This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2018.2889512, IEEE Transactions on Biomedical Engineering

Deep Convolutional Neural Networks for feature-less automatic classification of Independent Components in multi-channel electrophysiological brain recordings

Pierpaolo Croce, Filippo Zappasodi, Laura Marzetti, Arcangelo Merla, Vittorio Pizzella, Antonio Maria Chiarelli

***Abstract*— *Objective:* Interpretation of the Electroencephalographic (EEG) and Magnetoencephalographic (MEG) signals requires off-line artifacts removal. Since artifacts share frequencies with brain activity, filtering is insufficient. Blind source separation, mainly through Independent Component Analysis (ICA), is the gold-standard procedure for the identification of artifacts in multi-dimensional recordings. However, a classification of brain and artifactual Independent Components (ICs) is still required. Since ICs exhibit recognizable patterns, classification is usually performed by experts' visual inspection. This procedure is time-consuming and prone to errors. Automatic ICs classification has been explored, often through complex ICs features extraction prior to classification. Relying on Deep Learning ability of self-extracting the features of interest, we investigated the capabilities of Convolutional Neural Networks (CNNs) for off-line, automatic artifact identification through ICs without feature selection. *Methods:* A CNN was applied to spectrum and topography of a large data-set of few thousand samples of ICs obtained from multi-channel EEG and MEG recordings acquired during heterogeneous experimental settings and on different subjects. CNN performances, when applied to EEG, MEG, and combined EEG and MEG ICs, were explored and compared with state of the art feature-based automatic classification. *Results*: Beyond state-of-the-art automatic classification accuracies were demonstrated through cross-validation (92.4% EEG, 95.4% MEG, 95.6% EEG+MEG). *Conclusion:* High CNN classification performances were achieved through heuristically selection of machinery hyperparameters and through the CNN self-selection of the features of interest. *Significance*: Considering the large data availability of multi-channel EEG and MEG recordings, CNNs may be suited for classification of ICs of multi-channel brain electrophysiological recordings.**

***Index Terms*— Deep Learning, Electroencephalography, Magnetoencephalography, Independent Component Analysis, Artifact Identification**

I. Introduction

In an active postsynaptic neuron, a negative voltage between neural dendrites and other locations along the neuron is generated. Within a small brain compartment in which dendritic arborization are parallel and follow a main direction, such situation can be modelled as a current dipole generating an electromagnetic field [1]. Both electrical potentials and magnetic fields, generated from the dipole in this compartment, can be measured non-invasively by sensors located on or close to the scalp. The technology that measures electrical potentials is called Electroencephalography (EEG), while the technology that measures magnetic fields is called Magnetoencephalography (MEG). EEG and MEG provide sub-milliseconds temporal resolution of brain electrical activity and, despite providing a spatial resolution of a few centimeters, they are non-invasive tools widely employed in the interpretations of neuronal communication, computation and brain spatiotemporal networks organization [1]. Electrical potentials and related magnetic fields are particularly small in amplitude and are attenuated by volume conduction effect [2], thus, both EEG electrodes and MEG sensors need to be particularly sensitive. Such elevated sensitivity implies sensors' susceptibility to extracerebral electrical signals. Recorded electrical activity that is not of cerebral origin is termed 'artifact' and can be divided into of physiological and non-physiological origin. Common physiological artifacts are generated by muscle activity, eye movements, cardiac pulsation and respiration [3]. Common non-physiological artifacts are generated by sensors' malfunction, alternating current supply, subjects' movements and electric and magnetic field generated by other instrumentations close to the recording environment.

For a correct interpretation of the EEG and MEG signals, artifacts removal algorithms need to be applied off-line. Filtering techniques (band-pass filtering) are the simplest procedures for artifacts suppression. However, since artifacts often share common frequencies with brain activity, standalone filtering procedures are insufficient and more sophisticated algorithms are generally required. Indeed, many off-line approaches for the identification and removal of artifacts were explored and validated. We here name: adaptive filtering [4], Wiener filtering [5], Bayes filtering [6], regression [7], electrooculogram correction [8], wavelet transform (WT) [9], empirical mode decomposition (EMD)

P. Croce, F. Zappasodi, L. Marzetti, A. Merla, V. Pizzella and A.M. Chiarelli work within the Department of Neuroscience, Imaging and Clinical Sciences and the Institute for Advanced Biomedical Technologies, Università G. d'Annunzio, Chieti, 66100, Italy. (e-mail: antonio.chiarelli@unich.it)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2018.2889512, IEEE Transactions on Biomedical Engineering

[10], nonlinear mode decomposition (NMD) [11] and blind source separation [12]. Among the different algorithms developed, the blind source separation procedures rely on multi-dimensional (multi-channel) data, do not use any training data, and do not assume a priori knowledge about parameters of mixing convolutive and filtering systems, thus being the favorite procedures when multiple channels information are provided [13]. Within blind source separation, Independent Component Analysis (ICA) is the most common procedure. ICA is a fully data-driven statistical approach which exploits multi-channel EEG or MEG information, allowing to decompose the signal into independent (both in space and time) components (IC) [14]. After ICA is performed, it is possible to classify the ICs as sources related to brain activity or of artifact origin. After classification, artifactual ICs are subtracted from the measured signal allowing to obtain an estimate of artifact-free electrical brain activity. Since brain and artifactual ICs exhibit recognizable time, frequency and topographic patterns, classification is generally performed by experts' visual inspection. However, this procedure is time-consuming and prone to human error. In fact, semiautomatic and automatic procedures have been developed over time [3], [15]–[17] .

Automatic procedures currently rely on extracting ICs' features by means of quantities of interest derived from ICs time series (e.g., Variance, Skewness, Kurtosis, Maximum Amplitude, Range, Max First Derivative, Shannon and Deterministic Entropy), from topographic patterns (e.g. Horizontal Sobel, Large Horizontal Gradient, Range Filter, Three Dimensional Local Binary Pattern, Range Within Pattern, Spatial Distance of Extrema, Spatial Mean Activation, Border activation) and from spectra (e.g., 1/f Fit, Power within Frequency Bands) before feeding those features to classifiers (e.g., Support Vector Machine, Linear Discriminant Analysis) [18], [19].

With respect to the automatic classification problem, Artificial Neural Networks (ANNs) are particularly suitable for classification of ICs. In a nutshell, ANNs are collections of connected non-linear processing units, called artificial neurons, that resemble biological neural networks in animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules [20]. In Huang et al. [21], the authors suggested that advanced machine learning tools such as ANN can be effective in classification of ICs. An ANN perceptron was proposed by Sovierzoski et al. [22] to classify eye blink in EEG signal. The work in Nguyen et al. [23] proposed a wavelet neural network for removal of eyes ocular artifacts from EEG. Raduntz and colleagues [19] proposed an ANN for recognition of brain and artifactual components based on features of interest extracted from ICs. The natural evolution of ANNs are Deep Neural Networks (DNNs, also referred to as Deep Learning algorithms) that, thanks to the current high computational power, architectures and learning algorithms development, are increasing their popularity. DNNs are ANNs [20], [24] composed of many layers of neurons. Each of these layer uses the output from the previous layer as input and all, or part of the neurons from consecutive layers are connected. Each layer of DNNs learn representations of data with different levels of abstraction. DNNs representations are generally learned by modifying the network's parameters relying on the backpropagation algorithm that distributes the error calculated at the output back through the network layers [25]. DNNs structure should be instead heuristically selected a priori or determined through computationally demanding hyper-parameters optimization algorithms [26]–[28]. Algorithm development allowed for Deep Learning evolution: efficient learning algorithms were implemented to avoid local minima in the objective function and poor generalization (over-fitting) [29] as well as new neuron's activation functions (such as Rectified Linear Unit Function, ReLU function [30], [31]) that dampened the vanishing gradient problem [32] of multilayers ANN.

DNNs can perform very complex, non-linear, transformations and/or classifications, greatly increasing shallow ANN [33] and other classifiers performances (Linear Discriminant Analysis, Support Vector Machine, etc.) [34]. DNNs can reach unprecedented classification outcomes when applied to signals (e.g. speech and language processing) and/or to images [35]–[38]. Because of their performances, these algorithms are also receiving attention within the biomedical field [39], [40]. When dealing with mono-dimensional signals or multi-dimensional images, a further improvement to DNNs was the implementation of Convolutional Neural Networks (CNNs). CNNs are Neural Networks where neurons are connected to sliding portions of signals and/or images that are close in time and/or space [37], [41]. They allow to encode temporal and/or spatial information and to decrease the number of free parameters with respect to standard, fully connected, DNNs. Since neurons of CNNs are combined in groups and connected to sliding portions of the signals or images, the free parameters of the neurons, thanks to the learning procedure, generate peculiar filters, allowing for automatic pattern identifications and avoidance of a-priori feature selection. CNNs have recently enjoyed great success in image and video recognition as well as in biomedical signal/imaging applications [42].

CNNs were also employed on EEG and MEG recordings for task classification (see [43]–[45] for a review).

In this paper we report investigation of the capabilities of CNNs for off-line, fully automatic, feature-less, artifact identification applied to ICs extracted by means of ICA of a large data-set of multichannel electrophysiological recordings (EEG and MEG). Moreover, we compare the results with state-of-the-art feature-based classification procedures.

II. Methods

A. Independent Component Analysis

ICs from both EEG and MEG data were extracted with the FASTICA algorithm [14]. FASTICA allows the computation of an optimal number of independent components along with unmixing and mixing matrices. FASTICA relies on the maximization of non-Gaussianity using the fourth-order

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2018.2889512, IEEE Transactions on Biomedical Engineering

cumulant of the signal (kurtosis). Maximizing the norm of the kurtosis leads to the identification of non-Gaussian sources. The hyperparameters employed in the FASTICA ICs extraction applied to brain electrophysiological recordings are already validated and described elsewhere [46].

B. EEG and MEG dataset

EEG ICs were extracted from different datasets collected in our facilities with a full-head, 128 channels EEG system (Electrical Geodesic Inc, EEG System Net 300). EEG recording length varied from a minimum of 5 minutes to a maximum of 22 minutes (mean=11 min, standard deviation=3 min). Sampling frequency was set to 500 Hz for all the experiments that consisted of various tasks exploring rest, attention and movement. Raw common referenced signals were filtered between 1 Hz and 80 Hz and notch-filtered at 50 Hz (2nd order Chebyshev digital filters) prior to FASTICA evaluation. A total of number of 1067 ICs for the EEG dataset were obtained. The EEG ICs were classified by trained experts. The trained expert classification relied on time course, spectral and topographic characteristics of the ICs. 503 brain ICs and 564 artifact ICs were classified.

MEG ICs were extracted from two different datasets acquired with different instrumentation. The first ICs dataset were extracted from recordings collected in our facilities with a MEG system constituted of 153 DC superconducting quantum interference devices (SQUIDs) magnetometers placed on a helmet- shaped surface covering the whole scalp at a pace of around 3.2 cm [47]. The second dataset consisted of ICs extracted from resting state recordings performed within the Human Connectome Project whose data are freely available online [48]. MEG recordings length varied from a minimum of 5 minutes to a maximum of 15 minutes (mean=11 min, standard deviation=3 min). Sampling frequency was set at 1025 Hz for the in-house system whereas it was set at 2034 Hz for the connectome data. Data were acquired during various tasks exploring rest, attention and movement. Raw signals were filtered between 1 Hz and 130 Hz and notch-filtered at 50 Hz or 60 Hz (2nd order Chebyshev digital filter) prior to FASTICA application. A total number of 2572 ICs were extracted from the first dataset. A total number of 2177 ICs were extracted from the second dataset. A total of 4749 ICs for the MEG dataset were obtained. The MEG ICs were classified by trained experts. The trained experts' classification relied on time course, spectral and topographic characteristics of the ICs. 2019 brain ICs and 2730 artifact ICs were classified.

Figure 1 reports examples of normalized ICs time courses and related spectra, topographic maps, and fast Fourier transforms of the maps. Figure 1a reports ICs classified by experts as artifact related to heart activity (top row), eye movement (middle row) and sensor malfunction (bottom row). The artifacts were identified by means of typical temporal, spectral and topographic features (periodic signal with heart frequencies components, sudden temporal spikes, 1/f decay in the spectrum). Figure 1b reports instead ICs classified by experts as of brain origin. A 10 Hz activity peak is clear in the spectra together with dipole like topographic maps.

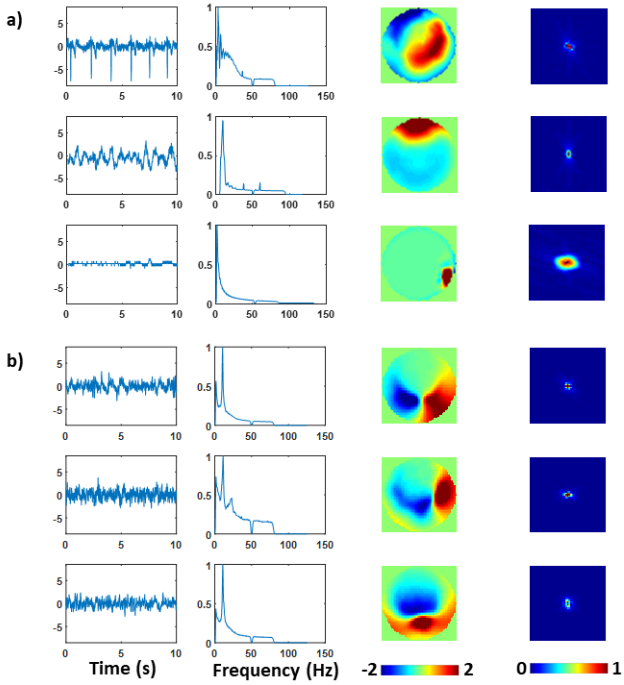

Fig. 1. Example of normalized ICs time courses (10 second window zoom, left column) and related normalized spectra (250 Hz sampling frequency, 0.12 Hz resolution, center-left column), normalized topographic map (51x51 pixels, center-right column), and fast Fourier transforms of the maps (51x51 pixels, right column). (a) ICs classified by experts as artifact related to heart activity (top row), ocular movements (middle row) and sensor malfunction (bottom row). The artifacts are identified by typical temporal, spectral and topographic features (periodic signal with heart frequencies components, sudden temporal spikes, 1/f decay in the spectrum) (b) ICs classified by experts as brain activity. A 10 Hz activity peak is clear in the spectrum together with topographic maps likely generated by a dipolar brain current source.

C. EEG and MEG data pre-processing

In order to develop an algorithm independent of the recording length, CNNs were applied to ICs spectrum and its weights topographic distribution, without considering the time domain signal.

Each IC time course was resampled at 250 Hz to provide a sufficient, but not redundant, frequency information (Figure 1). Power spectrum of the IC time course was computed using the following parameters: Hamming window, 2048 samples, 8.192 seconds, with a 50% overlap [49]. Each IC's spectrum (ICSPECT) consisted of 1025 points, spanning frequencies up to 125 Hz with a frequency resolution of ~0.12 Hz. IC weights were organized in a topographic fashion by considering EEG and MEG channels' locations for the different data-set and by warping them into a unitary circle by means of a commonly utilized EEGLAB algorithm [50]. Topographic images were constructed by interpolations of IC weights at the warped channel location in matrices of 51x51 pixels (ICMAP). The spatial sampling provided by the chosen matrices size was heuristically chosen after investigation of the frequency content of the topographic images obtained (Figure 1). Both topographic images and spectra were normalized based on z-score computation (subtraction of the sample mean and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2018.2889512, IEEE Transactions on Biomedical Engineering

division by the sample standard deviation) and Min-Max normalization before feeding them to the CNN.

D. EEG and MEG data selection for CNN input

When training CNNs, it is important to account for the different classes numerosity in the training set. In fact, unbalanced classes create preferences for the machinery to optimize a specific class (generally the one with the higher numerosity) distorting the learning phase. Generally, this problem can be overcome by employing class-specific weights in the objective function (lower weights for the class with higher numerosity and vice-versa) or by simply down-sampling the classes to the class with the lower numerosity. In this work, in order to provide balanced classes to the CNN classifier, EEG and MEG ICs were randomly down-sampled. When CNN classification performance was evaluated on standalone EEG ICs, a balance between brain and artifact ICs was obtained by down-sampling artifact ICs to 503. Thus, a total EEG dataset of 1006 ICs was obtained. When CNN classification performance was evaluated on standalone MEG ICs, a balance between brain and artifact ICs was obtained by down-sampling artifact ICs to 2019. Thus, a total MEG dataset of 4038 ICs was obtained. When CNN classification performance was evaluated on combined EEG and MEG ICs, a balance between brain and artifact ICs, as well as EEG and MEG dataset was obtained by down-sampling artifact EEG ICs, as well as brain and artifact MEG ICs, to 503. Thus, a total combined EEG and MEG dataset of 2012 ICs was obtained.

E. Convolutional Neural Network and Independent Components Classification

Evolution of DNNs, Convolutional Neural Networks are meant to deal with data in the form of multiple arrays, for example 2D images, but also 1D signals and sequences. Local connections, shared weights, pooling and the use of many layers are fundamental ideas behind CNNs. In general, a typical CNN architecture consists of an input layer and several successive layers that are divided in three categories: convolutional, pooling and fully connected layers (typically the last layers) [20].

The structure of the CNN employed in this work was heuristically chosen, in similarity with previously reported CNN structures on biological image analysis and classification [42]. The structure is reported in Figure 2. The CNN was composed of 2- parallel, 3 convolutional layers and one fully connected layer. The input layers were respectively composed of normalized topographic images of the ICA mixing weights (ICMAPs) and spectrum of ICs (ICSPECTs). The ICMAPs were 51x51x1 images whereas ICSPECT were 1025x1x1 images. The first two dimensions represent the height and the width of the images whereas the third dimension represents different levels of the maps (in general more than one level is considered in input when RGB images or time-frequency maps are analyzed, and more levels are generated in the convolutional layers by applying multiple filters on each image). The parallel structures consisted of convolutional, non-linear and pooling layers performing features extraction and dimension reduction. The convolutional layer generally performs convolutions between the original map and a set of weights (called a filter bank) by sliding the filter over the image and computing the dot product (convolution) between the filter and the map. In detail, convolutional layers consist of rectangular grids of neurons. Neurons in a given grid take inputs from a rectangular sliding section of the previous layer and the weights for this rectangular section are constant during the sliding process. Multiple grids for each convolutional layer can be employed generating different maps. Thus, the convolutional layer performs different (equivalent to the number of grids) image convolutions of the previous layer, where the weights of each grid specify the convolution kernel. The convolutional layer is often called feature extraction layer. The sliding step, that determinates the resolution of the extracted feature, is called stride. For each convolutional layer and for both ICMAPs and ICSPECTs we employed 4, 8 and 16 filters, respectively. All filters were of size 5x5 and 5x1 for ICMAPs and ICSPECT feature extraction, respectively with a stride of 1. The output of the convolutional layer is then passed through a non-linear transformation. As a non-linear processing function, we decided to employ the state of the art Rectified Linear Unit (ReLU) function, which was proven to dampen the vanishing gradient problem providing better performance than other non-linear functions (such as the hyperbolic tangent or the sigmoid function) [30]. Hidden neurons output, when ReLU function is employed, can be written as:

$$y = \begin{cases} 0, & wx + b \leq 0 \\ wx + b, & wx + b > 0 \end{cases} \quad (1)$$

where x is the input vector, w and b are the weight vector and bias, respectively, and y is the output vector. The pooling layer performs a spatial pooling (also called subsampling or down sampling) reducing the dimensionality of each feature map retaining the significant information. The pooling layer takes small rectangular regions from the convolutional layer and subsamples those regions to produce a single output from each region. There exist different procedures for spatial pooling. For the implemented CNN we employed the common procedure of MaxPooling, where the largest element from the rectified feature map is retained. A 4x4 MaxPooling and a 11x1 MaxPooling was implemented for the ICMAPs and ICSPECTs respectively, to reduce the dimensionality of each feature to 1 after the three convolutional layers and before the fully connected layer. In fact, the output of the last pooling layer acted as an input to the fully connected layer. The term "fully connected" implies that every neuron in the previous layer is connected to every neuron in the next layer. Fully connected layer is generally a Multi-Layer Perceptron that, in our case, used the ReLU function for non-linear transformation. The fully connected layer (32 neurons) was introduced to summarize and share information between the parallel convolutional structures (composed of 16 features each after the last feature extraction layer). Whereas the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2018.2889512, IEEE Transactions on Biomedical Engineering

convolutional layers with associated pooling stage acted as features extractors, the fully connected layers acted as classifier. Finally, a SoftMax activation function constituted the output layer to provide a probability of the inputs of being in the different classes of the dataset. A two neuron SoftMax layer was employed as output layer.

The operation of the SoftMax layer can be written as:

$$\begin{bmatrix} P_{brain} \\ P_{artifact} \end{bmatrix} = \begin{bmatrix} \frac{e^{w_1 x}}{\sum_{k=1}^{2} e^{w_k x}} \\ \frac{e^{w_2 x}}{\sum_{k=1}^{2} e^{w_k x}} \end{bmatrix} \quad (2)$$

where x is the input vector, from the fully connected layer, of the SoftMax layer, w_1 and w_2 are the weight vectors of the neurons and P_{brain} and $P_{artifact}$ are the probabilities of the IC of being of brain or artifactual origin.

In fact, the SoftMax function outputs the predicted probability of an IC of being of Brain or Artifactual origin.

All Weights of the CNN were initialized in a pseudo-random fashion employing a truncated normal distribution (mean=0, standard deviation=0.1, truncation at 2 standard deviations), whereas the biases were initialized to 0 [51].

The CNN was trained in a supervised learning approach [52]. In the supervised learning, CNN parameters, i.e. weights *w*s and biases *b*s, are adjusted relying on an objective function minimization procedure. The objective function measures the error (or distance) between the output scores and the desired scores. We employed the cross-entropy error as objective function. Cross-entropy (CE) is defined as:

$$CE = -\frac{1}{n}\sum_{i=1}^{n} y_i' ln y_i \quad (3)$$

where n is the data set numerosity, y is the output vector of the CNN ($[P_{Brain}\ P_{Artifact}]$ in the study) and y' is the correct classification, that in our case was supposed to be the one derived from the experts' visual inspection ([1 0] for Brain IC or [0 1] for Artifact IC). Cross-entropy metric takes into account the closeness of a prediction to the ground truth and is a more granular way to compute error than Classification Error or Mean Squared Error [53]. As optimization algorithm we employed the Adam Optimizer [29]. Adam Optimizer is a state-of the art learning algorithm that is different from classical stochastic gradient descent since it computes individual adaptive learning rates from estimates of the first and second moments of the gradients, dampening slow learning rate and/or local minima issues. Adam Optimizer parameters were set to: learning rate=10^{-4}, first moment exponential decay rate=$9 \cdot 10^{-1}$, second moment exponential decay rate=$9.99 \cdot 10^{-1}$, constant=10^{-8} [29]. In order to avoid possible over-fitting, a dropout regularization procedure (regularization procedure where randomly selected neurons are ignored during training) [54] was employed (dropout=0.75). The overall CNN learning process is summarized in the following steps:

1. Given input images the forward propagation step was performed; convolution, ReLU and pooling operations along with forward propagation in the fully connected and SoftMax layers were computed.
2. The total error at the output SoftMax layer was calculated based on cross-entropy.
3. Backpropagation was used to calculate the error gradient with respect to all weights and a modified gradient descent algorithm (Adam Optimizer) was employed to update filter weights and parameters values with the target of minimizing the output error. That is, the weights were adjusted proportionally to their contribution to the classification error of the network.

Steps were repeated for all images in the training set multiple times. The accuracy of the CNN was evaluated by counting the number of correct predictions after an ArgMax evaluation of probabilities (selection of the state with the highest probability) of the SoftMax output vector.

In order to reduce learning computational cost and to avoid local minima through noisy updates, the training set is generally divided in small batches that are utilized one at a time in each iteration. The number of iterations required to train the machinery on all the training set available constitutes an epoch.

The optimization procedure was iterated for 200 epochs with a batch size of 20. In order to address the CNN performance, for each training set considered, we performed a 10-fold cross validation procedure, thus separating the data-set in 10 randomly selected equal size subsamples and employing 9 subsamples, 90% of the data, for training and 1 subsample, 10% of the data, for testing, for 10 times [55].

The described CNN architecture, training and validation were implemented in Python through the open-source software library Tensorflow [56]. Further analysis was performed in Matlab.

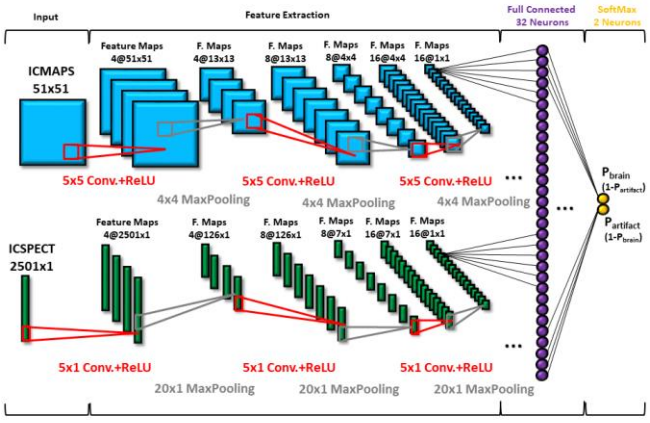

Fig. 2. CNN employed in the study. The CNN was composed of 2- parallel, input layers and 3 convolutional layers, and a fully connected plus a SoftMax layer. The input layers were composed of normalized topographic images of the ICA mixing weights (ICMAPs) and spectrum of ICs (ICSPECTs). The parallel structures consisted of convolutional, non-linear and pooling layers performing feature extraction and dimension reduction. The fully connected layer and the SoftMax layer acted as classifier on the features extracted from the convolutional layers (16 features from ICMAP and ICSPECT, respectively). In fact, a two neurons SoftMax layer constituted the output

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2018.2889512, IEEE Transactions on Biomedical Engineering

layer to provide a probability of the inputs of being either a brain or an artifact IC.

F. Independent Components Classification through Feature based analysis

We compared the feature-less CNN classification with feature-based state-of-the-art automatic algorithms. Specifically, we implemented classification procedures recently reported by Radunz and colleagues [19]. We employed IC features that were shown to provide the best classification performances in their work, namely vectorized images of the IC topography (resampled at 20x20 pixels) after the application of the range filter operator (3x3) concatenated with a down sampled version of the spectra. As classifiers, we employed different algorithms such as Linear Discriminant Analysis (LDA), Linear Support Vector Machine (SVM) and shallow ANN (constituted of an input layer, full connected layer, and SoftMax layer). For the ANN, the last two layers, the objective function and training procedure were identical to that of the CNN. A 10-fold cross validation was employed for estimating performances of all the classifiers.

III. RESULTS

Figure 3 reports average (and related standard error), 10-fold cross-validated, accuracy (evaluated after an ArgMax operation on the SoftMax output) and cross-entropy of the CNN in classifying brain or artifact ICs as a function of training epoch. At epoch 200, we obtained a classification accuracy of 92.4 ± 0.9 % for standalone EEG, of 95.5 ± 0.3 % for standalone MEG and 95.6 ± 0.3 % for combined EEG and MEG dataset.

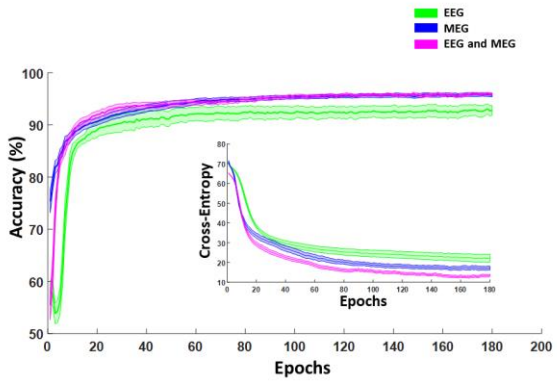

Fig. 3. Average (and related standard error), 10-fold cross-validated, accuracy (evaluated after an ArgMax operation on the SoftMax output) and cross-entropy of the CNN in classifying brain or artifact ICs as a function of training epoch. The CNN was trained on three datasets: standalone EEG (green), standalone MEG (blue), combined EEG and MEG (magenta).

Table 1 reports cross-validated accuracies for the implemented CNN and feature based classifiers.

	Feature-Based			Feature-Less
	LDA	**SVM**	**ANN**	**CNN**
EEG	78.8±0.7%	85.4±0.9%	89.2±1.3%	92.4±0.9%
MEG	92.1±0.6%	91.9±0.7%	94.6±0.3%	95.4±0.3%
EEG+ MEG	86.6±1.1%	89.5±1.3%	92.8±0.2%	95.6±0.3%

Table 1. Average (and related standard error), 10-fold cross-validated, accuracies employing state-of-the-art feature based classifiers (LDA, SVM, shallow ANN) and CNN on standalone EEG, standalone MEG and combined EEG and MEG datasets.

A repeated measures two-way analysis of variance (3X4, 2 within subjects' factors) showed that there was a significant difference on classification accuracy both as a function of the ICs considered (coming from either EEG, MEG or combined EEG and MEG recordings, $F(2,18)=49.3$, $p\sim0$) and the classifier employed (LDA, SVM, ANN or CNN, $F(3,27)=86.3$, $p\sim0$). Moreover, an interaction effect was present ($F(6,54)=8,8$ $p\sim0$).

Post hoc analysis, focused on the classifier effect, revealed an increase performance of CNN classifier with respect to all other feature-based analysis (CNN vs. LDA, $t(9)=15.9$, $p=6.5\cdot10^{-8}$, CNN vs. SVM, $t(9)=11.2$, $p=1.4\cdot10^{-6}$, CNN vs. ANN, $t(9)=5.1$, $p=6\cdot10^{-4}$).

Figure 4 reports evaluation of the CNN behavior and performance as a function of the SoftMax output (before the ArgMax operation). Figure 4a presents the relative frequency of the output probabilities of the CNN for the test set when considering all the cross-validations combined and the different recording modalities. Clear high occurrence of particularly low (P_{brain} or $P_{artifact}$ <0.1) or particularly high (P_{brain} or $P_{artifact}$>0.9) SoftMax output probabilities is present.

Figure 4b reports the average cross-validated CNN accuracy as a function (conditional) of the SoftMax output probabilities. High accuracy (between 97% and 99% for the recording data sets and probabilities examined) of the CNN after the ArgMax evaluation was obtained for particularly low (P_{brain} or $P_{artifact}$ <0.1) or particularly high (P_{brain} or $P_{artifact}$>0.9). SoftMax output probabilities whereas lower accuracy was obtained for SoftMax output probabilities closer to 0.5.

Figure 5 reports examples of ICMAPs and ICSPECTs (CNN inputs) for ICs classified by human experts as either brain activity (Figure 5a) or artifact (Figure 5b). Feature maps extracted from each convolutional layer are reported together with machinery output probability, The CNN correctly classified the ICs (Figure 5a, $P_{brain}=0.9708$; Figure 5b $P_{artifact}=1$).

Figure 6 instead reports examples of ICMAPs and ICSPECTs classified by human experts as either brain activity (a and b), or artifact (c and d). The CNN, after the ArgMax operation on the output of the SoftMax layer, correctly classified the ICs reported on the left side of the Figure (a and c), and erroneously classified the ICs reported on the ride side of the Figure (b and d).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2018.2889512, IEEE Transactions on Biomedical Engineering

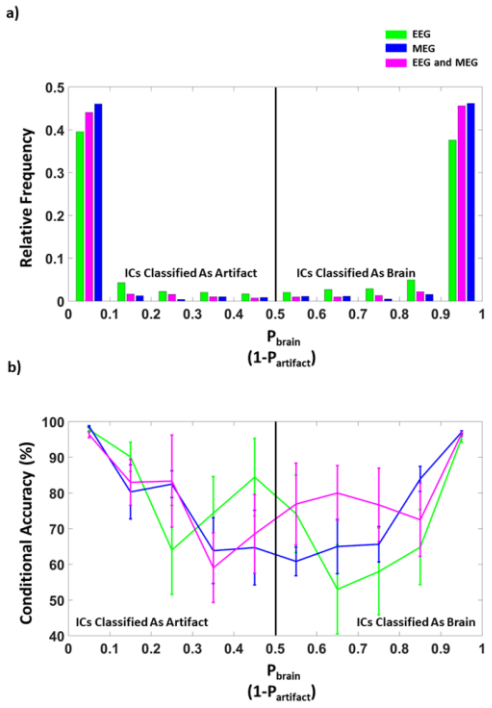

Fig. 4. (a) Relative frequency of the SoftMax output probabilities of the CNN for the test set when considering all the cross-validations combined and the different recording sets. (b) Average cross-validated CNN accuracy (evaluated after an ArgMax operation on the SoftMax output) as a function (conditional) of the SoftMax output probabilities.

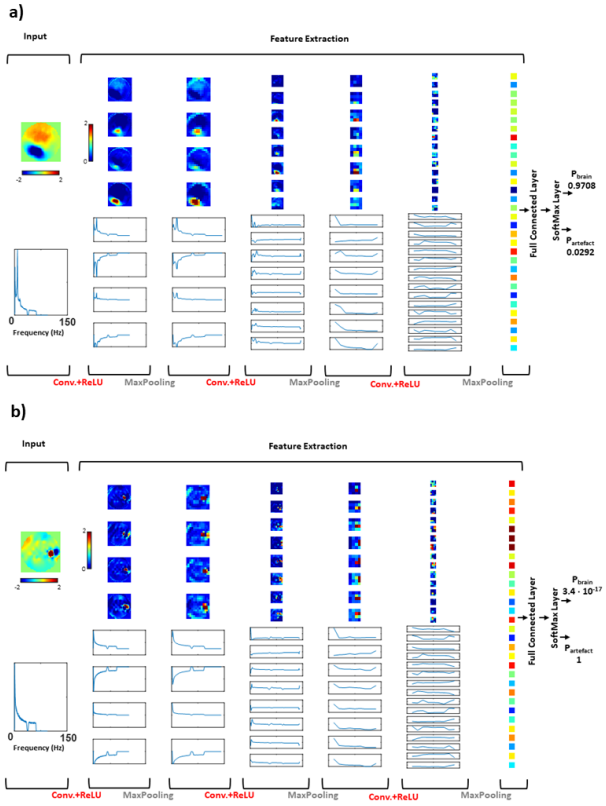

Fig. 5. Examples of ICMAPs and ICSPECTs (CNN inputs) for ICs classified by human experts as either brain activity (a) or artifact (b). Feature maps extracted from each convolutional layer are reported together with machinery output probability. The CNN correctly classified the ICs.

IV. DISCUSSION

Blind source separation techniques are generally considered the gold-standard procedures for offline removal of artifacts in multi-dimensional (multi-channel) brain electrophysiological recordings. In fact, they do not use any training data and do not have a priori assumptions of artifact and brain signal characteristics. ICA is a blind source separation procedure which exploits multi-channel EEG or MEG information, allowing to decompose the recordings in ICs of brain or artifact origin. However, ICA based algorithms for artifact identification and removal must rely on external classification of the ICs. Often, the ICs are classified based on human visual inspection of temporal, spatial and spectral characteristics. These procedures are time-consuming and prone to human error. Automatic procedures for IC classification are thus highly likeable. Development of automatic procedures in the last years generally focused on pre-classification complex feature extraction of the ICs temporal, spatial and spectral characteristics. In this study, we explored the possibility of applying Deep Learning, feature-less, classification algorithms based on Convolutional Neural Network (CNN). CNNs are Neural Networks where neurons are connected to sliding portions of signals and or images that are close in time and/or space [37], [41]. Since neurons of CNNs are combined in groups and connected to sliding portions of the signals or images, the free parameters of each neurons, after learning, generate peculiar filters that automatically identify patterns of interest avoiding a-priori feature selection. CNNs have recently shown their potentialities in image and video recognition as well as in biomedical signal/imaging applications [42]. In order to make the machinery classification as independent as possible of the recording length, we applied the CNN to spectra and topographic maps of the ICs disregarding the temporal signals (note that the Signal to Noise Ratio of the spectrum and topographic map still depends on the signal time length).

The hyperparameters of the CNN were heuristically selected in similarity with previously described CNN architectures [34], [37]. The CNN (through a last layer SoftMax operation) output the probabilities of an IC of being of brain or artifactual origin. Although the heuristically selection of hyperparameters and the absence of complex feature selection, beyond state-of-the-art performance of electrophysiological IC classification was demonstrated by applying the CNN to ICs extracted from recordings of different origin (EEG and MEG), and different datasets (from in house systems and Human Connectome Project) for a total dataset numerosity of few thousand samples. Cross-validated accuracy (estimated on an ArgMax operation of the SoftMax outputs) were 92.4 ± 0.9 % for standalone EEG, of 95.4 ± 0.3 % for standalone MEG and 95.6 ± 0.3 % for combined EEG and MEG dataset (Figure 3). The highest performance, reflecting in the highest accuracy and lower cross-entropy, was obtained for combined EEG and MEG data despite the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2018.2889512, IEEE Transactions on Biomedical Engineering

different recording systems and technologies. These results suggest the generalization capabilities of the procedure being robust to a change in the recording instrumentation and technology.

The evaluation of the SoftMax output probabilities prior to the ArgMax operation, highlighted a high occurrence of particularly low (P_{brain} or $P_{artifact}$ <0.1) or particularly high (P_{brain} or $P_{artifact}$>0.9) probabilities indicating the identification of clear patterns by the CNN (Figure 4a). This aspect was also reflected in extremely high classification accuracies (accuracies between 97% and 99%) for all the data-sets when only extreme SoftMax probabilities were considered.

In fact, it should be highlighted that, although classified erroneously, often errors in classification after the ArgMax procedure were driven by probabilities of being one of the two classes not far from 50% (Figure 5 and 6). This aspect, combined with an expert dependent classification of the IC when the IC characteristics are not clear, suggest that CNNs may be highly suited for feature-less classification of electrophysiological ICs.

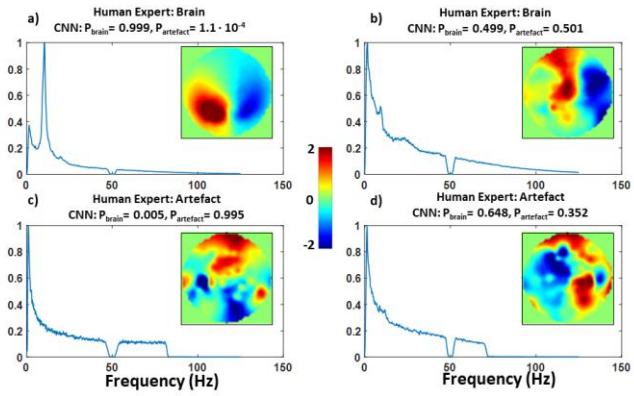

Fig. 6. Examples of CNN input ICMAPs and ICSPECTs classified by human experts as either brain activity (a and b), or artifact (c and d). The CNN correctly classified the ICs reported on the left side of the Figure (a and c), and erroneously classified the ICs reported on the ride side of the Figure (b and d).

It should be further highlighted that CNN, since they self-select the features of interest in a data-driven fashion, generally require large data-set for training. However large data should be available for EEG and MEG recordings, allowing increased performances of feature-less CNN classification beyond the current study thus making CNN the gold-standard classifier for this application. Further research for CNN classification should in fact focus on augmenting the training data-set beyond the current study, possibly combining recordings from multiple systems to provide good generalization. Finally, by increasing the data-set, CNN may be further employed for accurate classification of artifactual ICs of different source origin (e.g., heart, movement, muscle).

V. Conclusions

Beyond state-of-the-art performance of Independent Components (ICs) classification was demonstrated by applying a Convolutional Neural Network (CNN) on ICs extracted from brain electrophysiological multi-channel recordings of different origin (EEG and MEG), and different datasets. These performances were achieved by heuristically selection of hyper parameters and the absence of a priori complex feature extraction, relying on CNN self-selection of the feature of interest. The results suggest that CNNs may be highly suited for feature-less classification of ICs of brain electrophysiological recordings. Considering the high data numerosity of multi-channel EEG and MEG recordings, CNN classification could become the gold-standard procedure for this application.

Acknowledgment

This study was partially funded by grants:

-H2020, ECSEL-04-2015-Smart Health, grant n. 692470, Advancing Smart Optical Imaging and Sensing for Health (ASTONISH).

- BIAL foundation, grant n. 66/2016, Mindfulness meditation shapes synchronization of brain networks for effective perceptual decision making.

References

[1] G. Buzsáki, C. A. Anastassiou, e C. Koch, «The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes», Nat. Rev. Neurosci., vol. 13, n. 6, pagg. 407–420, giu. 2012.

[2] A. F. Jackson e D. J. Bolger, «The neurophysiological bases of EEG and EEG measurement: A review for the rest of us», Psychophysiology, vol. 51, n. 11, pagg. 1061–1071, nov. 2014.

[3] M. K. Islam, A. Rastegarnia, e Z. Yang, «Methods for artifact detection and removal from scalp EEG: A review», Neurophysiol. Clin. Neurophysiol., vol. 46, n. 4, pagg. 287–305, nov. 2016.

[4] A. G. Correa, E. Laciar, H. D. Patiño, e M. E. Valentinuzzi, «Artifact removal from EEG signals using adaptive filters in cascade», J. Phys. Conf. Ser., vol. 90, n. 1, pag. 012081, 2007.

[5] D. J. Doyle, «Some comments on the use of Wiener filtering for the estimation of evoked potentials», Electroencephalogr. Clin. Neurophysiol., vol. 38, n. 5, pagg. 533–534, mag. 1975.

[6] K. T. Sweeney, T. E. Ward, e S. F. McLoone, «Artifact Removal in Physiological Signals #x2014;Practices and Possibilities», IEEE Trans. Inf. Technol. Biomed., vol. 16, n. 3, pagg. 488–500, mag. 2012.

[7] G. Gratton, M. G. Coles, e E. Donchin, «A new method for off-line removal of ocular artifact», Electroencephalogr. Clin. Neurophysiol., vol. 55, n. 4, pagg. 468–484, apr. 1983.

[8] R. J. Croft e R. J. Barry, «Removal of ocular artifact from the EEG: a review», Neurophysiol. Clin. Neurophysiol., vol. 30, n. 1, pagg. 5–19, feb. 2000.

[9] M. Unser e A. Aldroubi, «A review of wavelets in biomedical applications», Proc. IEEE, vol. 84, n. 4, pagg. 626–638, apr. 1996.

[10] N. E. Huang et al., «The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis», Proc. R. Soc. Lond. Math. Phys. Eng. Sci., vol. 454, n. 1971, pagg. 903–995, mar. 1998.

[11] D. Iatsenko, P. V. E. McClintock, e A. Stefanovska, «Nonlinear Mode Decomposition: a new noise-robust, adaptive decomposition method», Phys. Rev. E, vol. 92, n. 3, set. 2015.

[12] C. J. James e C. W. Hesse, «Independent component analysis for biomedical signals», Physiol. Meas., vol. 26, n. 1, pagg. R15-39, feb. 2005.

[13] M. Pal, R. Roy, J. Basu, e M. S. Bepari, «Blind source separation: A review and analysis», in 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013, pagg. 1–5.

[14] A. Hyvärinen e E. Oja, «Independent component analysis: algorithms and applications», Neural Netw. Off. J. Int. Neural Netw. Soc., vol. 13, n. 4–5, pagg. 411–430, giu. 2000.

[15] G. Barbati, C. Porcaro, F. Zappasodi, P. M. Rossini, e F. Tecchio, «Optimization of an independent component analysis approach for artifact identification and removal in magnetoencephalographic signals», Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol., vol. 115, n. 5, pagg. 1220–1232, mag. 2004.

[16] A. Puce e M. S. Hämäläinen, «A Review of Issues Related to Data Acquisition and Analysis in EEG/MEG Studies», Brain Sci., vol. 7, n. 6, mag. 2017.

[17] D. Mantini, R. Franciotti, G. L. Romani, e V. Pizzella, «Improving MEG source localizations: an automated method for complete artifact removal based on independent component analysis», NeuroImage, vol. 40, n. 1, pagg. 160–173, mar. 2008.

[18] I. Winkler, S. Haufe, e M. Tangermann, «Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals», Behav. Brain Funct., vol. 7, pag. 30, ago. 2011.

[19] T. Radüntz, J. Scouten, O. Hochmuth, e B. Meffert, «Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features», J. Neural Eng., vol. 14, n. 4, pag. 046004, 2017.

[20] Y. LeCun, Y. Bengio, e G. Hinton, «Deep learning», Nature, vol. 521, n. 7553, pagg. 436–444, mag. 2015.

[21] H. S. Huang, N. R. Pal, L. W. Ko, e C. T. Lin, «Automatic identification of useful independent components with a view to removing artifacts from eeg signal», in 2009 International Joint Conference on Neural Networks, 2009, pagg. 1267–1271.

[22] M. A. Sovierzoski, L. Schwarz, e F. M. d Azevedo, «Binary Neural Classifier of Raw EEG Data to Separate Spike and Sharp Wave of the Eye Blink Artifact», in 2009 Fifth International Conference on Natural Computation, 2009, vol. 2, pagg. 126–130.

[23] H.-A. T. Nguyen et al., «EOG artifact removal using a wavelet neural network», Neurocomputing, vol. 97, pagg. 374–389, nov. 2012.

[24] J. Schmidhuber, «Deep Learning in Neural Networks: An Overview», Neural Netw., vol. 61, pagg. 85–117, gen. 2015.

[25] R. Hecht-nielsen, «III.3 - Theory of the Backpropagation Neural Network*», in Neural Networks for Perception, H. Wechsler, A c. di Academic Press, 1992, pagg. 65–93.

[26] D. J. C. MacKay, «Bayesian Methods for Backpropagation Networks», in Models of Neural Networks III, Springer, New York, NY, 1996, pagg. 211–254.

[27] J. Snoek et al., «Scalable Bayesian Optimization Using Deep Neural Networks», ArXiv150205700 Stat, feb. 2015.

[28] Y. Bengio, «Gradient-Based Optimization of Hyperparameters», Neural Comput., vol. 12, n. 8, pagg. 1889–1900, ago. 2000.

[29] D. P. Kingma e J. Ba, «Adam: A Method for Stochastic Optimization», ArXiv14126980 Cs, dic. 2014.

[30] G. E. Dahl, T. N. Sainath, e G. E. Hinton, «Improving deep neural networks for LVCSR using rectified linear units and dropout», in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pagg. 8609–8613.

[31] A. Maas, A. Hannun, e A. Ng, «Rectifier Nonlinearities Improve Neural Network Acoustic Models».

[32] R. Pascanu, T. Mikolov, e Y. Bengio, «On the difficulty of training Recurrent Neural Networks», ArXiv12115063 Cs, nov. 2012.

[33] M. Bianchini e F. Scarselli, «On the Complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures», IEEE Trans. Neural Netw. Learn. Syst., vol. 25, n. 8, pagg. 1553–1565, ago. 2014.

[34] A. M. Chiarelli, P. Croce, A. Merla, e F. Zappasodi, «Deep learning for hybrid EEG-fNIRS brain–computer interface: application to motor imagery classification», J. Neural Eng., vol. 15, n. 3, pag. 036028, 2018.

[35] R. Collobert e J. Weston, «A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning», in Proceedings of the 25th International Conference on Machine Learning, New York, NY, USA, 2008, pagg. 160–167.

[36] G. Hinton et al., «Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups», IEEE Signal Process. Mag., vol. 29, n. 6, pagg. 82–97, nov. 2012.

[37] A. Krizhevsky, I. Sutskever, e G. E. Hinton, «ImageNet Classification with Deep Convolutional Neural Networks», in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, e K. Q. Weinberger, A c. di Curran Associates, Inc., 2012, pagg. 1097–1105.

[38] K. Simonyan e A. Zisserman, «Very Deep Convolutional Networks for Large-Scale Image Recognition», ArXiv14091556 Cs, set. 2014.

[39] D. Ciresan, A. Giusti, L. M. Gambardella, e J. Schmidhuber, «Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images», in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, e K. Q. Weinberger, A c. di Curran Associates, Inc., 2012, pagg. 2843–2851.

[40] O. Ronneberger, P. Fischer, e T. Brox, «U-Net: Convolutional Networks for Biomedical Image Segmentation», ArXiv150504597 Cs, mag. 2015.

[41] N. Kalchbrenner, E. Grefenstette, e P. Blunsom, «A Convolutional Neural Network for Modelling Sentences», ArXiv14042188 Cs, apr. 2014.

[42] A. Esteva et al., «Dermatologist-level classification of skin cancer with deep neural networks», Nature, vol. 542, n. 7639, pagg. 115–118, feb. 2017.

[43] F. Lotte et al., «A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update», J. Neural Eng., vol. 15, n. 3, pag. 031005, 2018.

[44] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, e B. J. Lance, «EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces», J. Neural Eng., giu. 2018.

[45] R. T. Schirrmeister et al., «Deep learning with convolutional neural networks for EEG decoding and visualization», Hum. Brain Mapp., vol. 38, n. 11, pagg. 5391–5420.

[46] D. Mantini, M. G. Perrucci, S. Cugini, A. Ferretti, G. L. Romani, e C. Del Gratta, «Complete artifact removal for EEG recorded during continuous fMRI using independent component analysis», NeuroImage, vol. 34, n. 2, pagg. 598–607, gen. 2007.

[47] S. D. Penna et al., «Biomagnetic systems for clinical use», Philos. Mag. B, vol. 80, n. 5, pagg. 937–948, mag. 2000.

[48] L. J. Larson-Prior et al., «Adding dynamics to the Human Connectome Project with MEG», NeuroImage, vol. 80, pagg. 190–201, ott. 2013.

[49] P. Welch, «The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms», IEEE Trans. Audio Electroacoustics, vol. 15, n. 2, pagg. 70–73, giu. 1967.

[50] A. Delorme e S. Makeig, «EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis», J. Neurosci. Methods, vol. 134, n. 1, pagg. 9–21, mar. 2004.

[51] I. Sutskever, J. Martens, G. Dahl, e G. Hinton, «On the importance of initialization and momentum in deep learning», in International Conference on Machine Learning, 2013, pagg. 1139–1147.

[52] T. Hastie, R. Tibshirani, e J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2° ed. New York: Springer-Verlag, 2009.

[53] C. Robert, «Machine Learning, a Probabilistic Perspective», CHANCE, vol. 27, n. 2, pagg. 62–63, apr. 2014.

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, e R. Salakhutdinov, «Dropout: A Simple Way to Prevent Neural Networks from Overfitting», J. Mach. Learn. Res., vol. 15, pagg. 1929–1958, 2014.

[55] R. Kohavi e G. H. John, «Wrappers for feature subset selection», Artif. Intell., vol. 97, n. 1, pagg. 273–324, dic. 1997.

[56] M. Abadi et al., «TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems», ArXiv160304467 Cs, mar. 2016.