use deep architectures where many hidden layers or maps of neurons generate a lower-dimensional projection of the input space that corresponds, for example, to the signals generated by the network of sensors in monitoring applications. The successive hidden layers are able to build an effective high-level abstraction of the raw data. State-of-the-art DL processors present architectural advantages and benefits of novel training paradigms synergistic with other approaches, like compressive sensing and sparsity methods. The high number of neurons and links is reminiscent of brain networks and allows the storage of the essential features of the underlying input-output mapping. In biomedical signal processing, many diagnostic systems produce multivariate time series, and the automatic extraction of features without human intervention is of high interest for supporting clinical diagnoses and for highlighting latent aspects hidden in standard visual interpretation. For example, in medical imaging, small irregularities in tissues from prodromal to tumors can be detected in the successive levels of abstractions of the DL network. The development of efficient DL systems can have a significant impact on public health also because of the possibility of incorporating real-time information in the existing computational models. DL is indeed data-driven model endowed with the ability of generalization. In this chapter, DL methods are briefly presented from the historical perspective of neural network studies. Electroencephalographic (EEG) multivariate data are considered as many application domains spanning from brain-computer interface (BCI) to neuroscience take advantage of this noninvasive and cheap technique as the basis of brain studies. Two different DL architectures will be proposed that successfully solve difficult neurology problems.

2 The neural network approach

ML has the objective to yield computers the ability to autonomously learn and interact with their environment by exploiting data to learn optimal behaviors without the need for a specific programming step. Neural networks (NN) are machines explicitly designed

to possess this ability. NNs are a collection of elementary processing nodes suitably arranged in various topological architectures. The elementary node of the network is referred to as a neuron and includes a linear part taking a weighted linear combination of its inputs and a nonlinear part where a selected nonlinear function transforms the net input in the final output of the node. The inputs of the neuron come from other neurons and its output is distributed to other nodes. The organization of the neurons is hierarchical if the nodes are structured in layers. This kind of topology is somehow reminiscent of the organization of pyramidal neurons in the mammalian brain. Other kinds of topologies have been considered, in particular the maps and the grids, where the neurons are organized in 2D or 3D distributions. The most typical neural network architecture is referred to as multilayer perceptron (MLP): here the neurons are organized in successive layers. The input layer is linear and just distributes the inputs (possibly deriving from a network of sensors) to the successive layer; the output layer gives the final output of the processing chain; in between there is a layer of nodes that are not linked to inputs and outputs and for this reason is called "hidden." A pictorial illustration of the MLP NN is reported in Fig. 1A. The MLP is quite similar to the ADALINE previously proposed in signal processing literature apart from the presence of the nonlinearities at least in the hidden layer's nodes.

Various nonlinear functions have been proposed for approximation, pattern recognition, and classification problems. In MLP, the nodes in successive layers are connected and the connections are weighted. Among them, monotonic saturating functions (i.e., sigmoids) are a favorite choice. In dynamical problems, nonmonotonic functions have been proposed. The weights are typically randomly initialized.

Thus, the MLP is of a feedforward type. However, in recursive networks, there are also feedback links and this enriches the dynamical behavior of NNs.

The relevance of MLP in approximating input-output mappings has been increased by the proofs of some theorems, demonstrating that any continuous mapping can be approximately by MLP with at least one hidden layer whose output functions are sigmoid (or

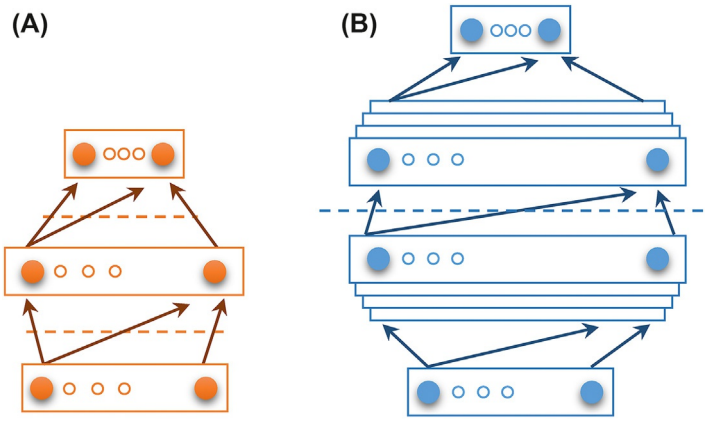

FIG. 1 (A) A shallow (one hidden layer) and (B) a deep (multiple hidden layers) neural network.

monotonic) functions. It is worth mentioning that this notion of universal approximation just states that the NN can learn, not what in practice it really does learn.

NNs are adaptive systems that are trained aiming to derive an optimal representation of the weights' matrices. The training is carried out through a specific "learning" procedure. The learning can be supervised (SL), unsupervised (UL), semisupervised (SSL), or reinforced (RL). In SL, NNs are forced to associate their outputs to real (or complex) valued targets fixed by a "teacher," through a procedure (typically gradient-based) that optimizes approximation errors (a "cost" function). The goal of SL is to derive optimal weights' matrices that minimize or achieve a small error. However, NNs are asked to generalize well in unseen input vectors, i.e., generating small errors also testing cases. In UL, there is no teacher, and the NN is asked to autonomously extract some underlying statistical regularities from the available data. In SSL, a prestage of UL is used to facilitate the following SL procedure. In the case of the availability of both labeled and unlabeled data, these procedures can help to extract additional information on the problem under analysis. In some applications, i.e., image "semantic" segmentation, an additional form of learning is considered, referred to as weakly supervised learning, where the relevant labels are weakly annotated in the training data because of the heavy annotation cost. In RL, the machine interacts with the surrounding environment and, following the observation of the consequences of its actions, it learns how to modify its own behavior in response to some form of "reward" received.

The different approximation theorems imply a "sufficient" number of hidden nodes to satisfy the universal approximation properties. However, this in turn implies a high number of degrees of freedom in the optimization procedure. The capacity of the NN is statistically bounded and underlies a relation between the number of weights to be determined and the number of "examples" available to train them. A high number of free parameters increases the descriptive complexity of NN, approximately related to the number of bits of information required to describe a NN. The complexity of NN limits generalization ability. Unsuccessful generalization performance reduced the impact of the NN approach, after an initial enthusiastic acceptance. Some techniques have been proposed to reduce the complexity of NNs, among which some concepts relevant to deep learning (DL) are the weight sharing, the regularization, and the forced introduction of data invariances.

3 Deep architectures and learning

DL methods iteratively modify more sets of parameters (the weights and the biases of the layer-to-layer matrices) by minimizing a loss/cost function aiming to define an optimal set. However, the performance of DL, and more generally, ML and NN approaches strongly depend on the quality of available data or the careful selection of a representation suitable for the task at hand. Most of the efforts in designing the processing chain are thus devoted to data preprocessing or domain transformation. Time series are commonly analyzed in time, frequency, or time-frequency domain. The related transformation constitutes an

engineering way to extract features from data that are apparent in a specific domain. DL represents the most significant and successful advance in ML over the last decade. A large part of the current appeal of DL techniques derives from the possibility of acquiring data representations that are not model-based but totally data-driven. This circumvents the need to hand-designed features. The hierarchically organized learned features are often richer and more powerful than the ones suitably engineered. DL is indeed an emerging methodology firmly rooted within the traditional ML community whose main objective is to design learning algorithms and architectures for extracting multiple-level representations from data. The representation is both hierarchical and distributed, as the relevant characteristics of a problem emerge gradually in successive levels (or layers) and are the collective result of weighting multiple nodes similarly to shallow NNs. These representations facilitate the pattern recognition tasks sometimes without the need for any feature engineering but just autonomously extracting them from the available data. This is because the multiple layers of successive latent representations are able to disentangle potential confounding factors in the input data, also reducing their complexity. Fig. 1B shows a deep architecture. The huge amount of researches recently carried out in the field by a large number of academic and industrial groups are motivated by the surprising successes achieved also in precommercial competitions and applications. AlphaGo and Watson are some relevant examples. Major IT companies (e.g., Google, IBM, Intel, Facebook, Baidu, and Microsoft) hold a large extent of patents in the field; they also made DL their core business. This resurgence of interest in the NN approach is related to the following evidences:

(1) General availability of large database (big data) coming from international initiatives and worldwide collaboration on projects;
(2) Availability of big computing power mainly associated with cloud computing and novel GPU extensions;
(3) Availability of novel algorithms and processing architectures, or advanced paradigms of computation, like quantum computing and memristor-based network implementations.

Indeed, as previously noted, the capacity of a NN chain is related to the number of free parameters whose estimation calls for large datasets. In turn, to process big data, powerful computing is needed.

Some of the DL schemes are biologically motivated. In essence, the brain visual cortex inspired hierarchical DL architectures. In particular, neurons found in the visual cortex of cats respond to specific properties of visual sensory inputs, like lines, edges, colors, and the successive layers extract combinations of such low-level features to derive higher-level features resulting in objects' recognition [1].

Several DL models have been proposed in the literature. In what follows, the most known are presented: deep belief networks (DBNs), stacked autoencoders (SAEs), and deep convolution neural networks (CNN).

3.1 Deep belief networks

DBN is a probabilistic generative model, composed of stacked modules of restricted Boltzmann machines (RBMs) (Fig. 2) [2]. An RBM is an undirected energy-based model with two layers of visible (v) and hidden (h) units, respectively, with connections only between layers. Each RBM module is trained one at a time in an unsupervised manner and using a contrastive divergence procedure [3]. The output (learned features) of each stage is used as input for the subsequent RBM stage. After, the whole network is commonly trained with supervised learning to improve classification performance (fine-tuning method).

3.2 Stacked autoencoders

The *Stacked Autoencoders* architecture is similar to DBNs, where the main component is the autoencoder (Fig. 3) [4]. An autoencoder (AE) is a NN trained with unsupervised learning whose attempt is to reproduce at its output the same configuration of input. A single hidden layer with the same number of inputs and outputs implements it. AE consists of two main stages: compression of the input space into a lower dimension space (encoding) and reconstruction of the input data from the compressed representation (decoding). In a stacked architecture, the encoded pattern is used as input for training the successive AEs. The SAE ends with an output layer trained with a supervised criterion. As DBN, the whole network can be fine-tuned to improve classification performance.

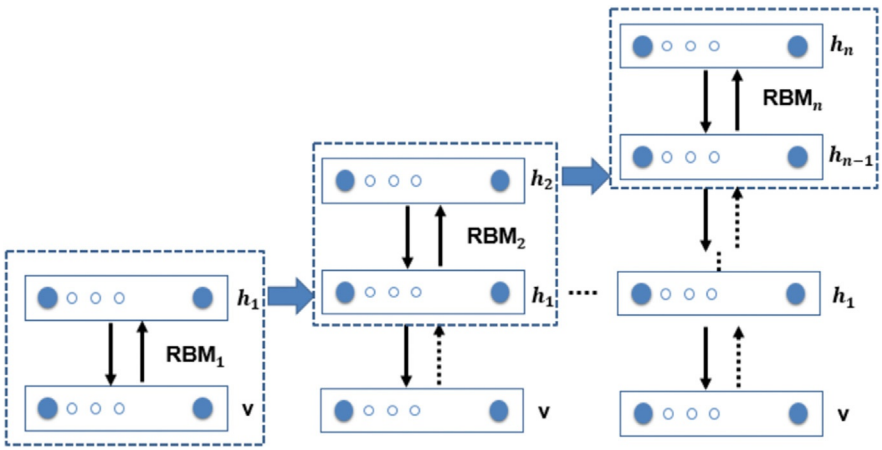

FIG. 2 Deep belief network (DBN) architecture composed of stacked restricted Boltzmann machines (RBMs). Each RBM consists of a visible layer **v** and a single hidden layer $\mathbf{h_n}$. RBM$_1$ is trained using the input data as visible units. The hidden layer $\mathbf{h_2}$ of RBM$_2$ is trained using the output of the previous trained layer $\mathbf{h_1}$ of the RBM$_1$. The output of $\mathbf{h_2}$ is the input of the next RBM$_3$ and so on. The trained layers $\mathbf{h_1}$, $\mathbf{h_2}$, …, $\mathbf{h_n}$ form the stacked architecture. Finally, the whole DBN is fine-tuned with a standard backpropagation algorithm.

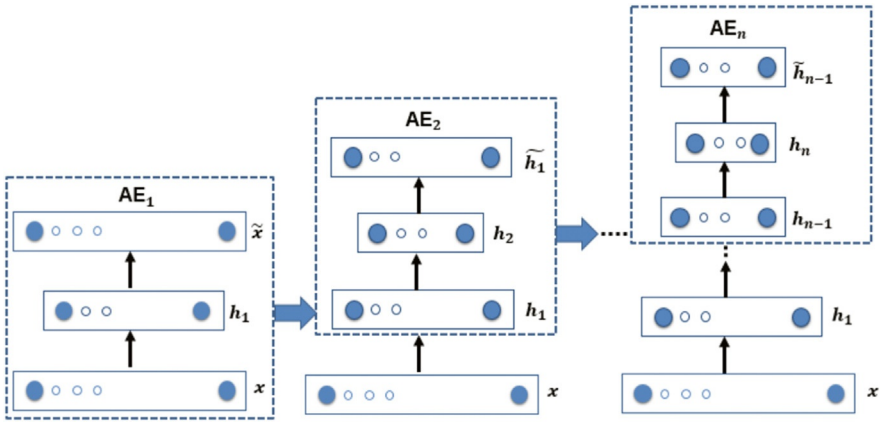

FIG. 3 Stacked autoencoders architecture. The first autoencoder (AE_1) maps the input instance **x** into a compressed representation $\mathbf{h_1}$ (coding operation) which is used to reconstruct the input data (decoding operation). After training AE_1, the *code* $\mathbf{h_1}$ is used as input to train AE_2, providing the *code* vector $\mathbf{h_2}$ and so on. The procedure is repeated for all AE_n autoencoders. The compressed representations $\mathbf{h_1}$, $\mathbf{h_2}$, …, $\mathbf{h_n}$ form the stacked architecture (SAE) which is typically fine-tuned using a conventional backpropagation algorithm.

3.3 Convolutional neural networks

Convolutional neural networks (CNN) are an alternative type of DNN that allow to model both time and space correlations in multivariate signals. They are attractive as they explicitly consider and take advantage of input topology. In SAE, for example, the inputs can be organized in any order without affecting the performance of the model. In biomedical signal processing, however, spectral and time-frequency representations of the original signals show strong correlations: modeling local correlations is easy with CNNs through weight sharing. CNNs are inspired by the visual cortex of the brain and have been widely applied in image and speech recognition. A CNN includes an automatic feature extractor (composed of multiple stages of convolution and pooling layers) and a standard MLP-NN which processes the features learned before for classification tasks (Fig. 4) [5]. The convolutional layer computes the dot product between the input image X and a set of K_j learnable filters. Each filter K_j sized $k_1 \times k_2$ moves across the input space performing the convolution with local subblocks of inputs, providing Y_j *feature maps* ($Y_j = \sum X^* K_j + B_j$, where B is the bias term). A rectified linear unit (ReLU) activation function is commonly applied to each feature map, to improve computational efficiency by inducing sparsity and to reduce the effect of the vanishing gradient problem [6,7]. The nonlinear convolutional layer is followed by the pooling layer which performs a maximum or average subsampling of the feature maps previously obtained in the previous step: a filter sized $\overline{k}_1 + \overline{k}_2$ moves across the input feature map taking the maximum (max pooling) or the average (average pooling) of the neighbor values selected by the filter. Finally, the learned feature maps are the input of a standard NN that performs classification tasks.

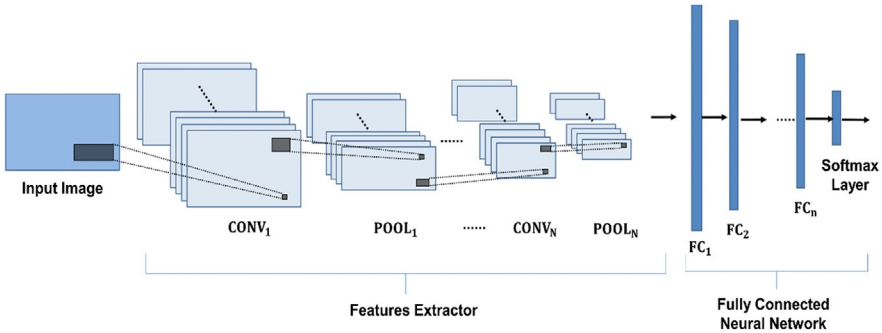

FIG. 4 CNN architecture. It includes a sequence of convolution (CONV) and pooling (POOL) layers followed by a standard fully connected neural network. In the convolutional layer, the input map convolves with *K* filters (or kernels), providing *K* feature maps. After applying a nonlinear activation function (sigmoidal or ReLU) to each feature map, the pooling layer is performed. The features learned are the input of a fully connected neural network followed by a softmax layer, which performs the classification tasks.

4 Electrophysiological time series

4.1 Multichannel neurophysiological measurements of the activity of the brain

The brain generates bio-electromagnetic fields related to the activity and the synaptic action of interacting neurons. This activity can be detected through sensors yielding neurophysiological measurements like electroencephalography (EEG) and magnetoencephalography (MEG).

4.2 Electroencephalography (EEG)

The EEG collects the measurements of the brain's electrical activity. It consists of recording and displaying, over the time, the voltage difference between two scalp sites: the location of interest and the "reference" location. Since its discovery in 1924 by Berger, EEG has become a routine examination in neurology [8] and the basic neurophysiological measurement of many brain-computer interface (BCI) applications [9].

EEG electrodes are located at the surface of the scalp (Fig. 5) and collect the electrical activity generated by networks of neurons. Extracellular current flow is generated because of the excitatory and inhibitory postsynaptic potentials produced by cell bodies and dendrites of pyramidal neurons. The *EEG waveforms* are mainly produced by layers of pyramidal neurons whose synchronous activity produces a bioelectromagnetic field, that propagates from the sources to the recording scalp electrodes. Fields propagate through tissues that have different conduction properties and overlap with the fields generated by other neuronal populations. As a result, the potentials recorded at a specific electrode site will reflect the combination of the contributions of different cortical sources (volume

FIG. 5 Standard EEG recording cap.

conduction effect). The issue of volume conduction is not trivial in EEG analysis as it can result in apparently high functional connectivity between channels thus leading to a wrong neurophysiological interpretation of the results. It has long been discussed in the literature and all of the proposed approaches (Laplacian filtering, source estimation, and the use of connectivity measures not sensitive to phase interactions, among others) introduce various limitations and may cancel relevant source activity at low spatial frequencies [10].

EEG waveforms are characterized by amplitude, shape, morphology, and frequency. Four major rhythms are commonly investigated when analyzing EEG signals: delta (0–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz) bands, which are associated with specific physiological and mental processes [11]. Alpha is the main resting rhythm of the brain, it is commonly observed in awake adults, especially in the occipital electrodes. In healthy subjects, theta rhythm appears at the early stages of sleep and the delta appears at deep-sleep stages. Beta waves appear because of anxiety or intense mental activity. The brain wave components of EEG signals can be investigated by frequency analysis or, when keeping track of the temporal evolution of EEG frequencies necessary, by time-frequency analysis [12].

The EEG electrode placement was standardized in a 10–20 system, which is a method to describe the location of electrodes over the scalp. This system is based on the relationship between the location of the electrode and underlying brain lobes. The distances between adjacent electrodes are either 10% or 20% of the total front-back (nasion-inion) or right-left distance of the skull.

EEG signals are important in the study of many neurological diseases. For example, Alzheimer's disease (AD) is a neurodegenerative disorder with a subtle, asymptomatic onset and a gradual progression toward the full-blown stage of the disease, when the clinical symptoms become noticeable [13]. AD affects the neuronal metabolic processes and leads to a loss of connections between nerve cells. Three main characteristics can be commonly observed in AD patients' EEG signals, compared to healthy controls: the slowing effect (the power at low frequencies increases, whereas the power at high frequencies decreases), the reduction of complexity and of synchrony between pairs of EEG signals [14,15]. Such effects come with the functional disconnection caused by the death of neurons [16,17].

In Creutzfeldt-Jakob disease (CJD), a progressive transmissible form of encephalopathy, patients develop dementia and a peculiar spongiform degeneration of cortical and subcortical gray matter [18]. EEG helps in diagnosing CJD, particularly in the middle/late stages of the disease when periodic sharp wave complexes (PSWC) become clearly visible [19]. PSWC may disappear at the later stages of the disease when spongiform changes involve the whole cerebral cortex [20,21].

Fig. 6 (top) shows the EEG traces of a patient affected by CJD, a patient affected by AD, and a Healthy Control (HC), recorded by the occipital channel O_1. The EEG was recorded in a comfortable eye-closed resting state. The signals were band-pass filtered at 0.5–32 Hz and sampled at 256 Hz. The CJD EEG (Fig. 6, top) exhibits the typical aforementioned PSWC sharp and wave complexes. The AD EEG (Fig. 6, middle) shows the "slowing effect," peculiar to cerebral degeneration due to AD.

The AD slowing effect is evident in Fig. 6 (middle) and Fig. 7A (middle). The dominant peak in the alpha band (8–13 Hz), peculiar of healthy subjects (HC) in an eye-closed resting state, is indeed present in the power spectral densities (PSD) of the HC, whereas it slowed and reduced in the AD patient as well as in the CJD patient. Delta band (0–4 Hz) looks prominent in both AD and CJD patients, as rather expected from clinical considerations. In Dl approaches, this clinical behavior can be automatically extracted and represented in suitable features.

4.3 High-density electroencephalography

EEG signals have very good temporal resolution but poor spatial resolution, because of both volume conduction effects and large interelectrode distance. Biophysical analyses reported that some information on the scalp electrical potentials is lost unless an intersensor distance of 1 to 2 cm is used [22–25]. With a standard 10–20 system, the average intersensor distance is 7 cm. Achieving a 1 to 2 cm sampling density would require 500 EEG channels distributed uniformly over the scalp. High-density-EEG (HD-EEG) offers a high temporal resolution, typical of EEG, in conjunction with a high spatial resolution (Fig. 8). HD-EEG with 256 channels provides adequate approximate spatial sampling [23,24]. An example of high-density EEG recording is shown in Fig. 9.

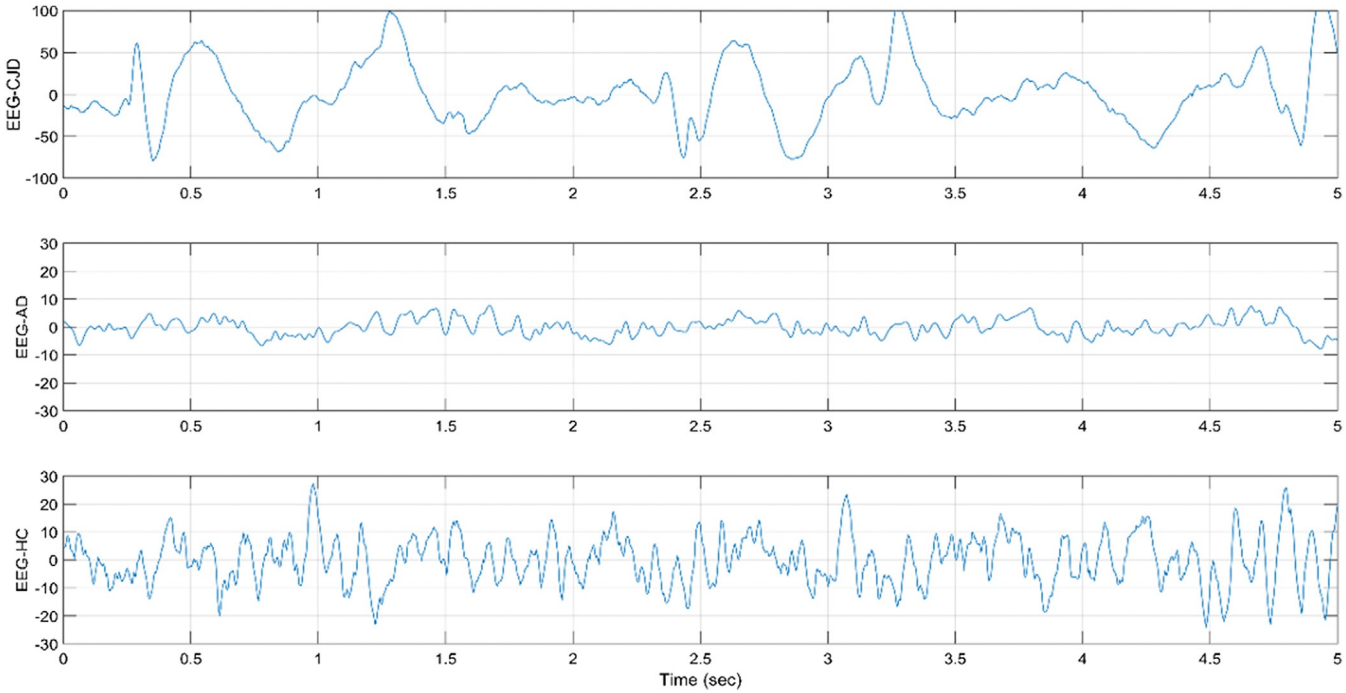

FIG. 6 EEG was recorded at location O1 (occipital) from a CJD patient *(top)*, an AD patient *(middle)*, and a healthy subject *(bottom)*.

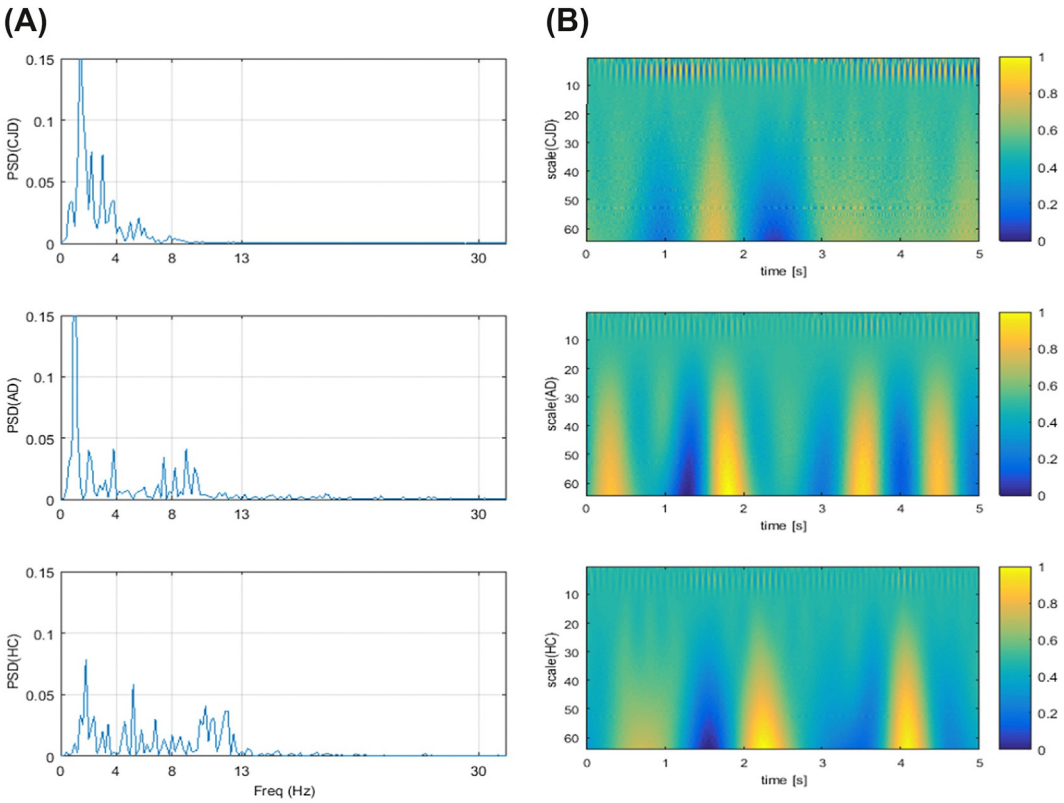

FIG. 7 (A) Power spectral densities (PSD) of the EEG signals (shown in Fig. 6) recorded from a CJD patient *(top)*, an AD patient *(middle)*, and a healthy subject, HC *(bottom)*. The abscissa represents the frequency and the ranges of the brain rhythms (delta, theta, alpha, and beta EEG rhythms are emphasized). (B) Time-frequency maps (TFM) for the same signals.

FIG. 8 High-density 256 channels EEG recording system.

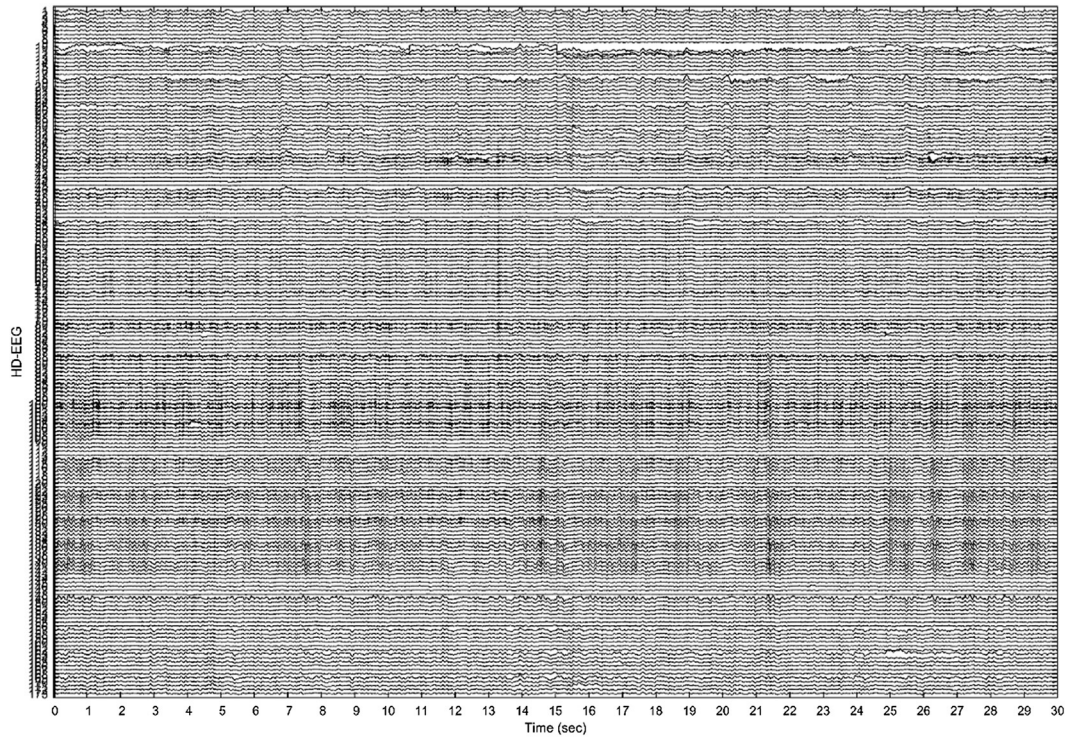

FIG. 9 A 30 s sample of high-density 256 channels EEG recording.

Fig. 10 depicts the 3D scalp topography of the power of the EEG recording shown in Fig. 9. The signal power was estimated for every EEG epoch, and then averaged over the time giving the average power value per channel. The obtained values were then mapped over the scalp, where the coloration of intersensor areas was estimated by interpolation [26].

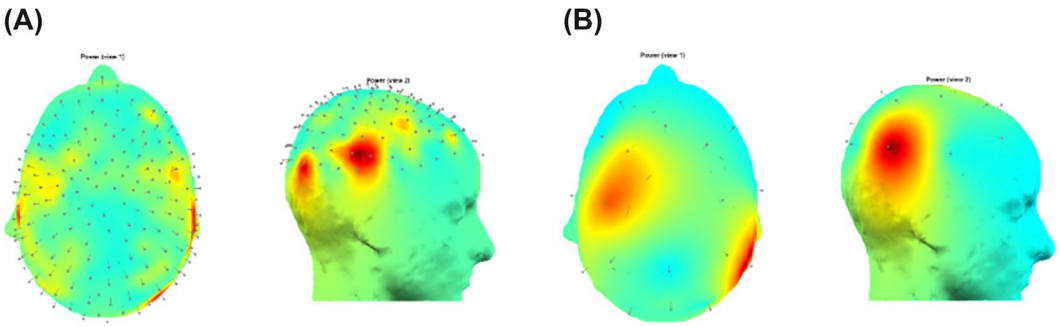

FIG. 10 3D scalp topography of the power of the EEG recording is shown in Fig. 9. (A) Topographic reconstruction was achieved by taking into account all the available scalp channels. (B) Topographic reconstruction was achieved by taking into account only the standard 19 channels.

Fig. 10 shows the interpolated power distribution topography for both HD-EEG and the standard 19-channel EEG. The significant interelectrodes distance in the standard 10–20 configuration implies the need for interpolation with unavoidable loss of precision. As an example, because of the interpolation effect, the two high-power areas (red areas) in the right parietal-occipital zone in Fig. 10A are clustered into one in Fig. 10B.

The potential of HD-EEG has been widely proven in the identification of the epileptogenic onset zone through electrical source imaging [27–29], but it is mostly unexplored in many other fields of application, like dementia. However, it requires a high computational effort and generates huge amounts of data, particularly in long-time monitoring. It is clear that DL approaches can be of great help to manage this kind of data.

4.4 Magnetoencephalography

Magnetoencephalography (MEG) is a functional neuroimaging technique for mapping brain activity by recording magnetic fields produced by electrical currents occurring naturally in the brain, using very sensitive magnetometers. Arrays of superconducting quantum interference devices (SQUIDs) are currently the most common magnetometer.

Although EEG and MEG signals originate from the same neurophysiological processes, many important differences can be highlighted. Magnetic fields are less distorted than electric fields by the different conductivity properties of the head tissues (the skull is insulating, whereas the scalp is conducting), resulting in a mild sensitivity to volume conduction effects and in a greater spatial resolution. This has relevant implications for connectivity analyses and source modeling. Furthermore, MEG measurements are absolute as they are independent of the reference choice.

However, EEG is far more affordable, manageable, and cheap than MEG, which caused its widespread availability, as compared to MEG technology, both in research and clinical practice.

5 Deep learning models for EEG signal processing

In recent years, DL architectures have been applied for the analysis of EEG recordings in cognitive neuroscience and neurology. In this research area, DL models have been developed to learn discriminating features from EEG signals recorded from patients with neurological disorders.

5.1 Stacked auto-encoders

DL methodologies are of growing interest to process complex signals like EEG or MEG, in both disease diagnosis and brain-computer interface (BCI) systems. These kinds of signals represent practical examples of noisy and nonstationary multivariate time series, being acquired simultaneously on multiple channels. Typically, EEG is acquired during a long time for diagnosis purposes and it presents some artefactual and noise activity that may reduce its reliability and visual interpretability. The DL approach to EEG/MEG signal

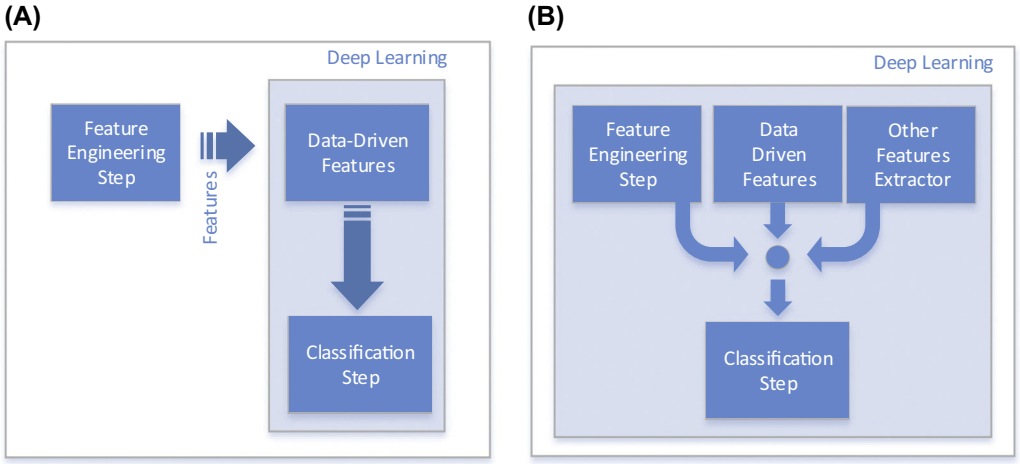

FIG. 11 (A) Serial and (B) parallel DL schemes.

processing can be proposed in two basic ways (see Fig. 11): (1) *in series*, in which a feature engineering step yields a high number of features that are then combined and reduced by a DL network before using them for classification; (2) *in parallel*, where the data-driven features and the engineered features form a unique input vector for the classification step. The serial step is preferred for reducing the information redundancy of the features. In a recent paper [30], a serial scheme based on SAEs has been proposed that includes a time-frequency transform of the input recordings, an intermediate step of data-driven feature combination, and a final classification stage. The SAE model has been proposed for discriminating EEGs of subjects affected by early stage CJD from other forms of rapidly progressive dementia (RPD).

Each AE is implemented by an MLP-NN that includes an encoding stage followed by a decoder. Higher-order features have been obtained by stacking two levels of nonlinear (sigmoidal) nodes. The output of the deepest hidden layer is the final feature vector used as input for the classification step. The processing chain is depicted in Fig. 12. It is worth noting that while the "engineered" features are extracted channel by channel, the higher-order features generated by the SAE are mixing information pertaining to all of the channels. This procedure can be limited to selected areas, i.e., frontal or parietal channels, to gain information on brain areas mostly relevant for the classification.

The output of each hidden layer of the SAE is given by:

$$\underline{h} = \underline{\sigma}\left(\underline{\underline{\Phi}}\,\underline{x}\right),$$

where $\underline{s}$ is the node nonlinearity, for example, the standard sigmoidal function:

$$\sigma = (1 + e^{-z})^{-1}.$$

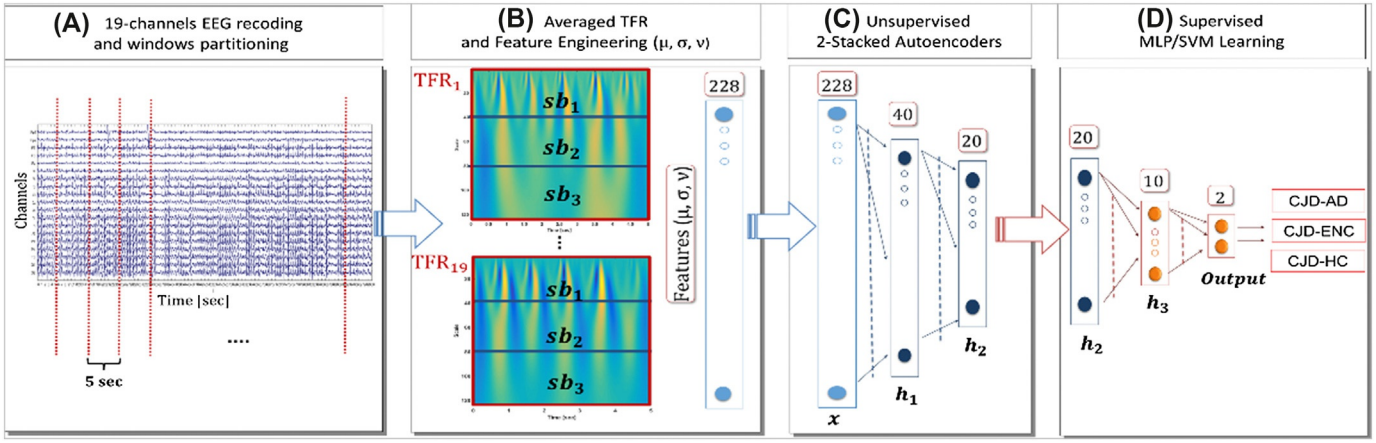

FIG. 12 Flowchart of the method proposed in Ref. [30]. (A) The 19-channel EEG recording is partitioned into *N* nonoverlapping 5 s windows. (B) For each EEG epoch, and for every EEG channel a time-frequency representation (TFR) is computed. The TFRs are averaged over epochs resulting in 19 averaged TFRs (one per channel). Each averaged TFR is subdivided into three subbands, and then, the mean (μ), the standard deviation (σ), and the skewness (ν) are estimated both for the subbands and for the whole TFR. Therefore, $12 \times 19 = 228$ engineered features are extracted and are used to train 2 stacked autoencoders (UL). (C) The first autoencoder (AE_1, 228:40:228) compresses the input representation in 40 parameters ($\mathbf{h_1}$). The second autoencoder (AE_2, 40:20:40) compresses the 40 learned features in 20 higher-level paraments ($\mathbf{h_2}$). Finally, a classifier with a single hidden layer ($\mathbf{h_3}$) of 10 neurons is trained (SL). The whole DL processor is possibly fine-tuned to improve the performance of the classification tasks: CJD-AD, CJD-ENC, CJD-HC.

$\underline{\underline{\Phi}}$ is the learned matrix of the encoding layer of the MLP-NN. The loss/cost function to be minimized through learning is given by:

$$\mathcal{L} = \|\underline{\widetilde{x}} - \underline{x}\|^2 + \lambda \left\|\underline{\underline{\varphi}}\right\|^2$$

where λ is the regularization coefficient, $\underline{\underline{\varphi}}$ is the matrices' weights, $\underline{x}$ is the input vector, and $\underline{\widetilde{x}}$ is the approximate (learned) output that reconstructs $\underline{x}$ through the training of the AE, i.e.,

$$\underline{\widetilde{x}} = \underline{\sigma}\left[\underline{\underline{\Psi}}\left(\sigma\left(\underline{\underline{\Phi}}\,\underline{x}\right)\right)\right]$$

Here, the aim of the MLP-NN is to learn a suitable representation of the input vector and it can be traded off with the quality of reconstruction of the input; accordingly, the choice of an optimal λ is not a strong constraint.

5.2 Summary of the proposed method for EEG classification

The DL approach can solve a binary classification problem (0: healthy subjects, 1: patient with disease) or a multiclass problem (i.e., either different stages of a degenerative brain disease or differentiation between diseases). This approach can include both feature-engineering and data-driven steps aiming to represent discriminative information from the available data hardly emerging from the visual inspection of the EEG recordings. The available EEG database (including all of the considered categories of subjects) passes through a processing chain that can be resumed as follows:

(1) Artifact rejection by clinical (visual) inspection: the segments of signal affected by evident artefactual components are cut from all the recording channels;
(2) The residual recordings are subdivided into nonoverlapping epochs of 5 s duration through a moving window procedure;
(3) A time-frequency analysis of the signals is carried out: the continuous wavelet transform (CWT) with Mexican Hat mother wavelet has been used in Ref. [30] but other time-frequency representations can be used; in particular, the empirical mode decomposition (EMD) can yield the advantage of being fully data-driven [31];
(4) Extraction of the relevant "engineered" features from the TFMs possibly taking into account the relevant brain rhythms (as an example, mean values standard deviations and skewness of the wavelet coefficients);
(5) Detection of evident outliers in the features; these values may be generated by segments of artifacts that have not been detected in step 1 and can be dropped out.

According to this procedure, [30], a vector of features has been generated from the TFMs; the resulting input vector includes 228 elements. The successive steps, based on the DL approach, combine the single-channel features, thus exploiting the multivariate nature of the EEG signal. A final classification stage, based on support vector machines (SVM) trained by SL, outputs the classification result. The DL-based system, after global

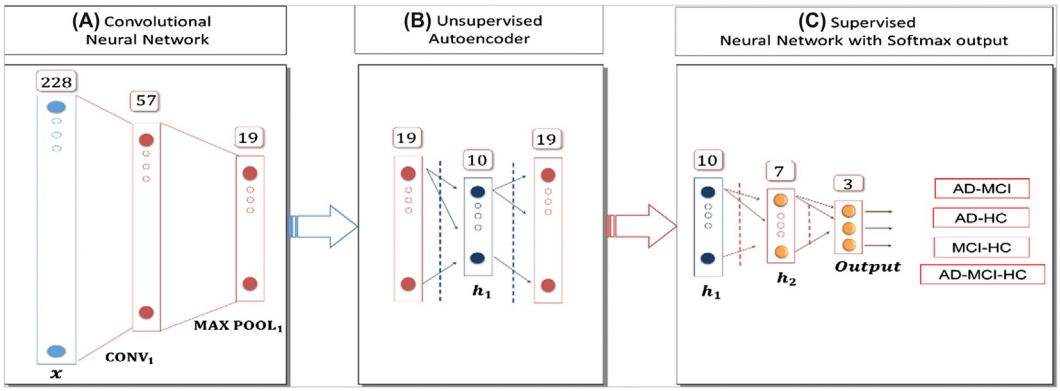

FIG. 13 Flowchart of the method presented in Ref. [32]. Extraction of the 228 features vector as shown in Fig. 12A and B. DL architecture: in the first stage, a convolutional layer (CONV1) performs the convolution operation between the 228 input features and 19 masks of 12 elements, outputting a vector of 57 features. Then, a max pooling layer (MAX $POOL_1$) reduces the feature vector size to 19. An autoencoder compresses the 19 outputs to 10 latent variables that form the input to the final fully connected MLP-NN with 7 hidden neurons, for binary (ADMCI, AD-HC, MCI-HC) and 3-way classification (AD-MCI-HC) tasks. In the figure, the output layer is specialized for the 3-way classification.

fine-tuning of the network by backpropagation, provided an average classification accuracy of around 90%, with similar sensitivity and specificity.

5.3 Deep convolutional neural networks

A second approach to DL-EEG has been introduced by [32]. It is based on a customized CNN for discriminating EEG recordings of subjects with Alzheimer's disease (AD), mild cognitive impairment (MCI), an early form of dementia, and healthy controls (HC). Fig. 13 shows the flowchart of the method. The EEG recordings have been transformed in the time-frequency domain, on an epoch by-epoch basis. Then, a grand total of 228 time-frequency features have been extracted and used as input vectors to the deep network. The DL model includes a convolutional layer followed by a sigmoidal nonlinearity, a max pooling layer, an autoencoder, and a classification single hidden layer MLP-NN. More than one convolutional-pooling stage can be used to generate higher-level features. The combined DL processor globally reduced the dimensionality of the input feature space from 228 to 10 latent features, providing very good performance in both binary (83% accuracy) and 3-way classification (82% accuracy) as reported in Table II of Ref. [32].

5.4 Other DL approaches

Other researchers have faced the problem of DL-EEG pattern classification. In particular, Zhao and He [33] proposed a 3-stacked restricted Boltzmann machines (RBM) structure for the classification of AD patients. They claimed to reach 92% of accuracy.

Other papers focused on the detection of epileptic seizures in EEG recordings. [34] modeled a DBN model for the classification and anomaly detection of epileptic patterns.

[35] developed a CNN for seizure prediction achieving zero-false alarm on 95% of patients analyzed, whereas Turner et al. [36] proved the effectiveness of the DBN algorithm in seizure detection (F-measure up to 0.9). Recently, DL networks have been also applied in the fast-growing field of brain-computer interface (BCI) as a rapid serial visual presentation (RSVP) task [37], steady-state visual evoked potential (SSVEP) classification [38], and P300 waves detection [39].

6 Future directions of research

The examples presented in this chapter aimed to show that DL can be a powerful tool to assist the solution of difficult biomedical signal processing problems like discriminating brain states to differentiate brain pathological conditions or to interpret tasks in BCI applications, like the classification of the left and right hand in motor imagery. The use of EEG is the gold standard in both cases. It is cheap and noninvasive and can be repeated easily being commonly well accepted by patients. Furthermore, the relevant data can be acquired through the increasingly popular wearable devices that allow capturing continuously physiological and functional data in both well-being and healthcare applications. This potentially rich information content can be transmitted through Bluetooth and smartphone channels for remote monitoring. This opens relevant possibilities in the immediate care of patients, for example, in life-threatening situations like epileptic absences. However, it also raises novel challenges to DL, like the resource-constrained use of low-power devices. The EEG is a complex signal, i.e., a multivariate nonstationary time series, which is inherently high-dimensional taking into account time, spectral, and spatial (channel) dynamic evolution. Recently, various DL architectures have been proposed to decode disease- or task-related information from the raw EEG recording with and without handcrafted features. Higher-level features extracted from DL can be analyzed, visualized, and interpreted to yield a different perspective with respect to conventional engineered features. Despite the exponential growth of research papers in DL, in most cases, a black-box approach is yet provided. In what follows, some of the critical issues of presently investigated DL are briefly summarized.

6.1 DL Explainability/interpretability

General methods to interpret how DL networks take decisions are basically lacking. There is no full theoretical understanding of how learning evolves in DL networks and how it generates their inner organization. This unsolved lack of ability to explain decisions to clinicians prevents the practical use of any predictive outcome. Some information-theoretic-based model has been proposed to “open the black box”: in particular, it has been suggested that the network optimize the information bottleneck trade-off between prediction and compression in the successive layers [40]. Essentially, it has been shown that DL spends most of the information available in the database of training for learning efficient representations instead of fitting the labels. This consideration seems to confirm

the importance of UL techniques, for which unsatisfactory algorithms have been devised so far [41]. Future advances in UL will focus on finding structural information on the input signals and in building generative models: generative adversarial networks are indeed highly promising directions of research [42].

6.1.1 explainable artificial intelligence (xAI)

The opaqueness of DL (AI)-models is denoted as *black-box behavior*. It has been recently explored by the so-called *explainable artificial intelligence (xAI)*. Explainability refers to the development of computational methodologies able to guarantee transparency in AI systems; for example, they can show which input area mostly contributed to achieve a specific performance and explaining how [43]. Some common xAI techniques are described in the following paragraphs.

6.1.2 Occlusion sensitivity analysis

Occlusion sensitivity analysis (OSA) is used to measure the *sensitivity* of a trained CNN to different regions of an input image [44]. OSA procedure consists of systematically occluding areas of the input data by using a moving grey mask and evaluating the related effect on the network output. For each position of the occluding mask, the occluded image is fed into the trained CNN to estimate the deterioration of the discrimination performance. Such modifications are used to plot the *heatmap* or *saliency maps*, which can reveal the most relevant regions of the input image for the classification task. In this representation, the areas mostly involved in recognizing a class are depicted with red coloration; in contrast, the areas less relevant for the discrimination task are depicted with a blue coloration.

6.1.3 Gradient-weighted class activation mapping (Grad-CAM)

The gradient-weighted class activation mapping (GradCAM) is one of the most employed xAI techniques typically used to better understand the CNN-based models [45]. Indeed, Grad-CAM allows to detect which input area is most significant for predictions. Let o^c the score of a certain class c. The gradient $\frac{\partial o^c}{\partial R^n}$ of the last convolutional layer is first calculated, where R^n are the features maps (with n number of features representations). Then, the global average pooling is evaluated to estimate the neuron importance weights $\widehat{w}_n^c$, defined as:

$$\widehat{w}_n^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial o^c}{\partial R_{i,j}^n}$$

where Z denotes the number of pixels in a feature map and (i,j) identifies the pixels. Finally, a weighted combination of R^n is computed:

$$\widehat{w}_n^c = ReLU\left(\sum_n \widehat{w}_n^c R^n\right)$$

where *ReLU* is the rectified linear unit transfer function. The result of this operation is known as the *Grad-CAM map* or *importance map* where the most important input areas for the classification are detected with coloration from blue (low importance) to red (high importance).

6.1.4 xAI approaches in biomedical engineering applications

In Ref. [46], the authors proposed an xAI approach for EEG-based brain-computer interface systems. Fig. 14 shows the flowchart of the method. In particular, the explainability of the developed CNN-based system was investigated to provide a better understanding of cortical sources activation when the brain is planning the hand's close/open (HC/HO) movement. To this end, an occlusion sensitivity analysis was performed to explore which cortical areas were mostly involved in the classification process and a *k*-means clustering technique was also employed to detect the highest saliency region. The cortical sources were then mapped to the cortical representation. Results showed that the central region

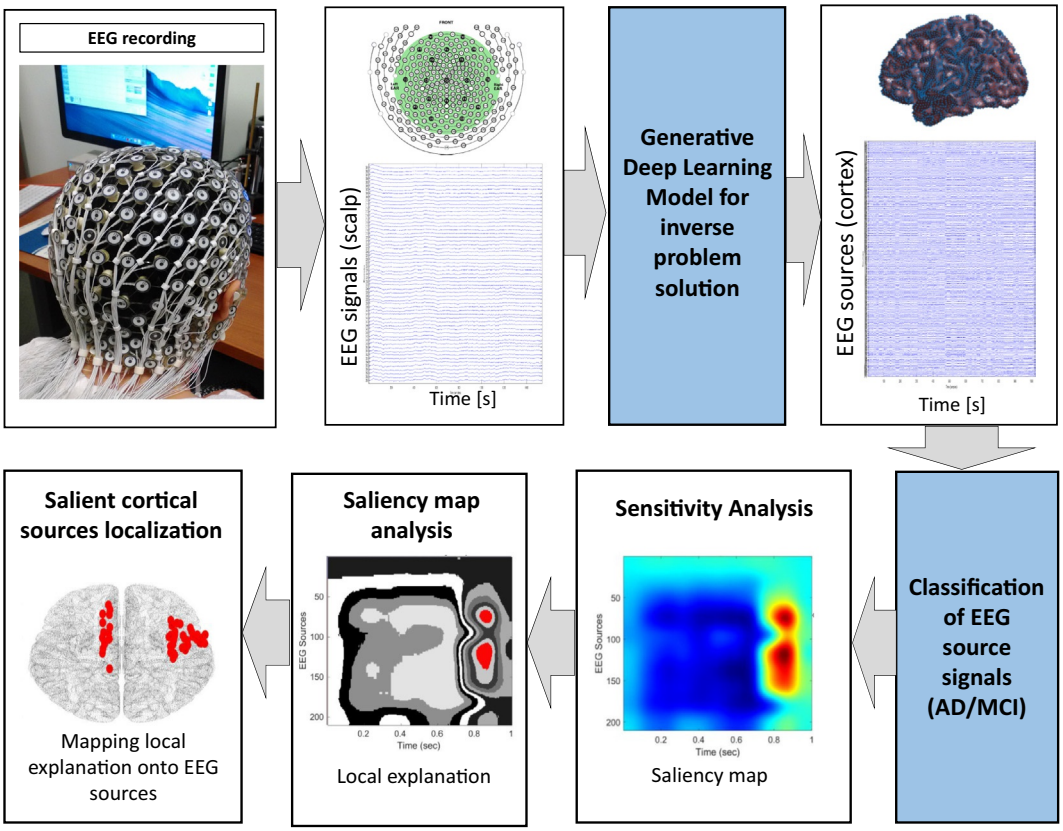

FIG. 14 Flowchart of the method presented in Ref. [46].

(close to the longitudinal fissure) and the right temporal zone of the premotor together with the primary motor cortex appear to be primarily involved.

Another approach based on xAI was introduced by Ref. [47] to longitudinally monitor subjects affected by mild cognitive impairment (MCI) by using high-density electroencephalography (HD-EEG). In particular, the proposed method consisted in mapping windows of HD-EEG into channel-frequency (HD-CF) representations using the power spectral density (PSD) estimation, subsequently used as input to a custom CNN, trained to classify the HD-CF maps as "T0" (MCI state) or "T1" (AD state). The Grad-CAM approach was applied to "explain" the approach taken by CNN. The methodology was capable of detecting which EEG channels (i.e., head region) and range of frequencies (i.e., subbands) resulted in more active in the progression to AD. Results showed that the activation of different EEG channels observed the main information that was included only in the subband. As an example, Fig. 15 reports the Grad-CAM maps of Subject 01 at time T0 (MCI state) and at time T1 (AD state).

6.2 Advanced learning approaches in DL

One of the problems with DL is the overfitting of the training data, as the number of free parameters is often quite high, compared to the size of the training set; in this case, DL performs poorly in generalization, i.e., on held-out tests and validation examples. This effect is particularly serious in a clinical setting, where the fresh data often refer to a

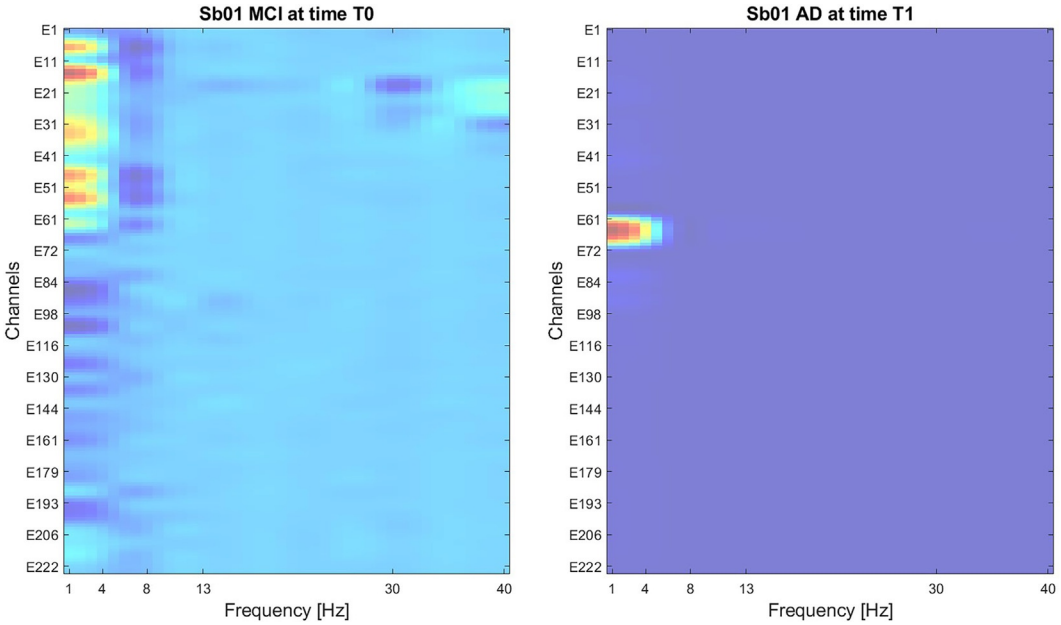

FIG. 15 Grad-CAM maps of Subject 01 at time T0 (MCI state) and at time T1 (AD state). *Red* color denotes areas with the highest relevance; vice-versa, *blue* color denotes areas with the lowest relevance [47].

novel patient. Several strategies for reducing the impact of overfitting have been proposed in the literature. One of these suggests randomly omitting half of the feature detectors on each training example [48]. The use of AEs with UL is also quite beneficial, particularly when the learning cost functions involve regularization terms, as in the dropout method. SAEs face the problem of vanishing gradients that become negligible during training; therefore, the DL network tends to learn the average of all the training examples. Furthermore, the AE treats all inputs equally and its representational capability aims to minimize the reconstruction error of each input. This is a shortcoming in the presence of noisy data. In addition, being trained without knowing the labels, the AE cannot discriminate between task-irrelevant and task-relevant information in the dataset. One of the approaches to deal with these problems is to make a pretraining of DL networks and to evaluate the architectural design (i.e., the number of layers and the related nodes) by means of information theoretical quantities, like entropy and mutual information [49]. In other works, the impact of the hidden nodes is measured simply by the variance of the output of the nodes. Indeed, any node with constant activation across different inputs fails to convey discriminative information about the input vectors. This observation can help reduce the size of the hidden layers during or posttraining. Another way proposed to minimize the limitations of deep architectures is to change the standard, biologically plausible, sigmoidal neuron nonlinearity, by substituting it with a strong nonlinear rectifier that helps to create a sparse representation with true zeros. The EEG is not sparse per se, but interchannel and intrachannel redundancy can be exploited to generate a block-sparse representation in suitable domains [50]. An evident advantage of rectifying neurons is that they allow the DL network to better disentangle information with respect to dense DL. In addition, by varying the number of active hidden neurons, and thus giving a variable-size flexible data structure, they allow to represent the effective dimensionality of the input data vectors.

6.3 Robustness of DL networks

DL is rapidly becoming a standard approach for the processing of medical and health data. In particular, DL provides the opportunity to automate the extraction of relevant features (as a difference with highly subjective interpretation of diagnostic data), integrate multimodal data, and combine the extraction stage with classification procedures. The classification performance is often limited as the available databases are typically small and incomplete, and preprocessing of data remains commonly a rather crucial step. The designed classifiers sometimes do not satisfy a check of robustness: it happens that trained models reduce their performance if small perturbations are applied to the examples. Adversarial perturbations are a relevant example. A rectifier-based sparse representation is typically robust to small input changes, as the set of nonzero features is well conserved. From geometric considerations, it has been shown that the high instability of DL networks is related to data points that reside very close to the classifier's decision boundary. As the robustness of DL is a critical requirement in a clinical setting, novel strategies for designing and training of DL schemes should be devised by the community in the years to come.

7 Conclusions

DL can yield appropriate tools for analyzing multivariate time-series data from a genuinely new perspective, which is both fully data-driven and automatic but can also take advantage of engineered features derived from standard signal processing methods, like frequency and time-frequency transforms. In recent years, the international community has shown enormous interest in DL and artificial intelligence, by funding several programs, in the public and private sectors. It is foreseen that these programs will favor a relevant growth of the economy and national gross value added (GVA). However, in this chapter it has been shown that DL is just a contingent development of ML and NN techniques originally proposed decades ago. This chapter focused on a rather limited aspect of DL, namely the processing of multivariate time series that is relevant for biomedical applications but also pertinent to the future development of IoT systems. The techniques here described can support a significant leap forward in the real-time processing of unstructured data and in clinical diagnosis. Open problems and limitations of DL have been discussed.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[2] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.

[3] G.E. Hinton, Training products of experts by minimizing contrastive divergence, Training 14 (8) (2006).

[4] D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pretraining help deep learning? J. Mach. Learn. Res. 11 (Feb) (2010) 625–660.

[5] A. Krizhevsky, I. Sutskever, G.E. Hinton, Image-net classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–s.

[6] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.

[7] M.D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q.V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, et al., On rectified linear units for speech processing, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 3517–3521.

[8] P.L. Nunez, R. Srinivasan, Electric Fields of the Brain: The Neurophysics of EEG, Oxford University Press, USA, 2006.

[9] S. Vaid, P. Singh, C. Kaur, EEG signal analysis for BCI interface: a review, in: 2015 Fifth International Conference on Advanced Computing Communication Technologies (ACCT), IEEE, 2015, pp. 143–147.

[10] G.R. Philips, J.J. Daly, J.C. Principe, Topographical measures of functional connectivity as biomarkers for post-stroke motor recovery, J. Neuroeng. Rehabil. 14 (1) (2017) 67.

[11] W.O. Tatum IV, Handbook of EEG Interpretation, Demos Medical Publishing, 2014.

[12] C.S. Herrmann, M. Grigutsch, N.A. Busch, EEG oscillations and wavelet analysis, in: T.C. Handy (Ed.), Event-Related Potentials: A Methods Handbook, MIT Press, 2005, pp. 229–259.

[13] F. Vecchio, C. Babiloni, R. Lizio, F.V. Fallani, K. Blinowska, G. Verrienti, G. Frisoni, P.M. Rossini, Resting state cortical EEG rhythms in Alzheimer's disease: toward EEG markers for clinical applications: a review, Suppl. Clin. Neurophysiol. 62 (2013) 223–236.

[14] J. Dauwels, S. Kannan, Diagnosis of Alzheimer's disease using electric signals of the brain. A grand challenge, Asia-Pac. Biotech News 16 (10n11) (2012) 22–38.

[15] J. Dauwels, K. Srinivasan, M. Ramasubba Reddy, T. Musha, F.B. Vialatte, C. Latchoumane, et al., Slowing and loss of complexity in Alzheimer's EEG: two sides of the same coin? Int. J. Alzheimer's Dis. 2011 (2011).

[16] F. Hatz, M. Hardmeier, N. Benz, M. Ehrensperger, U. Gschwandtner, S. Ruegg, C. Schindler, A.U. Monsch, P. Fuhr, Microstate connectivity alterations in patients with early Alzheimer's disease, Alzheimers Res. Ther. 7 (1) (2015) 78.

[17] J. Jeong, EEG dynamics in patients with Alzheimer's disease, Clin. Neurophysiol. 115 (7) (2004) 1490–1505.

[18] P. Parchi, A. Giese, S. Capellari, P. Brown, W. Schulz-Schaeffer, O. Windl, I. Zerr, H. Budka, N. Kopp, P. Piccardo, et al., Classification of sporadic Creutzfeldt-Jakob disease based on molecular and phenotypic analysis of 300 subjects, Ann. Neurol. 46 (2) (1999) 224–233.

[19] H.G. Wieser, K. Schindler, D. Zumsteg, EEG in Creutzfeldt–Jakob disease, Clin. Neurophysiol. 117 (5) (2006) 935–951.

[20] U. Aguglia, A. Gambardella, E. Le Piane, D. Messina, G. Farnarier, R. Oliveri, M. Zappia, A. Quattrone, Disappearance of periodic sharp wave complexes in Creutzfeldt-Jakob disease, Clin. Neurophys. 27 (4) (1997) 277–282.

[21] S. Gasparini, E. Ferlazzo, D. Branca, A. Labate, V. Cianci, M.A. Latella, U. Aguglia, Teaching neuroimages: pseudohypertrophic cerebral cortex in end-stage Creutzfeldt-Jakob Disease, Neurology 80 (2) (2013) e21.

[22] J. Malmivuo, R. Plonsey, Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields, Oxford University Press, USA, 1995.

[23] O.R. Ryynanen, J.A. Hyttinen, P.H. Laarne, J.A. Malmivuo, Effect of electrode density and measurement noise on the spatial resolution of cortical potential distribution, IEEE Trans. Biomed. Eng. 51 (9) (2004) 1547–1554.

[24] O.R. Ryynanen, J.A. Hyttinen, J.A. Malmivuo, Effect of measurement noise and electrode density on the spatial resolution of cortical potential distribution with different resistivity values for the skull, IEEE Trans. Biomed. Eng. 53 (9) (2006) 1851–1858.

[25] R. Srinivasan, D.M. Tucker, M. Murias, Estimating the spatial Nyquist of the human EEG, Behav. Res. Methods 30 (1) (1998) 8–19.

[26] A. Delorme, S. Makeig, EEGLab: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, J. Neurosci. Methods 134 (1) (2004) 9–21.

[27] V. Brodbeck, L. Spinelli, A.M. Lascano, M. Wissmeier, M.I. Vargas, S. Vulliemoz, C. Pollo, K. Schaller, C.-M. Michel, M. Seeck, Electroencephalographic source imaging: a prospective study of 152 operated epileptic patients, Brain 134 (10) (2011) 2887–2897.

[28] P. Mégevand, L. Spinelli, M. Genetti, V. Brodbeck, S. Momjian, K. Schaller, C.M. Michel, S. Vulliemoz, M. Seeck, Electric source imaging of interictal activity accurately localises the seizure onset zone, J. Neurol. Neurosurg. Psychiatry 85 (1) (2014) 38–43.

[29] S.F. Storti, I.B. Galazzo, A. Del Felice, F.B. Pizzini, C. Arcaro, E. Formaggio, R. Mai, P. Manganotti, Combining ESI, ASL and PET for quantitative assessment of drug-resistant focal epilepsy, NeuroImage 102 (2014) 49–59.

[30] F.C. Morabito, M. Campolo, N. Mammone, M. Versaci, S. Franceschetti, F. Tagliavini, et al., Deep learning representation from electroencephalography of early-stage Creutzfeldt-Jakob disease and features for differentiation from rapidly progressive dementia, Int. J. Neural Syst. 27 (02) (2017) 1650039.

[31] A. Zahra, N. Kanwal, N. Rehman, S. Ehsan, M.D.-M. KD, Seizure detection from EEG signals using multivariate empirical mode decomposition, Comput. Biol. Med. 88 (2017) 132–141.

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

[33] Y. Zhao, L. He, Deep learning in the EEG diagnosis of Alzheimer's disease, in: Asian Conference on Computer Vision, Springer, 2014, pp. 340–353.

[34] D. Wulsin, J. Gupta, R. Mani, J. Blanco, B. Litt, Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement, J. Neural Eng. 8 (3) (2011) 036015.

[35] P. Mirowski, D. Madhavan, Y. LeCun, R. Kuzniecky, Classification of patterns of EEG synchronization for seizure prediction, Clin. Neurophysiol. 120 (11) (2009) 1927–1940.

[36] J. Turner, A. Page, T. Mohsenin, T. Oates, Deep belief networks used on high-resolution multichannel electroencephalography data for seizure detection, in: 2014 AAAI Spring Symposium Series, 2014.

[37] R. Manor, A.B. Geva, Convolutional neural network for multi-category rapid serial visual presentation BCI, Front. Comput. Neurosci. 9 (2015).

[38] H. Cecotti, A. Graeser, Convolutional neural network with embedded Fourier transform for EEG classification, in: 19th International Conference on Pattern Recognition, ICPR 2008, IEEE, 2008, pp. 1–4.

[39] H. Cecotti, A. Graser, Convolutional neural networks for P300 detection with application to brain-computer interfaces, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 433–445.

[40] R. Schwartz-Ziv, N. Tishby, Opening the Black-Box of Deep Neural Networks via Information, arXiv:1703.00810v3, 2017.

[41] Y. Liu, J. Chen, L. Deng, Unsupervised sequence classification using sequential output statistics, in: Proc. NIPS, 2017.

[42] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, MA, 2016.

[43] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI, Inform. Fusion 71 (2021) 28–37.

[44] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, Cham, 2014, pp. 818–833.

[45] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[46] C. Ieracitano, N. Mammone, A. Hussain, F.C. Morabito, A novel explainable machine learning approach for EEG-based brain-computer interface systems, Neural Comput. & Applic. (2021) 1–14.

[47] F.C. Morabito, C. Ieracitano, N. Mammone, An explainable Artificial Intelligence approach to study MCI to AD conversion via HD-EEG processing, Clin. EEG Neurosci. (2021). 15500594211063662.

[48] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, Salakhutdinov, Improving Neural Networks by Preventing Co-adaption of Feature Detectors, 2012. arXiv:1207.0580v1.

[49] Y. Furusho, T. Kubo, K. Ikeda, Roles of pre-training in deep neural networks from information theoretical perspective, Neurocomputing (2017) 76–79.

[50] D. Labate, F. La Foresta, I. Palamara, G. Morabito, Bramanti, Z. Zhang, F.C. Morabito, EEG complexity modifications and altered compressibility in mild cognitive impairment and Alzheimer's disease, in: Proceedings of the 23rd Italian Workshop on Neural Networks (WIRN 2013), 2013.