International Journal of Neural Systems, Vol. 26, No. 0 (2016) 1650039 (15 pages)

DOI: 10.1142/S0129065716500398

Deep Learning Representation from Electroencephalography of Early-Stage Creutzfeldt-Jakob Disease and Features for Differentiation from Rapidly Progressive Dementia

Francesco Carlo Morabito*,**, Maurizio Campolo*, Nadia Mammone‖, Mario Versaci*, Silvana Franceschetti†, Fabrizio Tagliavini†, Vito Sofia‡, Daniela Fatuzzo‡, Antonio Gambardella§, Angelo Labate§, Laura Mumoli§, Giovanbattista Gaspare Tripodi¶, Sara Gasparini§,¶, Vittoria Cianci¶, Chiara Sueri¶, Edoardo Ferlazzo§,¶ and Umberto Aguglia§,¶

*University Mediterranea of Reggio Calabria, Italy

†Neurologic Institute "Carlo Besta", Milan, Italy

‡Institute of Neurology, University of Catania, Italy

§Magna Græcia University, Catanzaro, Italy

¶Regional Epilepsy Centre, Bianchi-Melacrino-Morelli Hospital
Reggio Calabria, Italy

‖IRCCS Centro Neurolesi Bonino-Pulejo
Via Palermo c/da Casazza, SS. 113, Messina, Italy
**morabito@unirc.it

Accepted 22 April 2016
Published Online 21 July 2016

A novel technique of quantitative EEG for differentiating patients with early-stage Creutzfeldt–Jakob disease (CJD) from other forms of rapidly progressive dementia (RPD) is proposed. The discrimination is based on the extraction of suitable features from the time-frequency representation of the EEG signals through continuous wavelet transform (CWT). An average measure of complexity of the EEG signal obtained by permutation entropy (PE) is also included. The dimensionality of the feature space is reduced through a multilayer processing system based on the recently emerged deep learning (DL) concept. The DL processor includes a stacked auto-encoder, trained by unsupervised learning techniques, and a classifier whose parameters are determined in a supervised way by associating the known category labels to the reduced vector of high-level features generated by the previous processing blocks. The supervised learning step is carried out by using either support vector machines (SVM) or multilayer neural networks (MLP-NN). A subset of EEG from patients suffering from Alzheimer's Disease (AD) and healthy controls (HC) is considered for differentiating CJD patients. When fine-tuning the parameters of the global processing system by a supervised learning procedure, the proposed system is able to achieve an average accuracy of 89%, an average sensitivity of 92%, and an average specificity of 89% in differentiating CJD from RPD. Similar results are obtained for CJD versus AD and CJD versus HC.

Keywords: Alzheimer's disease; CJD; EEG; classification; SVM; deep learning; continuous wavelet transform; subacute encephalopathies; dementia.

1. Introduction

Creutzfeldt–Jakob disease (CJD) is a rapidly progressive, fatal encephalopathy with a median survival of five months.[1,2] It is a rare disease, since it occurs worldwide at a rate of about 1 case per million population per year. It is classified into sporadic, genetic (mutation of the prion protein gene, PRNP) and transmissible forms. CJD is mainly characterized

by dementia and a typical pathological degeneration (spongiform changes) of cortical and subcortical gray matter.[3] Electroencephalography (EEG) is a useful tool in diagnosing CJD, namely in the middle/terminal stages of the disease when characteristic, periodic sharp wave complexes (PSWC) do appear.[2,4] PSWC may disappear at the later stages of the disease[5] when spongiform changes involve whole cerebral cortex.[6]

At early stages of CJD, EEG abnormalities are nonspecific and the whole electro-clinical picture may overlap with different rapidly progressive dementias (RPD). In particular, the challenge for neurologists is distinguishing early-stage CJD from CJD-mimics, such as Alzheimer's Disease (AD), autoimmune and infectious encephalopathies.[7,8]

In this paper, we analyze the most difficult problem of differentiating early-stage CJD from CJD-mimics, through an advanced quantitative EEG processing technique. The proposed system includes a feature extractor based on a time-frequency transform, a two-layer neural network (MLP-NN) used as data compressor of the vector of features extracted in the previous block, and an ensemble of classification neural networks (NN) that map the reduced feature vector in a binary representation of the categories of patterns analyzed.

The EEG signal processor here proposed is able to extract informative features representing sparse oscillatory events from transformed EEG in the time-frequency (wavelet) domain. These features are extracted from all the available recordings (i.e. from all the electrodes) and then compacted through a deep learning (DL) procedure.[9–12]

Recently, DL architectures[9,10] have raised attention in various fields due to their representational power. Motivated by similar research in other clinical applications,[13] we decided to exploit DL for extracting a better feature representation, aimed to enhance the classification (differentiation) accuracy. In practice, we propose a stacked Auto-Encoder,[14,15] which is able to discover a latent representation from the EEG time-frequency low-level features. To the best of our knowledge, this is the first study that considers DL for feature representation in CJD and CJD-mimics. Our experimental results on a database, which, despite the limited size, is unusually large for early-stage CJD studies, proves the effectiveness of the method.

The remainder of the paper is organized as follows. Section 2 describes the recorded experimental EEG database. It also describes the proposed technique for extracting relevant features from EEG and the deep machine learning approach. The results are presented in Sec. 3 and discussed in Sec. 4. Finally, Sec. 5 draws some conclusions.

2. Materials and Methods

2.1. *Study population* (*subjects*)

The data analyzed in this work came from three different Italian centers (Magna Graecia University of Catanzaro and Regional Epilepsy Center, Reggio Calabria; Neurologic Institute "Carlo Besta", Milano; Neurologic Institute, University of Catania).

Among 195 CJD patients consecutively examined in the last 15 years in the three centers, 23 were observed at very early stage and showed no PSWC on EEG recordings.

The EEG recordings of 3 out of 23 CJD-patients were excluded due to continuous artifacts and the EEG recordings from 20 CJD patients were finally considered. A total of 11 out of these 20 CJD patients had a diagnosis of "probable CJD",[2] 5/20 had genetic CJD, all with PRNP 200 K mutation,[16] and 4/20 were diagnosed "definite CJD".[17]

Among 75 patients with RPD consecutively examined in the last 15 years at Magna Graecia University of Catanzaro and Regional Epilepsy Centre, Reggio Calabria, 58 patients were excluded because of insufficient diagnostic data or continuous artifacts on EEG, and EEG from the remaining 17 patients were finally considered. Five out of these 17 patients had "idiopathic limbic encephalitis"[18] (1 with serum anti-Hu antibodies, 1 with serum anti-GABA antibodies, 1 with serum anti-LGI1/VGCK antibodies, 1 with serum anti-GAD antibodies and 1 seronegative limbic encephalitis), 2/17 had "Hashimoto's encephalopathy",[19,20] 4/17 had acute neuropsychiatric complication of systemic autoimmune disease[21,22] (systemic lupus erythematosus: 3 patients; Wegener's granulomatosis: 1 patient), the remaining 6/17 patients had laboratory-confirmed viral encephalitis (3 due to simplex herpes virus type 1, 1 to cytomegalovirus, 1 to enterovirus and 1 to adenovirus).

Among 21 patients fulfilling diagnostic criteria for "probable AD with documented decline"[23] who

Table 1. Demographic features of the evaluated groups of subjects.

Patients	Sample	Gender (M)	Mean age (SD)
CJD	20	8 M	59.4 (14.8)
AD	13	8 M	79.2 (7.0)
RPD	17	9 M	63.1 (8.1)
HC	26	11 M	59.5 (12.4)

Note: CJD = Creutzfeldt-Jakob Disease; AD = Alzheimer Disease; RPD = rapidly progressive dementias; HC = Healthy Controls. The four groups were not different in terms of gender distribution ($p = 0.58$), while mean age was significantly higher in AD group ($p < 0.001$).

were consecutively examined in the last 12 months at Magna Graecia University of Catanzaro and Regional Epilepsy Centre, Reggio Calabria, eights patients were excluded because of continuous artifacts on EEG, and 13 patients were included and their EEGs considered.

Moreover, EEGs recorded from 26 healthy controls (HC) were also evaluated. The final database included EEG recordings from 76 subjects belonging to four different diagnostic categories: 20 with CJD, 17 with RPD, 13 with AD and 26 HC.

The demographic data of these four groups are summarized in Table 1.

The EEGs were acquired after a median of 16 weeks (range 1–28) from clinical onset of the disease for the CJD group, four weeks (range 1–8) from clinical disease-onset for the RPD group and 18 months (range 5–120) of disease evolution in the AD group.

By the time of EEG acquisition, 13 of 20 CJD patients were taking medications (either in monotherapy or in combination); in particular: 6/20 antihypertensive drugs (*amlodipine* 2, *enalapril* 2, *atenolol* 1, *perindopril* 1), 4/20 antidepressant (*paroxetine* 1, *sertraline* 1, *venlafaxine* 2), 3/20 benzodiazepines (2 *alprazolam*, 1 *lorazepam*), 1/20 oral antidiabetic (*metformin*). Among 13 AD patients, 10 were receiving cholinesterase inhibitors (*rivastigmine* 4, *donepezil* 4, *galantamine* 2), 2 benzodiazepines (*lorazepam*), 1 neuroleptic (*risperidone*).

Among 17 patients with RPD, 4/17 antihypertensive drugs (*valsartan* 2,*bisoprolol* 1,*verapamil* 1), 4/20 antidepressant (*sertraline* 3,*mirtazapine* 1). Apart from benzodiazepines, none of the other medications is expected to affect the EEG. Finally, none of HC were assuming any drugs. All patients or caregivers signed informed consent.

2.2. *Methods*

2.2.1. *Electrophysiological recordings*

All EEG were recorded with scalp electrodes placed at 19 standard locations according to the International 10–20 System (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz and Pz) with referential montages using G2 (between electrodes Fz and Cz) as the reference. The EEG recordings were acquired in the morning, in a comfortable resting state, with eyes closed. The technician kept the subject alert in order to prevent the drowsiness. Every recording lasted around 20 min. The EEG was high-pass filtered at 0.5 Hz, low-pass filtered at 70 Hz, plus a 50 Hz notch filter with slope of 12 dB/Oct, and then downsampled to 256 Hz (see also Ref. 24). Manual cleaning of the recordings excluded EEG frames with evident artifacts identified by visual inspection.

2.2.2. *The DL representation scheme*

NN topologies are often organized in layers. Shallow NN are easy to learn through many well-known algorithms. Although Deep NN architectures are recognized to generate better representation of input–output mappings than shallow NN, their use is not prevalent because of the substantial lack of computationally appealing learning methods, and the difficulty of generating the necessary quantity of examples to train them.

In recent years, however, in many applications the trend has changed, because of the growing availability of data (big data), and the proposal of good learning schemes.

Advanced Deep NN now use suitable algorithms, big data and exploit the strong computational power of the GPU to solve complex problems. Among the relevant learning algorithms, we just mention the Restricted Boltzmann Machine, the Deep Convolutional NN, and the Contractive Encodings, which is the preferred scheme for the present study. Well-trained deep NN provide superior performance with respect to standard NN provided the convergence is ensured and the overfitting problem is limited.

In our problem, we need to extract from long EEG recordings a convenient representation that

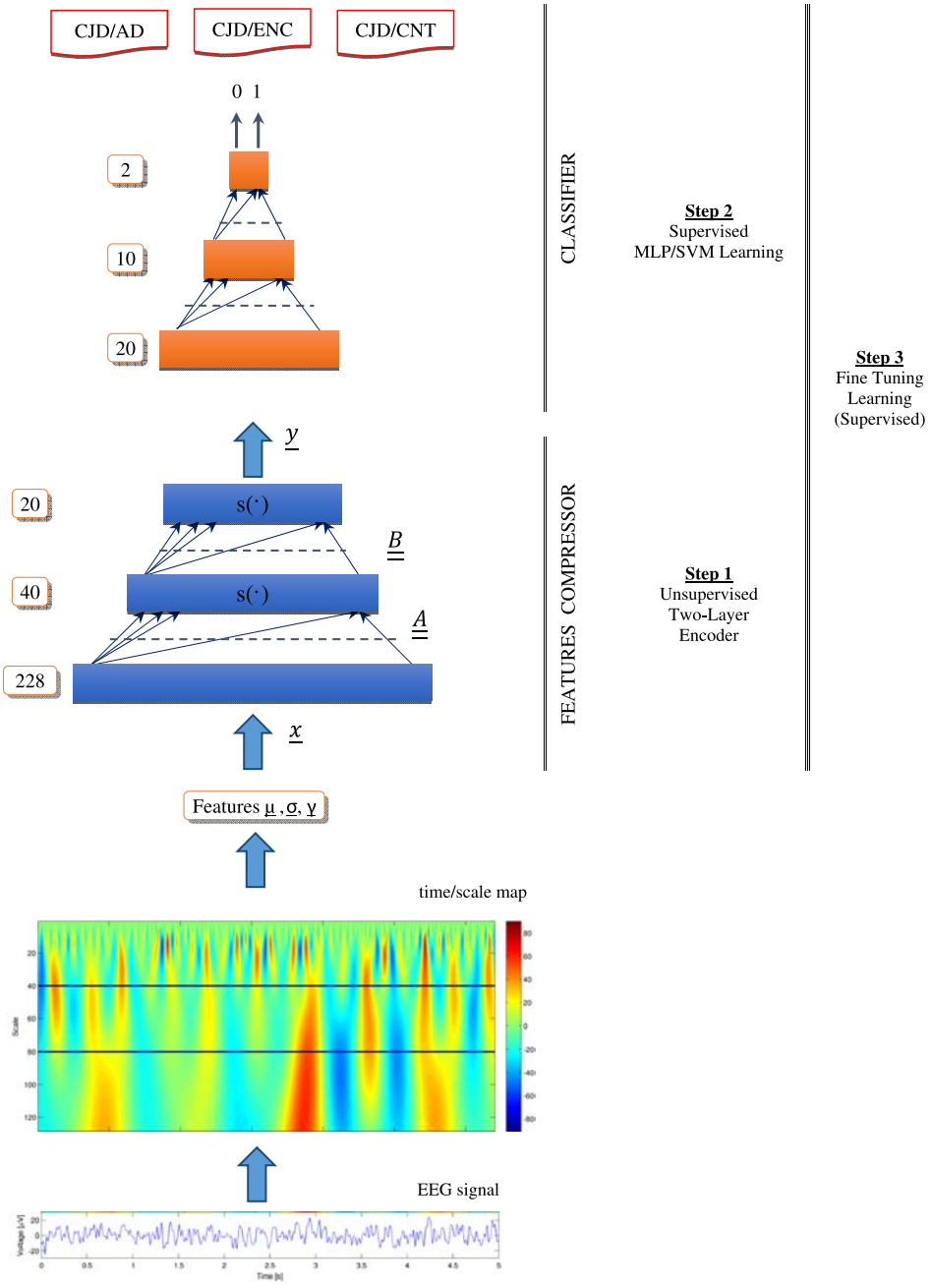

Fig. 1. The EEG signal for each channel is transformed by CWT in the time-frequency domain. The related features are extracted from the maps and form the input vector feeding the auto-encoder (DL processor). Through unsupervised training, the matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are learned. The vector of 20 outputs of the auto-encoder are the "super-features" used to train the classifier, which eventually decide the category of the analyzed subject. Finally, the weights of the whole DL processor can be fine-tuned through a supervised training step.

preserve most of the information embedded by exploiting simultaneously all the channels, but, at the same time, not focusing on possible features related to the underlying noise. The scheme we propose here is the simplest one, based on contractive encodings and a simple variant of regularized back-propagation learning.

Figure 1 depicts the general design scheme of the three-step processing chain here proposed: it includes a channel-by-channel time-frequency analysis of the

EEG recordings by using CWT, a two-layer bottleneck neural network architecture (MLP-NN) based on unsupervised learning, and a classification network based on supervised learning (MLP-NN or support vector machine (SVM)).

If needed, on the basis of the required accuracy, the coefficients (weights) of the whole processor are finally fine-tuned through supervised learning.

The feature compressor encoder drawn in Fig. 1 is obtained as the bottom (encoding) part of the stacked auto-encoder depicted in Fig. 2. The auto-encoder consists of an encoder and a decoder. The encoder is in charge of transforming the input $\underline{x}$ to a hidden representation $\underline{y}$ via the matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ that are learned by the standard back-propagation algorithm. The decoder has the objective to determine some approximate $\underline{\underline{\tilde{A}}}$ and $\underline{\underline{\tilde{B}}}$ matrices such that $\underline{\hat{x}}$ is as similar as possible to $\underline{x}$.

2.2.3. *Time-frequency analysis of EEG*

Most clinically valuable information is not immediately available from the conventional graphical EEG representation: if this information is present, it is possible, however, that it might become apparent in other domains, like time-frequency plane.[25–32] To extract features from the available data in the time-frequency domain, we decided to use the continuous wavelet transform (CWT). The choice of CWT with respect to discrete wavelet transform (DWT), which is relatively low complexity and more efficient, at least with a large number of scales, is here motivated as follows: (1) the CWT is self-similar at all scales, and thus any scale change does not alter the resolution, as a difference with the discrete-time analysis that involves subsampling at increased scales; (2) CWT yields a redundant representation of the EEG traces most useful to reduce noise through the procedure described in Sec. 2.2.5.

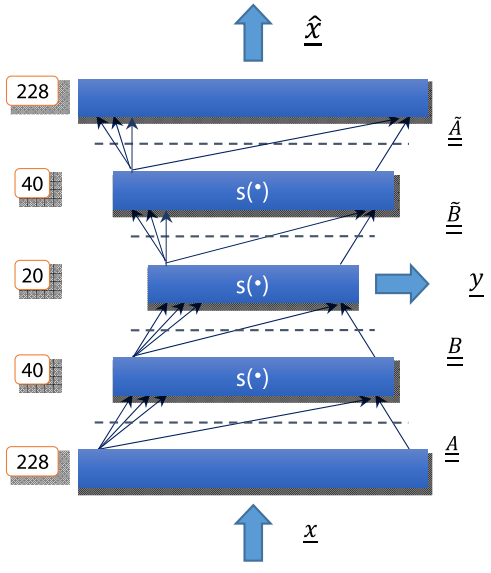

Fig. 2. Stacked Auto-Encoder: this multilayer scheme converts the feature input vector ($\underline{x}$) to a compressed vector ($\underline{y}$) that forms the input of the classification procedure. The objective of the learning process here is to reconstruct at the final output a good approximation ($\underline{\hat{x}}$) of the input ($\underline{x}$). The reconstructed vector is obtained through the approximation matrices $\underline{\underline{\tilde{A}}}$ and $\underline{\underline{\tilde{B}}}$. $s(\bullet)$ indicates a sigmoidal nonlinearity.

A wavelet is a "window" function of finite predefined length with zero mean value. The "mother" wavelet can be intended as a prototype for generating other functions through some scaling (dilation/compression) and time-translation operation. The scale factor is here denoted by α and the time delay by τ. The CWT is the integral of the EEG signal multiplied by the scaled and shifted versions of the selected mother wavelet over the epoch duration[26,28]:

$$\mathrm{CWT}(\alpha,\tau) = \frac{1}{\sqrt{\alpha}} \int_{-T}^{T} \mathrm{eeg}(t)\Psi^* \left(\frac{t-\tau}{\alpha}\right) dt. \quad (1)$$

The transformed signal in (1) is a function of two variables (α, τ). $\Psi(\bullet)$ indicates the mother wavelet. $2T$ is the length of the considered epoch. CWT coefficients are obtained by an inner product of the EEG signal with the selected basis functions. A relatively high value of the coefficient is given in the product with the wavelet if there exists a spectral component of the signal corresponding to the value of α at a location τ. The scale α is inversely proportional to the frequency (a large scale implies a small frequency and vice versa). However, this correspondence is intended in a quite broad sense, and the frequency is indeed normally referred to as "pseudo-frequency". In particular, by indicating with f_c the center frequency of the selected mother wavelet (which is a rough characterization of the leading dominant frequency of the wavelet), and the underlying sampling frequency with f_s, the pseudo-frequency f_α corresponding to the scale α is given by: $f_\alpha = f_c f_s/\alpha$.

In our study, we use the so-called "Mexican hat" function, as mother wavelet. This is useful in feature

extraction, since it is related to the second derivative of a Gaussian function, and thus it is sensitive to second derivatives of the inputs (2). In addition, we selected this function as prototype wavelet because of its good adherence to the brain signals profile, and, in particular, to the form of PSWC.[27,45]

$$\Psi(t) = \pi^{\frac{-1}{4}}(1-t^2)e^{\frac{-t^2}{2}}. \quad (2)$$

2.2.4. *DL of high-level EEG features*

DL refers to computational architectures composed of multiple processing layers that are able to learn representations of data with multiple levels of abstraction. In order to explore the power of the DL methodology independently from sophisticated learning algorithms, we decide to exploit here the simplest possible technique among the various proposed in the literature.[9,33] In particular, we use the well-known "bottleneck multilayer neural network (MLP-NN)" as a stacked auto-encoder (see Fig. 2).

Since our objective is to generate some higher order features, we stack two levels of nonlinear (sigmoidal) nodes and consider the outputs of the deepest hidden layer ($\underline{y}$) as feature input vector for the subsequent classification step.

The output of the deepest hidden layer of the MLP-NN has the form:

$$\underline{y} = \underline{s}(\underline{\underline{B}}\,\underline{s}(\underline{\underline{A}}\,\underline{x})), \quad (3)$$

where $\underline{s}$ is a component-wise sigmoid nonlinearity, e.g.:

$$s = \frac{1}{1+e^{-z}}. \quad (4)$$

$\underline{\underline{A}}$ and $\underline{\underline{B}}$ are the learned matrices of the first and second layer of the MLP-NN, randomly initialized and then determined through standard back-propagation. The cost function to be minimized through the learning process is here given by:

$$\mathrm{C} = \|\underline{\hat{x}} - \underline{x}\| + \lambda\|\underline{\underline{w}}\|^2 \quad (5)$$

where $\lambda > 0$ is the regularization (Tikhonov) coefficient, $\underline{\underline{w}}$ represents the matrix of weights, $\|\underline{\underline{w}}\|^2$ is a smoothness penalty term that tends to limit the growth of the matrix entries, w_{ij}. For $\lambda = 0$, the standard root mean square (RMS) cost function is recovered. We select here the value of λ through a semi-heuristic technique based on cross-validation: in other words, the "optimal" λ is obtained as giving the smallest "leave-one-out" cross-validation error. It has been shown that the training of a stacked auto-encoder with the penalty term produces features that later on helps fine-tuning.[14] In the formula, $\underline{x}$ acts both as input of the DL compressor module and as target vector for the auto-encoder, while $\underline{\hat{x}}$ is the approximate (learned) output, which should reproduce $\underline{x}$ as similarly as possible at the end of the training step, expressed by:

$$\underline{\hat{x}} = \underline{\underline{\tilde{A}}}^{-1}\underline{s}(\underline{\underline{\tilde{B}}}^{-1}\underline{y}). \quad (6)$$

It is worth noting that the derivation of an optimal value for λ is not so critical here, since the goal of the MLP-NN auto-encoder is just to build a compact representation of the input and not to make a high-quality reconstruction of $\underline{x}$.

The unsupervised training of the auto-encoder has the additional advantage of simplifying the training of the full DL scheme, since it moves the weights of the representation to a region more related to the actual inputs: in other words, the "gradient dilution" effect, often met during training of deep many-layered NN, is largely reduced.[33] The final step of the proposed DL representation scheme is indeed a "fine-tuning" of the weights, which is carried out through back-propagation (see right part of Fig. 2).

2.2.5. *Summary of the proposed method for EEG classification*

The EEG database (for all the considered categories of subjects) is processed according to the following steps:

(i) Artifact rejection through clinical (visual) inspection: the segments of signal affected by evident artefactual components are discarded;
(ii) Decomposition of each residual recording in nonoverlapping epochs of 5 s duration (moving window technique);
(iii) Per epoch time-frequency analysis of each signal through the CWT (with the Mexican-Hat wavelet as mother wavelet);
(iv) Subdivision of the CWT map in three parts grossly related to the brain rhythms;
(v) Computation of the mean value (μ), standard deviation (σ) and skewness (γ), of the wavelet coefficients for the three above considered subbands (SB) as well as for the whole CWT map;
(vi) Plot of the above parameters for all of the epochs included in each signal and visual

inspection for detection of outliers (evident outliers have been considered to be generated by segments of artifacts that have not been detected in step 1 and thus excluded from the analysis);

(vii) Computation of the averaged value of the statistical quantities on the whole recording length;

(viii) A vector of features is thus generated that includes the above described averaged values: they are three per electrode (μ , σ and γ for the three SB) plus the average values of μ, σ and γ taken from the whole CWT map; the resulting vector has thus a length of $12 \times 19 = 228$ elements.

The above-described procedure represents a feature-engineering step aiming to extract discriminative information from the available data that fails to emerge from the mere visual inspection of the EEG time-traces. Although the averaging step reduces the impact of local (time) distribution of frequencies, so partially limiting the usefulness of the same time-frequency analysis, the availability of the three estimated statistical quantities gives synthetic information on the underlying probabilistic distribution. In particular, the CWT map averaged over all the epochs and over all the subjects' exhibits at a visual inspection the presence of different periodic component in the different categories here analyzed. The proposed procedure, grounded on a machine learning approach, allows us to exploit simultaneously the entire electrode traces of the EEG signal, thus overcoming the limitations of the standard single-channel processing.

2.2.6. *SVM for classification*

SVMs are supervised learning processors used for both classification and regression problems. The aim of an SVM is to generate a hyperplane, which can separate two classes of data in an optimal way. An SVM builds a representation of the training data as points in a suitable space, mapped so that the examples of the classes are divided by a gap as wide as possible. The SVM algorithm was originally introduced by Vapnik *et al.*[34,35] Boser *et al.*[36] proposed a method to create nonlinear classifiers by applying the so-called "kernel trick" to maximum-margin hyperplanes. The kernel functions can be selected as linear, polynomial or radial basis functions. Cortes and Vapnik proposed the current standard (soft margin).[37] In SVM, intuitively, a good separation is achieved by the hyperplane that maximizes the distance to the nearest training-data point of any class (the so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.[38,39] In this paper, the SVM approach has been used both as a classifier "stand-alone" and as the final block of a DL scheme.

2.2.7. *Complexity analysis of the EEG though permutation entropy (PE)*

The behavior of the EEG signals can be investigated in time domain through a channel-by-channel complexity analysis. We carried out this analysis by using the PE concept.[40] The study highlighted that only some channels show statistical discriminative power with respect to the analyzed categories. However, we cannot exclude that, because of inter-individual diversity, in some cases different electrodes could also have discriminative power. This is also a possible consequence of the limited size of the database here processed. Accordingly, we considered all of the channels notwithstanding their statistical significance within the available database. On the other hand, one of the strengths of machine learning approaches is indeed the possibility of easily managing large input vectors. However, the PE can be useful as an additional marker for improving the discrimination between CJD and AD or HC. Since through DL we introduce an extensive data compression procedure, the PE marker for each channel can be easily added to the input vector of the bottleneck MLP-NN.

3. Results

All the experiments but the DL architecture and learning have been implemented in MatLab® R2015a. The stacked auto-encoder has been designed and implemented with NeuralWorks Professional II Plus, NeuralWare®, an interactive environment that allows a friendly management of NN topologies and parameters. We used a moving window technique to subdivide each long term EEG into epochs whose duration has been fixed to 5 s for all the computations here carried out.

Figure 3 shows the time-frequency maps (i.e. the wavelet coefficients) related to a 5 s epoch for a sample subject and for a representative channel belonging to the four different categories of subjects here analyzed. We subdivide the associated wavelet map into three nonoverlapping bands, grossly indicated in the figure. These time-frequency plots unveiled that the signal power is differently distributed in the considered SB. Indeed, the number of relevant wavelet coefficients is different, i.e. the "sparsity" of the EEG is different for the four categories, and the characteristics "bumps" of the representation are differently distributed and located in the plots.

Some channels have shown statistically significant differences among the categories. However, we cannot consider as features all of the coefficients for each channel and each epoch, also taking into account the high redundancy of the coefficients' matrix. In order to gain insights on the full time recording without retaining all the coefficients, we decide to estimate the mean values of the coefficients in the three SB and then to take the value averaged on all the epochs.

In order to discriminate among the four categories of interest, we use some average quantities extracted from the EEGs. Since the EEG of AD patients is known to show a slowing effect with respect to healthy age-matched controls (HC), a feature globally controlling the time-complexity of the signal can be useful to help classification. Based on previous works, the feature considered for measuring the complexity of the EEG is the PE, i.e. a quantity, computed in the time domain. However, this feature is largely insufficient to solve the problem of differentiating CJD from RPD: thus, we propose here the use of features extracted by a time-frequency analysis. The input vector of the DL scheme is formed by taking 228 features extracted from the time-frequency maps.

They are the average value, the standard deviation and the skewness of the wavelet coefficients computed in each SB. We also added the same statistical quantities computed on the overall map. It is worth mentioning that the statistical features are not sufficiently sensitive for classifying the subjects when used either in isolation or in small groups. This is because no feature is able to capture the complete frequency information and can only extract some partial information (e.g. low, medium or high scale).

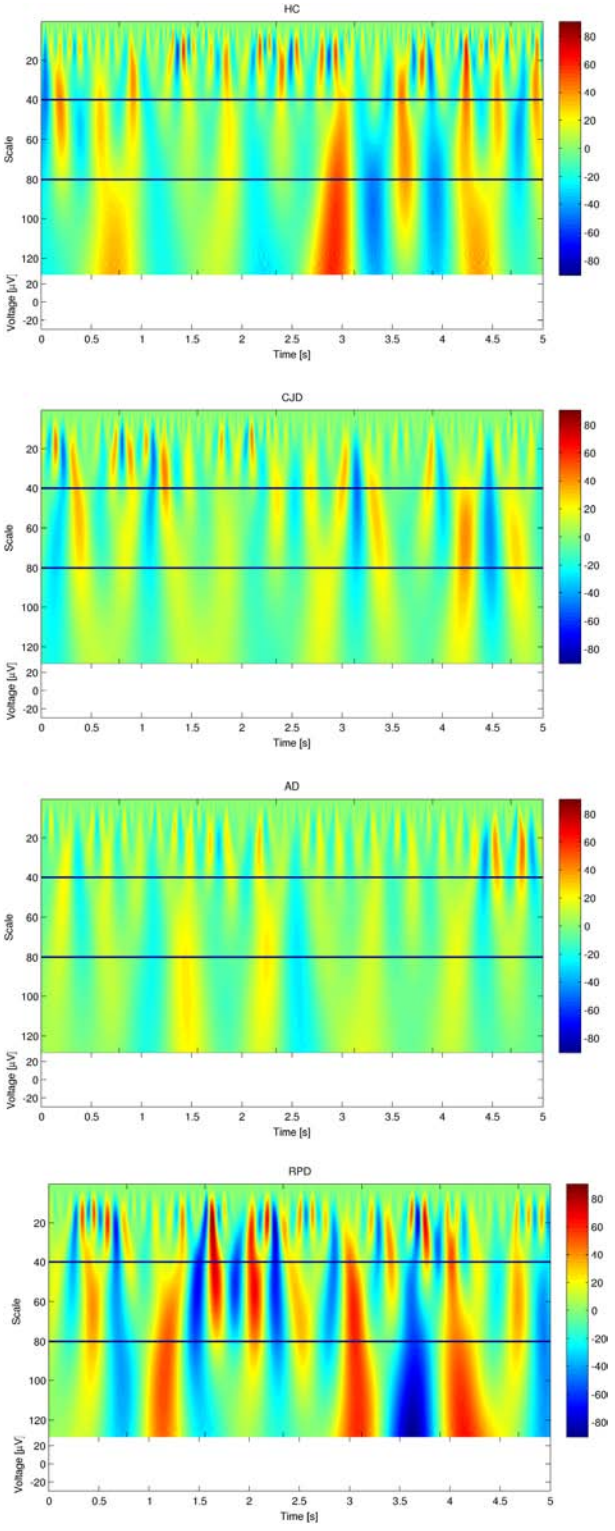

Fig. 3. Time-frequency maps for sample subjects, one for each study group here considered (HC, AD, CJD, RPD). Each figure refers to a 5 s epoch of the recorded EEG signal. The maps show that the distribution of the "bumps" are quite different for the considered groups.

Table 2. Statistical significance of the mean wavelet coefficients in the three defined SB for the 19 channels and the patients belonging to the categories CJD and RPD.

	Fp1	Fp2	F7	F3	Fz	F4	F8	T3	C3	Cz	C4	T4	T5	P3	Pz	P4	T6	O1	O2
SB1	0.48	0.19	0.94	0.24	0.94	0.95	0.38	0.03	0.15	0.25	0.02	0.07	0.01	0.01	0.01	0.01	0.01	0.01	0.01
SB2	0.86	0.80	0.07	0.18	0.81	0.44	0.11	0.01	0.24	0.21	0.22	0.01	0.01	0.02	0.02	0.02	0.01	0.01	0.01
SB3	0.22	0.11	0.11	0.38	0.52	0.18	0.07	0.02	0.30	0.26	0.04	0.01	0.01	0.04	0.01	0.01	0.01	0.01	0.01

The CWT approach helps to extract separate frequency scale components by exploiting fast and slow bell-shaped components through the Mexican-hat wavelet. The resulting number of features is appropriately reduced through the DL approach; this also enhances the training performance of the full model as it controls overfitting.

3.1. *Statistical relevance of PE and time-frequency features for categories discrimination*

A preliminary statistical analysis is carried out to check if a single feature measuring the average complexity of the EEG recording (i.e. the mean PE) can point out significant differences among the groups. This analysis is carried out on all of the electrodes' time-traces (i.e. over 19 channels) by using ANOVA. PE is estimated channel-by-channel, and the average PE over all channels as well as the RMS difference of PE values over the channels belonging to the right and left hemispheres has been calculated in all subjects. The following variables are analyzed: age at inclusion, sex, total average PE and RMS interhemispheric PE differences. Chi-square and one-way ANOVA have been performed to assess differences among groups, followed by a post-hoc test (Tukey–Kramer) for multiple comparisons. Outliers were detected with Tukey method. Significance level was set to 0.05.

The mean PE per channel shows that the differences among groups are statistically significant ($p < 0.05$) only for a limited subset of channels in the different combinations of the groups; the PE averaged on all of the electrodes shows significant differences for CJD versus HC ($p < 0.01$) and for CJD versus AD ($p < 0.05$), not for CJD versus RPD. The CJD and RPD groups did not differ significantly for age and sex. RMS interhemispheric PE difference was higher in CJD patients as compared to AD ($p < 0.03$). We also consider the statistical relevance of the time-frequency features here proposed (i.e. mean, standard deviation and skewness of the wavelet coefficients in the three SB). In Table 2, we report the p-value for the different electrodes and the three considered SB for the two groups CJD and RPD. Table 2 shows that the groups can be differentiated in all the three SB at the electrodes T3, P3, Pz, P4, T5, T6, O1 and O2. At C4 and T4, there are significant differences just in two SB. Age and sex did not show any statistically significant effect. The significance level for all tests has been fixed to $p = 0.05$.

3.2. *Classification with SVM*

We trained a two-class SVM for the following categories: CJD versus RPD; CJD versus HC; CJD versus AD. Thus, each classifier aims to discriminate CJD from one of the three other categories. SVM, similarly to NN models are prone to overfitting, particularly in the case of a database of limited size like the one we use here. This invariably implies a different performance on the training and on the validation databases. For the three sub-parts of the database, the classification performance in terms of percentage of correctly and incorrectly classified subjects (i.e. sensitivity and specificity) as well as of the global accuracy (i.e. the number of correctly classified patterns over the total number of patterns) are measured.

After a number of experiments, for the final SVM model we choose a Gaussian kernel, with two tuning parameters, i.e. the regularization parameter, γ that rules the trade-off between the training error minimization and the smoothness of the solution, and σ^2, which is the squared bandwidth. A Bayesian framework was used to optimize the tuning parameters. The optimal regularization parameter and the kernel parameters were estimated through a procedure of optimization of the cost function on the second and the third level of inference, respectively.[37]

For each classifier, 75% of the data have been used for training and the remaining 25% for testing. Both datasets were uniformly distributed, each containing a balanced number of input patterns from both classes. Due to the small size of the dataset, 100 training sets (with the corresponding 100 testing sets) were setup following the above-mentioned requirements. The SVM has been trained and tested 100 times for each classification problem. The procedure has been carried out on both a standard SVM with 228 inputs and a SVM preceded by a dimensionality reduction step based on standard principal component analysis (PCA). In the second case, 20 PCs have been retained.

In summary, in the test (validation) step, without PCA, the average accuracy is higher than 83%, the average sensitivity is higher than 85% and the average specificity is higher than 82%. With the PCA preprocessing the figures are respectively higher than 88%, 89% and 87% (CJD versus RPD database). For the CJD versus HC database, the performance are similar for the PCA+SVM processor, whereas the accuracy and the sensitivity are not so good without PCA; in the case of CJD versus AD database, we get a good performance just in terms of sensitivity. However, as we clarified in the previous paragraph, the performance of the last two classifiers can be highly improved by including further inputs derived from the mean complexity (averaged PE/RMS interhemispheric PE). In this case, the classification performance improves to over 90%.

A SVM scheme is also used for training on a database with 20 inputs and 2 binary outputs, where the 20 input variables are the outputs of the DL processor described in the next section.

3.3. *DL of high-level features*

The success of machine learning methods is dependent on the appropriateness of the features on which they are applied. Taking advantage of the time-frequency analysis, we have seen that, with rather limited design efforts, we are able to build good classifiers by using SVM, since the discriminative information appears to be well represented in the selected vector of features. However, in a clinical setting, we could expect that novel data sources come unlabeled, and thus an unsupervised representation could be appealing.

In view of this, we decide to build a DL-based NN processor. Accordingly, we have considered a two-layer stacked auto-encoder trained with a standard backpropagation algorithm. A regularizing term is added to the cost function in order to limit the overfitting of training data. This is carried out by introducing a L_2 penalty term on the weight matrices, as shown in Sec. 2.2.4. It penalizes the dependence of the output on few input variables, so ensuring stability of the output to small changes in the input.

A design step is the selection of the multilayer feedforward neural network (MLP-NN) topology: we try here to trade-off the computational complexity of the training of a "deep" structure and the inability of a single layer to find higher level representations of the input feature vector.[41] The optimal choice appears a double layer encoder followed by a double layer decoder. The optimal number of the extracted high-level features could be selected through a pruning procedure, but it seems inessential to find a minimal representation, since we need then to feed a classifier NN whose training is easier if the input vector has some redundancy. Thus, we decide to consider 20 features for the deepest level. The size of the intermediate layer is obtained by trial-and-error.

The resulting network is a 228-40-20-40-228 MLP-NN; finally, we use the output of the deepest hidden layer (20 nodes) as input vector of the following classifiers. We use an annealing procedure for the learning rates that basically starts from high values (0.1) to reach very small values (0.01) at the end of learning. We introduce the DL approach to extract from the data some high-level features that can be more useful to stabilize the classification procedure. The 20 latent variables extracted from mid-layer can be considered as "super-features", i.e. higher-level features well representing the problem at hand. As an example, in a problem of image recognition, the basic feature is the pixel, the next level feature is a line (edge), and a third-level feature is a figure bounded by edges. The whole database including the four categories of subjects has been used to train the compressor by using a k-fold cross-validation technique (we selected $k = 3$ because of the limited size of the database).

We achieve a RMS reconstruction error of less than 5% on both the training and test datasets. The

test performance are also assessed by a standard leave-one-out technique with random selection of the test pattern repeated 30 times. It is worth noting that, in this case, the objective of the training is to extract a set of features that is able to represent the input patterns in a condensed form not necessarily achieving a perfect reconstruction of the input. In addition, by limiting the reconstruction performance of the auto-encoder scheme in the training phase (early stopping), we reduce the risk of overfitting. The 20 outputs of the deepest hidden layer yield a distributed representation of the patterns belonging to the four categories; however, it is difficult to interpret the contribution of each individual neuron to the problem solution. To evaluate the discriminative effect of each hidden neuron, we estimate the average value and the variance of the hidden neurons output distribution and then we are able to select couple of neurons particularly effective in separating the different classes. It is worth mentioning that the very nature of the learning step with regularizer generates hidden neurons that specialize to discriminate between couples of different categories. In our example, neurons 2, 8, 13, 17, 18, 19 and 20 are able to statistically discriminate CJD from RPS ($p < 0.01$); neurons 8, 12 and 20 HC from AD ($p < 0.01$); neurons 7 and 12 CJD from AD ($p < 0.01$), and so on.

Figure 4 shows the scatter diagram referring to a couple of "super-features" selected within the 20 outputs of the two-layer encoder, for the case (CJD versus RPD). In this 2D representation, it is highlighted the possibility of defining a good separation between the two represented categories (CJD versus RPD). It can be noted that a good classification performance can be achieved not with standing we analyze just a two-variable case.

The final classification stage (both based on MLP or SVM) can exploit all the 20 "super-features" to enhance the classification performance.

Then, we built three different datasets for the three different classifiers, each of size 20-10-2 (a double-binary output has been selected). Both the SVM and the MLP approaches have been used to make the classification. The performance is quite similar, but the MLP approach is superior in the test phase.

Table 3 shows the performance of the classifiers. It is yet to be noted that the performance of the

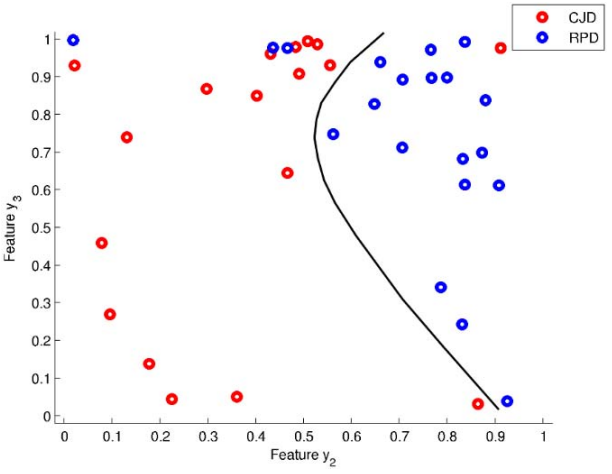

Fig. 4. Distribution of the projection of the vector of 20 "super-features" on the plane formed by 2 of the 20 components (CJD versus RPD). The superimposed curve shows that a good discrimination between the two categories is possible through a second-degree polynomial by taking into account just two components (here, an accuracy of 89% is achieved).

(CJD versus AD) and (CJD versus HC) classifier can be improved by adding an input variable measuring the complexity of the signal, according to the results illustrated in Sec. 3.1.

It is worth noting that the performance of the SVM classifier on the (CJD versus RPD) case are worst with respect to the previously reported case of SVM on 228 inputs both with and without PCA preprocessing. We interpret this circumstance as follows: the 20 "super-features" extracted from the auto-encoder contain discriminative information on the four categories, not just on two of them. Accordingly, although they are able to represent more compactly the general four-class problem, they are less efficient in the two-class problem.

The evident advantage is that this architecture is more useful for the processing of novel previously unseen patterns. Table 4 reports the performance of the DL processor supported by a final fine-tuning supervised learning step.

Although the performance is clearly improved, the tuning of the network parameters have been carried out on separate schemes for the three subcases, namely CJD versus RPD; CJD versus HC; CJD versus AD. This is clearly a limitation for the check of future unlabeled subjects (see next section).

Table 3. Performance on the validation step of the 20-10-2 classifier for the three problems at hand.

	SVM (20-10-2 classifier, test results)			MLP (20-10-2 classifier, test results)		
	Avg Acc (%)	Avg Sens (%)	Avg Spec (%)	Avg Acc (%)	Avg Sens (%)	Avg Spec (%)
CJD versus RPD	77	76	81	83	81	85
CJD versus AD	83	93	73	83	92	75
CJD versus HC	76	74	72	81	83	80

Table 4. Performance on the validation step of the 20-10-2 classifier for the three problems at hand, with supervised fine tuning step.

	MLP (20-10-2 classifier, test results)		
	Avg Acc (%)	Avg Sens (%)	Avg Spec (%)
CJD versus RPD	89	92	89
CJD versus AD	88	94	85
CJD versus HC	87	86	84

3.4. *On-line classification of novel subjects*

Once an EEG comes, through the moving window technique, the epoch-by-epoch (5 s) time-frequency map is computed and the average features are extracted. The input vector is passed through a double layer encoder 228-40-20 with nonlinear (sigmoidal) transfer function. The distance of the final vector of 20 "super-features" from the centroids estimated for the four groups are calculated. Then, we consider the two "nearest" classes and, if one of the two is CJD, we select the related previously trained classifier in order to take the decision about the class.

4. Discussion

The differentiation of CJD from other forms of dementia is a relevant clinical problem, particularly in the early stage of the disease.[46–50]

In this study, we explored the ability of machine learning-based processors to discriminate between the EEG of subjects belonging to four different groups (CJD, RPD, AD, HC). In the recent literature, a very limited number of papers can be found regarding the application of machine learning approaches to CJD. In particular, a SVM classification analysis has been carried out by taking into account samples of cerebrospinal fluid (CSF) from suspected CJD patients.[42] Quantitative EEG has been considered in two recent works, mainly focused on independent component analysis (ICA) for estimating the source of PSWC in the brain of three CJD patients.[43,44]

The performance of any machine learning is strongly dependent on the way the available data are represented (i.e. features). Hierarchical network models, such as the DL architectures, are considered as emerging methods within the community. DL are widely used in unsupervised learning to discover multiple levels of data representation, where the highest levels represent concepts that are more abstract. They transform observed variables (input of the procedure) in latent variables, which express both the relevant aspects of the given data and the underlying feature-generation process. Typically, through unsupervised learning, observable data are clustered according to some unknown nontrivial statistical characteristics. In the case of limited labeled data, as in this work, because of the unavailability of large database referring to CJD, a preliminary unsupervised step is needed to reduce the risk of insufficient generalization performance.

In this work, we found that a time-frequency wavelet analysis of the EEG data yields interesting features to differentiate CJD from RPD. To our best knowledge, this is the first time that wavelet analysis has been used in the study of resting-state CJD data. However, CWT gives redundant and messy information that ultimately reduces the efficiency of machine learning schemes. The set of low-level features produced by CWT forms the input of a DL procedure.

We observed that higher level features generated by a DL procedure enhance the statistical differences in EEG activity between CJD patients and the other considered categories. Since CJD is a rare disease and, in addition, we just considered the EEG of patients in the early stage of the disease (i.e. when the characteristics PSW complexes typical of CJD are not yet present), the database at our disposal is

of limited size. This clearly reduces the reliability of machine learning approaches since their performance in the generalization phase can be relatively unsatisfactory. However, the DL approach allowed us to design a processing system that reduces the impact of the limited size of the available data. Indeed, we considered a stacked double-layer network ("bottleneck" encoder) that was able to reduce the number of features of an order of magnitude. This reduction is beneficial since it reduces the number of free parameters of the classifier that become comparable to the database size. In addition, since this two-layer network is trained through unsupervised learning algorithms, i.e. on "unlabeled" examples, the design of the DL network gave us the opportunity to generate a system which is able to roughly estimate the probability of a previously unseen EEG to belong to a specific group. This quantity is obtained through a clustering procedure and a measure of distance from the group centroids.

With remarkable values up to 85% for the accuracy, sensitivity and specificity in the validation test when discriminating CJD and RPD, this study proved the interest of machine learning approaches for supporting the clinical diagnosis in early stage of the disease.

One of the limitations of the presented DL approach is that the "super-features" identified as output of the two-layer auto-encoder are difficult to be used in clinical settings, being a nonlinear combination of averaged time-frequency quantities. This is also related to the use of a regularizer during training that tends to distribute over the most part of the network's nodes the learned representation. The clinical significance of the individual inputs to the classifier is thus very scarce.

5. Conclusions

We carried out a retrospective analysis of a relatively high number of cases of patients affected by CJD, in the early stage of the disease, by differentiating them from RPD just using quantitative EEG. The study exploits the power of DL approach combined with a time-frequency analysis of the EEG recordings. The statistical significance of a set of features extracted both in time (complexity) and in time-frequency (wavelet coefficients) domain has been assessed: we found that just a limited number of channels could be of help to solve the problem. Because of the limited number of cases available, we decided not to exclude other channels from the analysis. A DL processor was in charge of reducing the dimensionality of the high number of features. The results achieved are quite relevant by taking into account the difficulty of a clinical diagnosis just based on a visual inspection of the EEG. These findings suggest that although it is difficult to discriminate CJD in its early stage from RPD by just inspecting the EEG, the time-frequency maps contain consistent information that can have relevant diagnostic value. The machine learning approach can be of help to overcome the practical limitations of the time-frequency approach.

Acknowledgements

Nadia Mammone's work was funded by the Italian Ministry of Health, project code: GR-2011-02351397.

References

1. M. D. Geschwind, H. Shu, A. Haman, J. J. Sejvar and B. L. Miller, Rapidly progressive dementia, *Ann. Neurol.* **64**(1) (2008) 97–108.
2. I. Zerr, K. Kallenberg, D. M. Summers, C. Romero, A. Taratuto, U. Heinemann, M. Breithaupt, D. Varges, B. Meissner, A. Ladogana, M. Schuur, S. Haik, S. J. Collins, G. H. Jansen, G. B. Stokin, J. Pimentel, E. Hewer, D. Collie, P. Smith, H. Roberts, J. P. Brandel, C. van Duijn, M. Pocchiari, C. Begue, P. Cras, R. G. Will and P. Sanchez-Juan, Updated clinical diagnostic criteria for sporadic Creutzfeldt-Jakob disease, *Brain* **132** (2009) 2659–2668.
3. P. Parchi, A. Giese, S. Capellari, P. Brown, W. Schulz-Schaeffer, O. Windl, I. Zerr, H. Budka, N. Kopp, P. Piccardo, S. Poser, A. Rojiani, N. Streichemberger, J. Julien, C. Vital, B. Ghetti, P. Gambetti and H. Kretzschmar, Classification of sporadic Creutzfeldt-Jakob disease based on molecular and phenotypic analysis of 300 subjects, *Ann. Neurol.* **46** (1999) 224–233.
4. H. G. Wieser, K. Schindler and D. Zumsteg, EEG in Creutzfeldt-Jakob disease, *Clin. Neurophysiol.* **117** (2006) 935–951.
5. U. Aguglia, A. Gambardella, E. Le Piane *et al.*, Disappearance of periodic sharp wave complexes in Creutzfeldt-Jakob disease, *Clin. Neurophysiol.* **27** (1997) 277–282.
6. S. Gasparini, E. Ferlazzo, D. Branca, A. Labate, V. Cianci, M. A. Latella and U. Aguglia, Teaching neuroimages: Pseudohypertrophic cerebral cortex in end-stage Creutzfeldt-Jakob disease, *Neurology* **80** (2013) e21.

7. M. H. Rosenbloom and A. Atri, The evaluation of rapidly progressive dementia, *Neurologist* **17** (2011) 67–74.
8. R. W. Paterson, C. C. Torres-Chae, A. L. Kuo, T. Ando, E. A. Nguyen, K. Wong, S. J. DeArmond, A. Haman, P. Garcia, D. Y. Johnson, B. L. Miller and M. D. Geschwind, Differential diagnosis of Jakob-Creutzfeldt disease, *Arch. Neurol.* **69** (2012) 1578–1582.
9. J. Schmidhuber, Deep learning in neural networks: An overview, *Neural Netw.* **61** (2015) 85–117.
10. Y. Bengio, Y. LeCun and G. Hinton, Deep learning, *Nature* **521** (2015) 436–444.
11. Y. Bengio, Learning deep architectures for AI, *Found. Trends Mach. Learn.* **2**(1) (2009) 1–127.
12. L. Deng and D. Yu, Deep learning: Methods and applications, *Found. Trends Signal Process.* **7**(3–4) (2014) 197–387.
13. H.-II Suk, D. Shen, Deep learning-based feature representation for AD/MCI classification, *Med. Image Comput. Comput. Assist. Interv.* **16**(2) (2013) 583–590.
14. H. Schulz, K. Cho, T. Raiko and S. Behnke, Two-layer contractive encodins for learning stable nonlinear features, *Neural Netw.* **64** (2015) 4–11.
15. G. Hinton and R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science* **313**(5786) (2006) 540–507.
16. H. S. Lee, N. Sambuughin, L. Cervenakova, J. Chapman, M. Pocchiari, S. Litvak, H. Y. Qi, H. Budka, T. del Ser, H. Furukawa, P. Brown, D. C. Gajdusek, J. C. Long, A. D. Korczyn and L. G. Goldfarb, Ancestral origins and worldwide distribution of the PRNP 200 K mutation causing familial Creutzfeldt–Jakob disease, *Am. J. Hum. Genet.* **64** (1999) 1063–70.
17. S. J. Collins, P. Sanchez-Juan, C. L. Masters, G. M. Klug, C. van Duijn, A. Poleggi, M. Pocchiari, S. Almonti, N. Cuadrado-Corrales, J. de Pedro-Cuesta, H. Budka, E. Gelpi, M. Glatzel, M. Tolnay, E. Hewer, I. Zerr, U. Heinemann, H. A. Kretszchmar, G. H. Jansen, E. Olsen, E. Mitrova, A. Alpérovitch, J. P. Brandel, J. Mackenzie, K. Murray and R. G. Will, Determinants of diagnostic investigation sensitivities across the clinical spectrum of sporadic Creutzfeldt-Jakob disease, *Brain* **129** (2006) 2278–2287.
18. A. Vincent, C. G. Bien, S. R. Irani and P. Waters Autoantibodies associated with diseases of the CNS: New developments and future challenges, *Lancet Neurol.* **10** (2011) 759–772.
19. J. Y. Chong, L. P. Rowland and R. D. Utiger, Hashimoto encephalopathy: Syndrome or myth? *Arch Neurol.* **60** (2003) 164–171.
20. E. Ferlazzo, M. Raffaele, I. Mazzù and F. Pisani, Recurrent status epilepticus as the main feature of Hashimoto's encephalopathy, *Epilepsy Behav.* **8** (2006) 328–330.
21. P. H. Schur, Neuropsychiatric manifestations of systemic lupus erythematosus, UpToDate, Post, TW (Ed), UpToDate, Waltham, MA (2015).
22. E. Ferlazzo, A. Gambardella, M. Bellavia, S. Gasparini, L. Mumoli, A. Labate, V. Cianci, C. Russo and U. Aguglia, Positivity to p-ANCA in patients with status epilepticus, *BMC Neurol.* **14** (2014) 148.
23. G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jr Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub and C. H. Phelps, The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, *Alzheimers Dement.* **7**(3) (2011) 263–269.
24. E. Ferlazzo, N. Mammone, V. Cianci *et al.*, Permutation entropy of scalp EEG: A tool to investigate epilepsies: Suggestions from absence epilepsies, *Clin. Neurophysiol.* **125** (2014) 13–20.
25. H. Adeli and M. Ghosh Dastidar, *Automated EEG-Based Diagnosis of Neurological Disorder* (CRC Press, 2010).
26. A. Grossmann and J. Morlet, Decomposition of Hardy functions into square integrable wavelets of constant shape, *SIAM J. Math. Anal.* **15** (1984) 723–736.
27. U. R. Acharya, S. Vinitha Sree, A. P. C. Alvin and J. S. Suri, Application of non-linear and wavelet based features for the automated identification of epileptic EEG signals, *Int. J. Neural Syst.* **22**(2) (2012) 1250002-1-1250002-14.
28. F. B. Vialatte, C. Martin, R. Dubois *et al.*, A machine learning approach to the analysis of time-frequency maps, and its application to neural dynamics, *Neural Netw.* **20**(2) (2007) 194–209.
29. H. Adeli, S. Ghosh-Dastidar and N. Dadmehr, A spatio-temporal wavelet-chaos methodology for EEG-based diagnosis of Alzheimer's disease, *Neurosci. Lett.* **444**(2) (2008) 190–194.
30. A. Ahmadlou, H. Adeli and A. Adeli, Fractality and a wavelet-chaos methodology for EEG-based diagnosis of Alzheimer's disease, *Alzheimer Dis. Assoc. Disorders* **25**(1) (2011) 85–92.
31. Z. Sankari and H. Adeli, Probabilistic neural networks for EEG-based diagnosis of Alzheimer's disease using conventional and wavelet coherence, *J. Neurosci. Methods* **197**(1) (2011) 165–170.
32. W. Y. Hsu, Assembling a multi-feature EEG classifier for left-right motor data using wavelet-based fuzzy approximate entropy for improved accuracy, *Int. J. Neural Syst.* **25**(8) (2015) 1550037.
33. H. Schulz, K. Cho, T. Raiko and S. Behnke, Two-layer contractive encodings for learning stable nonlinear features, *Neural Netw.* **64** (2015) 4–11.

34. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, NY, 1995).
35. V. Vapnik, S. Golowich and A. Smola, Support vector method for function approximation, regression estimation, and signal processing, in *Advances in Neural Information Processing Systems 9*, eds. M. Mozer, M. Jordan and T. Petsche (MIT Press, Cambridge, MA, 1997), pp. 281–287.
36. B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers, in *Proc. Fifth Annual Workshop on Computational Learning Theory–COLT '92* (1992), p. 144.
37. C. Cortes and V. Vapnik, Support vector networks, *Mach. Learn.* **20** (1995) 273–297.
38. Y. Zhang and W. Zhou, Multifractal analysis and relevance vector machine-based automatic seizure detection in intracranial, *Int. J. Neural Syst.* **25**(6) (2015) 1550020.
39. E. Castillo, D. Peteiro-Barral, B. Guijarro Berdinas and O. Fontenla-Romero, Distributed one-class support vector machine, *Int. J. Neural Syst.* **25**(7) (2015) 1550029.
40. F. C. Morabito, D. Labate, F. La Foresta, A. Bramanti, G. Morabito and I. Palamara, Multivariate multi-scale permutation entropy for complexity analysis of Alzheimer's disease EEG, *Entropy* **14** (2012) 1186–1202.
41. M. Berglund, T. Raiko and K. Cho, Measuring the usefulness of hidden units in Boltzmann machines with mutual information, *Neural Netw.* **64** (2015) 12–18.
42. W. Hulme, P. Richtarik, L. McGuire and A. Green, Optimal diagnostic tests for sporadic Creutzfeldt-Jakob disease based on support vector machine classification of RT-QuIC data, arXiv.org, q-bio (2012).
43. P. S. Wang, Y. T. Wu, C. I. Hung, S. Y. Kwan, S. Teng and B. W. Soomg, Early detection of periodic sharp wave complexes on EEG by independent component analysis in patients with Creutzfeldt-Jakob disease, *J. Clin. Neurophysiol.* **25**(1) (2008) 25–31.
44. C.-I. Hung, P.-S Wang, B.-W Soong, S. Teng and J.-C. Hsieh, Blind source separation of concurrent disease-related patterns from EEG in Creutzfeldt-Jakob disease for assisting early diagnosis, *Computational Neuroscience, Optimization and Its Applications Series* **38** (2012) 57–74.
45. Z. Vahabi, R. Amirfattahi, F. Ghassemi and F. Shayegh, Online epileptic seizure prediction using wavelet-based bi-phase correlation of electrical signal tomography, *Int. J. Neural Syst.* **25**(6) (2015) 1550028.
46. F. C. Morabito, M. Campolo, D. Labate, G. Morabito, L. Bonanno, A. Bramanti, S. de Salvo, A. Marra and P. Bramanti, A longitudinal EEG study of Alzheimer's disease progression based on a complex network approach, *Int. J. Neural Syst.* **25**(2) (2015) 1550005.
47. H. Adeli, S. Ghosh-Dastidar and N. Dadmehr, Alzheimer's disease and models of computation: Imaging, classification, and neural models, *J. Alzheimer's Dis.* **7**(3) (2005) 187–199.
48. H. Adeli, S. Ghosh-Dastidar and N. Dadmehr, Alzheimer's disease: Models of computation and analysis of EEGs, *Clin. EEG Neurosci.* **36**(3) (2005) 131–140.
49. M. Ahmadlou, H. Adeli and A. Adeli, New diagnostic EEG markers of the Alzheimer's disease using visibility graph, *J. Neural Transm.* **117**(9) (2010) 1099–1109.
50. Z. Sankari, H. Adeli and A. Adeli, Wavelet coherence model for diagnosis of Alzheimer's disease, *Clin. EEG Neurosci.* **43**(3) (2012) 268–278.