Computers in Biology and Medicine 106 (2019) 71–81

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals

Nicola Michielli[a], U. Rajendra Acharya[b,c,d], Filippo Molinari[a,*]

[a] Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy
[b] Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Singapore
[c] Department of Biomedical Engineering, School of Science and Technology, SUSS University, Clementi, 599491, Singapore
[d] School of Medicine, Faculty of Health and Medical Sciences, Taylor's University, Malaysia

ARTICLE INFO

Keywords:
Recurrent neural networks
Long short-term memory
Sleep analysis
EEG signals
Neurocognitive performance

ABSTRACT

Automated evaluation of a subject's neurocognitive performance (NCP) is a relevant topic in neurological and clinical studies. NCP represents the mental/cognitive human capacity in performing a specific task. It is difficult to develop the study protocols as the subject's NCP changes in a known predictable way. Sleep is time-varying NCP and can be used to develop novel NCP techniques. Accurate analysis and interpretation of human sleep electroencephalographic (EEG) signals is needed for proper NCP assessment. In addition, sleep deprivation may cause prominent cognitive risks in performing many common activities such as driving or controlling a generic device; therefore, sleep scoring is a crucial part of the process. In the sleep cycle, the first stage of non-rapid eye movement (NREM) sleep or stage N1 is the transition between wakefulness and drowsiness and becomes relevant for the study of NCP.

In this study, a novel cascaded recurrent neural network (RNN) architecture based on long short-term memory (LSTM) blocks, is proposed for the automated scoring of sleep stages using EEG signals derived from a single-channel. Fifty-five time and frequency-domain features were extracted from the EEG signals and fed to feature reduction algorithms to select the most relevant ones. The selected features constituted as the inputs to the LSTM networks. The cascaded architecture is composed of *two* LSTM RNNs: the first network performed 4-class classification (i.e. the five sleep stages with the merging of stages N1 and REM into a single stage) with a classification rate of 90.8%, and the second one obtained a recognition performance of 83.6% for 2-class classification (i.e. N1 vs REM). The overall percentage of correct classification for *five* sleep stages is found to be 86.7%. The objective of this work is to improve classification performance in sleep stage N1, as a first step of NCP assessment, and at the same time obtain satisfactory classification results in the other sleep stages.

1. Introduction

Neurocognitive performance (NCP) represents the mental/cognitive human capacity in performing a specific task [1]. Numerically and accurate evaluation of the subject's NCP is currently an open problem in several fields, such as rehabilitation, neurology, psychology/psychiatry and base research studies. NCP assessment relies primary on accurate information extraction from the electroencephalographic (EEG) signal and on its interpretation and classification. It is difficult to develop the study protocols in which the subject clearly changes its NCP in a known predictable way. Sleep analysis may be considered as an example of time-varying NCP, since the functional aspects of the brain vary in different sleep stages. Sleep plays an essential role in human health because it represents one of the primary functions of the human brain. The life of a human subject is constituted of sleep cycle for one third of its duration and the quality of sleep may be influenced by sleep-related disorders like insomnia, hypersomnia, narcolepsy, sleep apnea, breathing-related disorders, depression and circadian rhythm disorders [2]. Sleep deprivation, considered as a result of a sleep pathology or stress-related disorder, causes prominent cognitive risks in performing many common activities such as driving or controlling a generic device [3]. In fact, according to the National Highway Traffic Safety Administration in the USA, the reduction of reaction times due to drowsiness while driving causes between 56,000 and 100,000 car accidents, resulting in more than 1500 deaths and 71,000 injuries annually [4]. In this context, the variable called sleep onset period (SOP), i.e. the period interposed between weak wakefulness and drowsiness, becomes very important for the study of NCP. In the sleep stage classification, the

[*] Corresponding author. Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy.
E-mail address: filippo.molinari@polito.it (F. Molinari).

https://doi.org/10.1016/j.compbiomed.2019.01.013
Received 13 December 2018; Received in revised form 15 January 2019; Accepted 15 January 2019

non-rapid eye movement (NREM) sleep stage 1 (N1), considered as the first stage of sleep cycle, represents the center of the SOP [5]. For this reason, an accurate scoring of sleep stages, with a particular focus on stage N1, is considered a crucial part of the process.

Several polysomnographic (PSG) signals are acquired for sleep scoring: the EEG signals for monitoring brain activity, the electrooculographic (EOC) signals for eye movements and the electromyographic (EMG) signals to measure muscle tone. In general sleep signals are visually scored by experts according to two available guidelines: the Rechtschaffen and Kales's (R&K) standard [6] and the manual proposed, in a more recent period, by the American Academy of Sleep Medicine (AASM) [7]. The main change is in terminology: in the AASM manual the state of sleep is split into five sleep stages: wakefulness (stage W), non-rapid eye movement (NREM) sleep stage 1 (N1), NREM sleep stage 2 (N2), NREM sleep stage 3 (N3) and rapid eye movement (REM) sleep stage. The two R&K stages S3 and S4 have been combined into a single stage N3, also called slow wave sleep (SWS) stage. The AASM rules define the characteristic waves for each of the five sleep stages:

- W (Wakefulness): stage W is characterized by alpha (8–12 Hz) and beta (16–30 Hz) waves;
- N1 (NREM 1): stage N1 is scored when theta (4–8 Hz) waves are evident, and vertex sharp waves may be present;
- N2 (NREM 2): stage N2 is scored when high voltage biphasic waves (K-complexes) and sleep spindles (12–16 Hz) are noted and theta waves are present;
- N3 (NREM 3): stage N3 is characterized by high amplitude ($> 75\,\mu V$) delta (0.5–4 Hz) waves;
- REM: stage REM is scored when theta and sawtooth (2–6 Hz) waves are evident and alpha waves may be present.

Sleep scoring is a complex procedure, because differences among the stages are often very subtle. Several authors proposed automatic classification systems to support the scoring made by sleep specialists. These methods are based on *two* main strategies: *i)* multi-channel and *ii)* single-channel recording. In the first approach *i)*, a different number of PSG signals are used (more than one EEG channel, the EMG signal and the EOG signals especially for REM detection). But this kind of scoring imposes limitations on the subject's movements and could be a limit for NCP assessment in real conditions, such as during driving. In the second approach *ii)*, only a single EEG channel is used to extract informative features. This approach reduces the instrumental complexity and eases the experimental recordings. The standard procedure for an automatic sleep stage classification (ASSC) system is composed of four stages: *i)* data acquisition, *ii)* signal pre-processing, *iii)* feature extraction and *iv)* classification [8]. The *iii)* feature extraction process is based on the estimation of characteristic parameters from the pre-processed EEG signals of stage *ii)*. These parameters can be computed in time, frequency, time-frequency or complexity/nonlinear domain [9]. In few ASSC systems, there is another step before the classifier, called feature selection or dimensionality reduction [10]. This additional step is useful to reduce the computation cost by removing the most redundant features (feature selection) or to generate new features in a lower-dimensional subspace (dimensionality reduction). According to our survey, the most popular algorithms used to perform feature selection in the ASSC system are: sequential forward and backward selection methods [11], minimum redundancy maximum relevance (mRMR) [12], relief algorithm [13] and principal component analysis (PCA) [14] and linear discriminant analysis (LDA) [15] as dimensionality reduction techniques. Automatic sleep stage classification methods include: support vector machine (SVM) [8,16–18], gaussian mixture model (GMM) [19], bootstrap aggregating (Bagging) [20,21], J-means clustering [22], random forest classifier [23], k-means clustering [24] and artificial neural networks (ANNs) [25,26]. Recently deep learning methods have been applied to sleep stage classification. Recurrent neural networks (RNNs) applied to sequence data [27,28] and convolutional neural networks (CNNs) [29,30], most commonly applied to image data, are the most relevant algorithms in this field. RNNs have several variants including long short-term memory (LSTM) [31], gated recurrent unit (GRU) [32] and bidirectional RNNs [33].

The current overall performance in terms of stage N1, which is the most complex stage to identify, in most of the works published, is less than 40%. To the best of our knowledge, the only proposed method, which employed a RNN classifier for the automatic sleep stage classification, obtained a classification accuracy of 36.7% [27]. In this work, a novel approach using a cascaded RNN architecture with LSTM units is proposed to classify the sleep EEG signals to overcome the current limitation of multi-channel approaches and the low N1 sleep accuracy. The importance of this stage becomes relevant in the context of the NCP assessment; in fact, the next step towards the NCP should be the identification of the hypnagogic state, which is defined as the contact point between waking and sleeping and is considered as the opposite of the hypnopompic state. Hori et al. [5] studied the time and spatial-domain transitions of the EEG signals during the hypnagogic state and proposed a new sleep scoring made by nine stages: the first two stages correspond to stage W in AASM standard, EEG stages 3–8 correspond to sleep stage N1 and the last stage corresponds to sleep stage N2. It's clear from this new classification that most of the hypnagogic EEG stages are classified as stage N1 in the AASM standard. For this reason, the objective of this work is to improve the low classification performance in sleep stage N1 and, at the same time, to obtain satisfactory results in the other sleep stages. In the following section, an exhaustive description of the method is presented.

2. Materials and methods

In this paper, we propose a novel automatic classification model, consisted of two different RNNs with LSTM units. The first performed 4-class classification (W, N1-REM, N2 and N3), while the second performed binary classification (N1 vs REM). Both RNNs shared the first three steps: data acquisition, signal pre-processing and feature extraction from single-channel EEG signals. Subsequently, a feature selection or feature transformation method was adopted to reduce the number of input features for neural network. Two different methods were considered: for the first RNN the mRMR algorithm was used for feature selection, while for the second RNN, the PCA was employed for dimensionality reduction. Finally, the two RNNs were connected in cascade, with the aim of classifying five different sleep stages. The workflow of the proposed strategy is schematically described in Fig. 1.

2.1. Data acquisition

The dataset used in this study was the Sleep-EDF database, which is publicly available from the PhysioBank [34]. The sampling rate of the EEG signals was 100 Hz. For our study, we selected twelve recordings (SC4001E0, SC4002E0, SC4011E0, SC4012E0, SC4021E0, SC4031E0, SC4051E0, SC4061E0, SC4112E0, SC4122E0, SC4131E0, SC4182E0) of 10 healthy Caucasians aged 26–33 years. The other subjects were excluded from the analysis because their recordings contained movements and unreported epochs. In addition, we merged the stage 3 and 4 into a single stage N3, as it is currently recommended by the AASM standard. Only the EEG Fpz-Cz signals were used as single-channel in this work because K-complexes and sleep-spindles (typical patterns of stage N2) could be recorded in central/frontal regions and during stage N1, vertex sharp waves may be present which often occurs in central/frontal brain areas, according to the AASM guidelines. The number of 30 s epochs for each of the five sleep stages are shown in Table 1.

In this study, different from most of studies (ASSC systems), we randomly selected a relatively small number of epochs in stage W (8%). This because the characteristics of the W epochs are very well defined and have relatively low-variability. This makes the W epochs the easiest to recognize without the need for many epochs in the dataset.

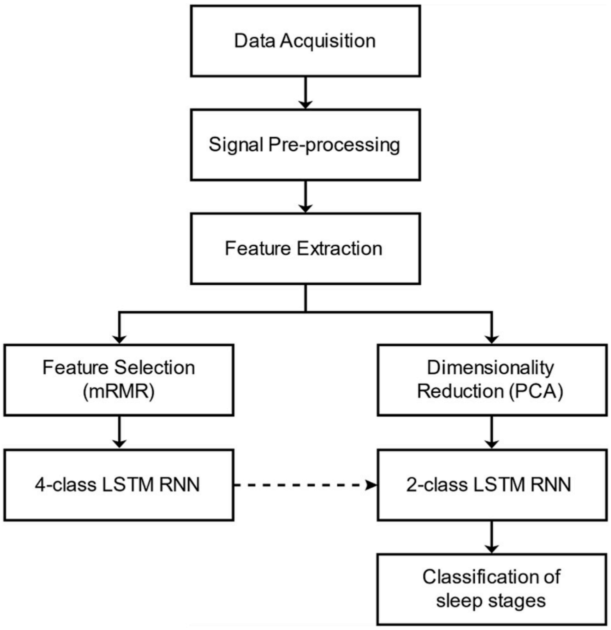


Fig. 1. Schematic representation of the proposed workflow.

Table 1
The data distribution in various stages of sleep.

Sleep stages	W	N1	N2	N3	REM	Total
Number of epochs in stages	850	920	4960	1690	1860	10280

2.2. Signal pre-processing

All EEG time-series data were filtered to remove the frequency components outside the range of 0.3–45 Hz. Next, the 30 s filtered epochs were subdivided in blocks of 1 s duration, thus for each epoch we obtained 30 time-segments. A number of timesteps equals to 30 was set for the RNN classifier.

2.3. Feature extraction

In the feature extraction process, a total of 55 features were identified from a single EEG channel (Fpz-Cz). All features were computed for each 1 s filtered epoch for two reasons: the first is the non-stationarity of the EEG signal and the second is the size of the input sequence at each time step for the RNN classifier.

2.3.1. Time-domain features

Statistical parameters: mean (1st raw moment), variance (2nd central moment), skewness (normalized 3rd central moment), kurtosis (normalized 4th central moment) and median were the first time-domain features extracted from EEG signals. The mathematical expressions of the statistical moments are reported below:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

$$var(x) = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{2}$$

$$skew(x) = \frac{\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^3}{var(x)^{3/2}} \tag{3}$$

$$kurt(x) = \frac{\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^4}{var(x)^2} \tag{4}$$

Other time-domain features: peak-to-peak amplitude (difference between the maximum positive and negative amplitudes), absolute maximum value, number of zero crossings (number of time signal crosses time-axis), root mean square (RMS) and average rectified value (ARV). The last two parameters are defined as:

$$RMS(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |x_i|^2} \tag{5}$$

$$ARV(x) = \frac{1}{N}\sum_{i=1}^{N} |x_i| \tag{6}$$

In this analysis were also used Hjorth parameters, i.e. Hjorth mobility (HM) and Hjorth complexity (HC), introduced by Bo Hjorth in 1970 [35]:

$$HM(x) = \sqrt{\frac{var(dx/dt)}{var(x)}} \tag{7}$$

$$HC(x) = \frac{HM(dx/dt)}{HM(x)} \tag{8}$$

Next, we employed six infinite impulse response (IIR) Butterworth band-pass filters in 0.5–4 Hz, 4–8 Hz, 8–12 Hz, 12–16 Hz, 16–30 Hz and 30–45 Hz to separate the delta, theta, alpha, sigma (or sleep spindle sub-band), beta and gamma waves, of each 1 s epoch, respectively. The peak-to-peak amplitude was computed for each of the six sub-bands (A_{delta}, A_{theta}, A_{alpha}, A_{sigma}, A_{beta} and A_{gamma}) in which the EEG signal was decomposed. We also computed the energy from each 1 s epoch filtered using the IIR filters in the six significant EEG sub-bands (E_{delta}, E_{theta}, E_{alpha}, E_{sigma}, E_{beta} and E_{gamma}). The wave energy is defined as the sum of squared magnitude of each signal component:

$$Energy = \sum_{i=1}^{N} |x_i|^2 \tag{9}$$

In addition, five energy ratios were computed for the first five sub-bands referred to the energy in gamma sub-band. Hence, the total number of time-domain features is 29.

2.3.2. Frequency-domain features

Spectral estimation was performed to transfer the time series to the frequency domain. We used the non-parametric approach: Power Spectral Density (PSD) values were computed from the signal samples multiplied by a window function. In this study Welch's method [36] was used to achieve PSD. Next, mean frequency (MNF), spectral entropy (SE) and Renyi entropy (RE) [37] were computed for each 1 s epoch. The mathematical expressions are reported below.

$$MNF = \frac{\sum_i P_i \cdot f_i}{\sum_i P_i} \tag{10}$$

$$SE = \sum_i p_i \cdot \log\left(\frac{1}{p_i}\right) \tag{11}$$

$$RE = -\log\left(\sum_i p_i^2\right) \tag{12}$$

where P is the power spectrum and p is the normalized power spectrum. The relative spectral powers in the six significant frequency sub-bands (P_{delta}, P_{theta}, P_{alpha}, P_{sigma}, P_{beta} and P_{gamma}) were computed from the obtained PSD values, as the ratios between the absolute power in each frequency sub-band and the total area of PSD function over frequencies. From these features, 15 power ratios were also derived. The last two features, used in this study, were the products of the relative powers in

Table 2
List of extracted features.

N	Feature	N	Feature	N	Feature	N	Feature
1	Peak-to-peak amplitude	15	Amplitude (alpha sub-band)	29	E_{beta}/E_{gamma}	43	P_{delta}/P_{gamma}
2	Arithmetic mean	16	Amplitude (sigma sub-band)	30	Mean frequency	44	P_{theta}/P_{alpha}
3	Absolute maximum value	17	Amplitude (beta sub-band)	31	Spectral entropy	45	P_{theta}/P_{sigma}
4	Median	18	Amplitude (gamma sub-band)	32	Renyi entropy	46	P_{theta}/P_{beta}
5	Variance	19	Energy (delta sub-band)	33	Power (delta sub-band)	47	P_{theta}/P_{gamma}
6	Skewness	20	Energy (theta sub-band)	34	Power (theta sub-band)	48	P_{alpha}/P_{sigma}
7	Kurtosis	21	Energy (alpha sub-band)	35	Power (alpha sub-band)	49	P_{alpha}/P_{beta}
8	Root mean square	22	Energy (sigma sub-band)	36	Power (sigma sub-band)	50	P_{alpha}/P_{gamma}
9	Average rectified value	23	Energy (beta sub-band)	37	Power (beta sub-band)	51	P_{sigma}/P_{beta}
10	Number of zero crossings	24	Energy (gamma sub-band)	38	Power (gamma sub-band)	52	P_{sigma}/P_{gamma}
11	Hjorth mobility	25	E_{delta}/E_{gamma}	39	P_{delta}/P_{theta}	53	P_{beta}/P_{gamma}
12	Hjorth complexity	26	E_{theta}/E_{gamma}	40	P_{delta}/P_{alpha}	54	$P_{delta}\cdot P_{theta}$
13	Amplitude (delta sub-band)	27	E_{alpha}/E_{gamma}	41	P_{delta}/P_{sigma}	55	$P_{alpha}\cdot P_{beta}$
14	Amplitude (theta sub-band)	28	E_{sigma}/E_{gamma}	42	P_{delta}/P_{beta}		

two low-frequency sub-bands (delta and theta) and in two high frequency sub-bands (alpha and beta). In total, we extracted 26 features in frequency domain. All extracted features are summarized in Table 2.

At the end of this process we extracted for each 30 s epoch, of each sleep stage, 55 features for each time step; hence our data matrix had dimensions of 55 × 30.

2.4. Feature selection

The minimum redundancy maximum relevance (mRMR) algorithm [12] was used to perform feature selection. This method balances the need for including the most relevant features (i.e. those that are correlated with the target, in our case the sleep stages) and the need for excluding the most redundant variables (i.e. those that are strongly correlated with each other). Feature relevance and redundancy are characterized in terms of mutual information. If we consider two discrete random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass function $p(x)$ and $p(y)$; their mutual information $I(X;\,Y)$ is defined as [38]:

$$I(X;\,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,\,y)\cdot\log\frac{p(x,\,y)}{p(x)\cdot p(y)} \tag{13}$$

We used the logarithmic base 2, thus the units of mutual information were bits. In the mRMR algorithm, the purpose is to search a feature set S with m features $\{x_i\}$, which should maximize the difference between the relevance defined as the mutual information between individual features x_i and target class c and the redundancy, defined as the mutual information between features x_i and feature x_j. The objective is to maximize the following mRMR cost function:

$$\phi_{mRMR} = \frac{1}{|S|}\sum_{x_i\in S} I(x_i;\,c) - \frac{1}{|S|^2}\sum_{x_i,x_j\in S} I(x_i;\,x_j) \tag{14}$$

where $|S|$ is the number of features in the feature subset S. After having selected sequential features, according to this algorithm, we computed for each incremental feature subset the corresponding value of the mRMR cost function and we defined the optimal subset size as the global maximum point, as shown in Fig. 2.

According to this supervised algorithm, the following 11 features were selected: median, skewness, root mean square, Hjorth mobility, amplitude (theta sub-band), amplitude (sigma sub-band), amplitude (beta sub-band), amplitude (gamma sub-band), E_{beta}/E_{gamma}, power (theta sub-band) and power (beta sub-band).

2.5. Dimensionality reduction

In this work, principal component analysis (PCA) is employed for dimensionality reduction. PCA is an unsupervised statistical technique for the reduction of the dimension of features in a new lower-dimensional subspace [14]. PCA produces a smaller number of uncorrelated variables from the original set of correlated variables, without loss of information. These new variables are called *principal components*. This procedure can be used as a form of dimensionality reduction if the eigenvectors corresponding to smaller eigenvalues are discarded. Before applying PCA, it's standard practice to perform feature scaling in order to have features with zero mean and comparable ranges of values. Dimensionality reduction was performed by selecting only a subset of principal components to retain the 95% of variance. In this work, 27 principal components were selected according to variance criterion and were stored in the matrix of coefficients U; thus, the new feature data matrix for the i -th 30 s sleep EEG epoch was computed as:

$$X_{pca}^{(i)} = U^T\cdot X^{(i)} \tag{15}$$

The new reduced data matrix was used as input for the second LSTM RNN which performed binary classification (N1 vs REM). Fig. 3 shows the absolute value (normalized between 0 and 1) of the coefficients of three principal components (those that were more correlated with the target) which multiplied the original feature values in the linear combination.

2.6. Recurrent neural networks

RNN architecture is a powerful deep learning classification method especially applied to sequential data. RNNs are currently state of the art methods in natural language processing (NLP) and speech recognition [39]. In fact, language data can be considered sequences, such as words (sequence of letters), sentences (sequences of words) and documents (sequence of sentences). RNNs are a particular form of standard artificial neural networks (ANNs) with the advantage of modelling time series with long range structural dependencies [40]. The basic idea in RNNs is to add time delay unit and a feedback connection so that information from previous state can be used in the subsequent state.

In the RNN architecture the input layer is a sequence layer which takes the input a sequence of vectors $\{x^{<1>},\ldots,x^{<t>},\ldots,x^{<T>}\}$, which contain all the features for each timestep; then the network computes a sequence of hidden activations $\{a^{<1>},\ldots,a^{<t>},\ldots,a^{<T>}\}$ and the output vector $\{\hat{y}^{<1>},\ldots,\hat{y}^{<t>},\ldots,\hat{y}^{<T>}\}$ for T timsteps. The first activation $a^{<0>}$ is usually a vector of zeros. The activation and the output prediction at time t are expressed as:

$$a^{<t>} = g(W_a\cdot[a^{<t-1>},\,x^{<t>}] + b_a) \tag{16}$$

$$\hat{y}^{<t>} = g(W_y\cdot a^{<t>} + b_y) \tag{17}$$

where the vector in square brackets is the vector concatenation of

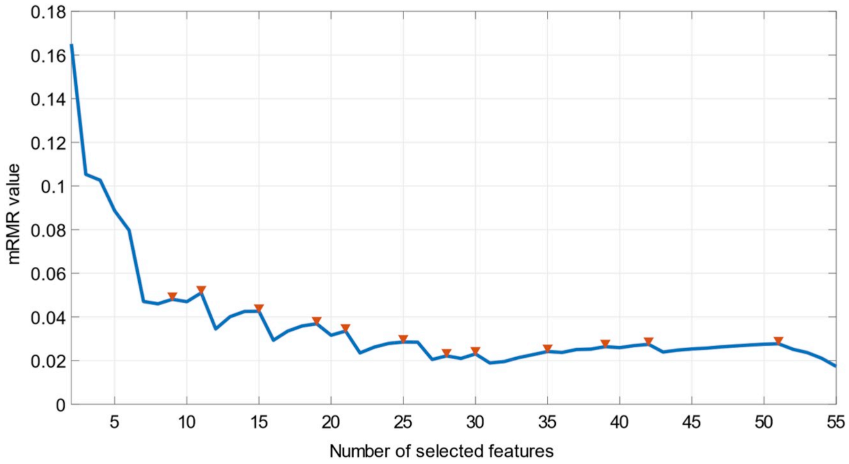


Fig. 2. mRMR cost function value for different numbers of selected features with local maxima.

activation at previous timestep and input at current timestep, W_a and W_y are the activation and output weight matrix respectively, while b_a and b_y are the activation and output bias term respectively. The operator g represents a generic activation function, which may be different for the activation and the output estimation. The characteristic of RNN is that each neuron of the hidden layer receives the activation of the previous time step to compute the activation of the current time step. So, in the RNN, the prediction of the output at the current time step $\hat{y}^{<t>}$ is done not only with the information in the input $x^{<t>}$, but also with the information from $x^{<1>}$ to $x^{<t-1>}$ through the activation $a^{<t-1>}$ at previous timestep. This architecture is called unidirectional RNN, because it uses information from the earlier sequence inputs to estimate the prediction at a certain time step, but no information later in the sequence, like in the bidirectional RNN. The equations 16 and 17 define the forward propagation in the RNN. In the backward propagation the weights and bias terms are iteratively updated using an optimization algorithm. In this case, the partial derivatives of the cost function are computed by propagating back through all timesteps. For this reason, this process is called *backpropagation through time* (BPTT). The main difficulty in training RNN is due to the vanishing gradient problem [41]: partial derivatives become very small in deep layers for large timesteps and the network parameters (weights and bias terms) can't change in the subsequent iterations and consequently the network stops learning. In order to solve this problem, RNN unit is replaced by a gated cell called long short-term memory (LSTM) unit.

The LSTM unit was introduced by Hochreiter and Schmidhuber in 1997 [31] and it represents a modification to the standard RNN that makes it much better capturing long-term dependencies and allows to address the problem of vanishing gradient. The LSTM memory cell consists of five components: the memory cell $c^{<t>}$ (a new variable computed for each timestep), the candidate value $\tilde{c}^{<t>}$ for replacing the memory cell at each timestep and three gates defined as update gate Γ_u, forget gate Γ_f and output gate Γ_o. The memory cell is useful to remember certain values even for a long time during the training process. The three gates can assume only values between 0 and 1 and for each of them a weight matrix and a bias term will be updated during the training process. The forget gate allows to decide what information may be thrown away and its expression is reported below:

$$\Gamma_f = \sigma(W_f \cdot [a^{<t-1>}, x^{<t>}] + b_f) \tag{18}$$

The update gate allows to decide whether or not to replace the memory cell with the candidate value. It can be expressed as:

$$\Gamma_u = \sigma(W_u \cdot [a^{<t-1>}, x^{<t>}] + b_u) \tag{19}$$

Finally, the output gate is the section where the activation at the current timestep is generated and can be defined as:

$$\Gamma_o = \sigma(W_o \cdot [a^{<t-1>}, x^{<t>}] + b_o) \tag{20}$$

In the previous expressions, σ represents the sigmoid function. The equations that govern the behavior of the LSTM unit are reported below:

$$\tilde{c}^{<t>} = \tanh(W_c \cdot [a^{<t-1>}, x^{<t>}] + b_c) \tag{21}$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>} \tag{22}$$

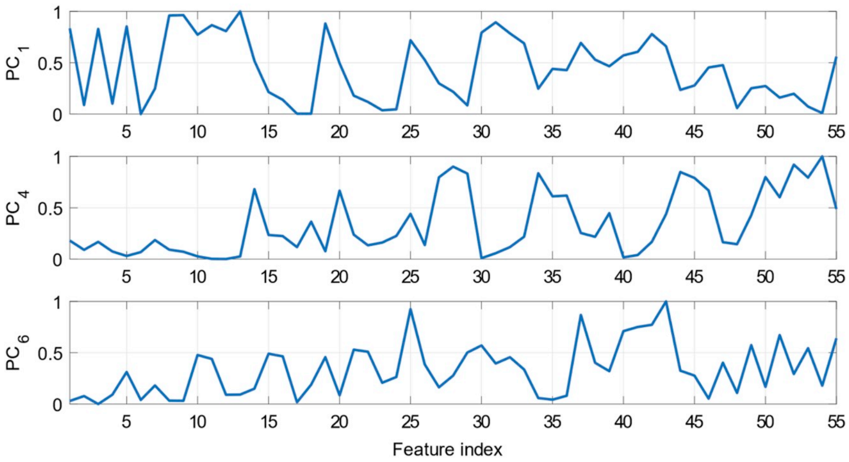


Fig. 3. PC_1 , PC_4 and PC_6 coefficients respect to the original features.

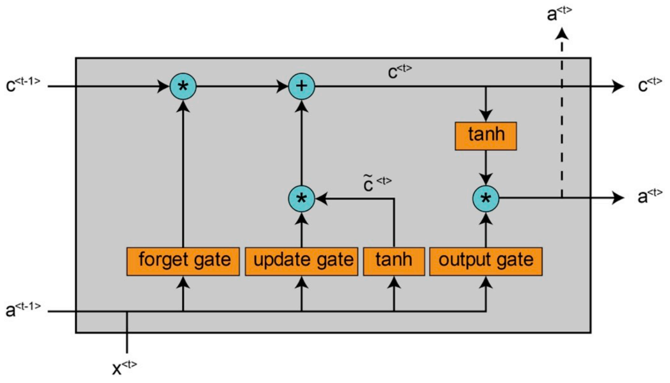


Fig. 4. Long short-term memory (LSTM) unit.

$$a^{<t>} = \Gamma_o * \tanh(c^{<t>}) \tag{23}$$

where W_c and b_c are the cell weight matrix and bias term respectively, the notation $*$ denotes the Hadamard (i.e. element-wise) product and *tanh* is the hyperbolic tangent function. The LSTM unit and its internal structure is reported in Fig. 4.

The idea of using RNNs for EEG classification came from the temporal and highly non-linear nature of the EEG signal [9]. In this study, we employed a cascade of two RNNs with LSTM units. The first network took the input the features selected by mRMR algorithm and performed 4-class (W, N1-REM, N2 and N3) classification (the N1 and REM epochs were merged into a single class), while the second network used the input new features computed by PCA of the correctly classified N1-REM epochs by the first RNN and classified these epochs into two classes (N1 and REM). Both RNNs proposed in this study presented the same structure: the input layer was a sequence layer with 30 timesteps; the LSTM layers were used to learn the features from EEG signals; the fully connected (FC) layer was used to convert the output size of the previous layers into the number of sleep stages to recognize; the softmax layer computed the probability of each target class over all possible target classes and, in the end, the classification output layer computed the cost function. The main advantage of using the softmax activation function is its output probability range. This function provides output values between 0 and 1 and the sum of all probabilities is equal to one. Its mathematical expression is reported below:

$$\hat{y}_j^{(i)} = \frac{e^{z_j^{(i)}}}{\sum_{j=1}^{C} e^{z_j^{(i)}}} \tag{24}$$

The superscript i refers to the generic training example, the subscript j denotes the generic neuron of the FC layer, z is the output value of the FC layer and C is the number of target classes. The cost function, minimized during the network training process, is a function of all weights W and bias terms b, and is expressed as the average (on training set) of cross entropy functions for C mutually exclusive classes:

$$J(W,\, b) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C} y_j^{(i)}\cdot\log\left(\hat{y}_j^{(i)}\right) \tag{25}$$

where M is the number of training examples, y denotes the true label and $\hat{y}$ is the value estimated by the network. The value of C was 4 for the first RNN and 2 for the second RNN. Fig. 5 shows the structure of the two RNNs with LSTM units proposed in this paper.

The main difference between the two networks was that in the first structure there was a single LSTM layer with a sequence-to-label architecture, while in the second RNN there were two LSTM layers: the first had a sequence-to-sequence architecture and the second had a sequence-to-label architecture.

The adaptive moment estimation (ADAM) algorithm was adopted in this work for backpropagation. ADAM optimization algorithm is a combination of gradient descent with momentum, based on the exponentiality weighted average, and root mean square propagation (RMSProp) algorithm. Thus, it is designed to combine the advantages of the momentum, that works faster than the standard gradient descent algorithm and RMSProp, which solves the optimization problem in non-stationary conditions [42]. The main hyperparameters used for ADAM training algorithm were the same used in the reference: learning rate (α = 0.001), gradient decay factor (β_1 = 0.9), squared gradient decay factor (β_2 = 0.999), and epsilon ($\varepsilon = 10^{-8}$) for numerical stability.

3. Results

In this study, to determine the best RNN model, 1000 different RNNs with LSTM units for both classification problems (4-class and 2-class) were developed in MATLAB environment (MATLAB and Neural Network Toolbox Release 2018b, The MathWorks, Inc., Natick, MA, USA). For the first RNN, all tested architectures had the input sequence layer with a size of 11 (the number of features selected by mRMR algorithm) and a fully connected layer of 4 units (the number of sleep stages to classify). For the second RNN, all tested architectures had the input sequence layer with a size of 27 (the number of principal components) and a fully connected layer of 2 units (for binary classification). The other parameters (i.e. the number of LSTM layers and LSTM units for each layer) were different amongst the networks. In addition, only for the second RNN, we tried to use different classification thresholds applied to the output of the softmax layer to determine whether values different from the standard probability threshold (i.e. 0.5) could increase the N1 stage accuracy, keeping almost unchanged the number of correctly classified epochs of REM sleep. The performance of the recurrent classifiers was assessed by computing the percentage of correct classification (PCC), i.e. the number of correctly classified epochs in all classes normalized by the total number of epochs in the dataset, and the per-class sensitivity (Se), specificity (Sp) and accuracy (Acc). The two best RNN structures were identified according to the following selection process. For the problem of 4-class classification, the RNN model with the best N1-REM sensitivity was selected as the first LSTM RNN, while, for the purpose of binary classification, the network with highest N1 true positives (and at the same time with an acceptable value for REM stage) was selected as the second LSTM RNN. The best RNN model performing 4-class classification had one LSTM layer with 101 hidden units, while the best RNN model for 2-class classification had two LSTM layers: the first was a sequence-to-sequence architecture with 125 hidden units and the second was a sequence-to-label architecture with 98 hidden units. In addition, for the second RNN the classification threshold was set to 0.6. The detailed information for each layer of the two proposed network models is reported in Tables 3–4.

The entire dataset was split into a training (80% of data), validation (10%) and test (10%) set. A stratified 10-fold cross-validation technique was used to test the performance of two RNNs. Stratification ensures that the class distribution from the whole dataset is preserved during the training, validation and test sets. The entire dataset, summarized in Table 1, was randomly split into ten parts; nine parts of the data (9252 epochs) were used as training (8224 epochs) and validation (1028 epochs) sets and the remining one part (1028 epochs) was used for testing. This process was repeated *ten* times using different parts for training, validation and testing in each case. For the first RNN we used the entire dataset and merged the N1 and REM epochs into a single class (a total of 2780 epochs); subsequently only a portion of these epochs (i.e. epochs which were correctly classified by the first RNN) was used for the performance evaluation of the second RNN. The training set, during backpropagation, was split into smaller training subsets called mini-batches to speed up the optimization algorithm. Mini-batch size of 512 and 256 for the first and the second RNN were used respectively. The validation set was used to stop training automatically when the validation accuracy stopped increasing in order to avoid overfitting [43]. Both the training and validation results for the two RNNs are

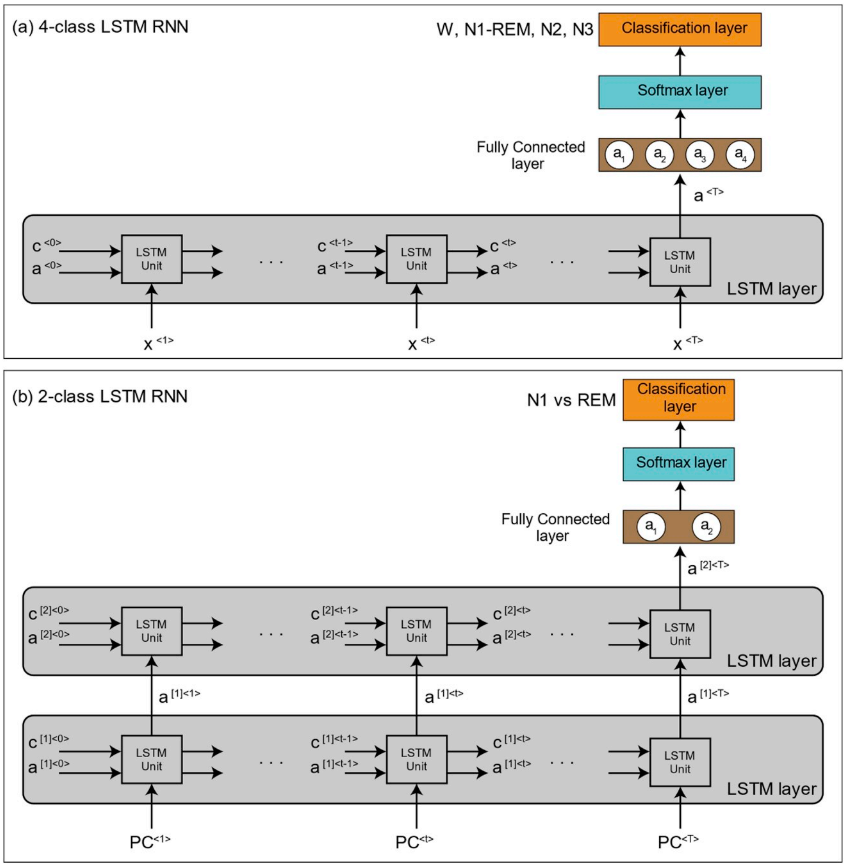


Fig. 5. a) RNN architecture for 4-class classification. b) RNN for 2-class classification.

Table 3
RNN model with LSTM units performing 4-class (W, N1-REM, N2 and N3) classification.

Layer number	Layer type	Properties
Layer 1	Sequence input layer	11 input features
Layer 2	LSTM layer	101 hidden units, sequence-to-label architecture
Layer 3	Fully connected layer	4 units
Layer 4	Softmax layer	Softmax activation function
Layer 5	Classification output layer	4 classes

Table 4
RNN model with LSTM units performing 2-class (N1 vs REM) classification.

Layer number	Layer type	Properties
Layer 1	Sequence input layer	27 principal components
Layer 2	LSTM layer	125 hidden units, sequence-to-sequence architecture
Layer 3	LSTM layer	98 hidden units, sequence-to-label architecture
Layer 4	Fully connected layer	2 units
Layer 5	Softmax layer	Softmax activation function (threshold = 0.6)
Layer 6	Classification output layer	2 classes

reported in Fig. 6. The solid line is the average accuracy of the 10 folds, while the shaded area and the error bars indicate the standard deviation for the training and validation set, respectively.

Both feature selection and reduction algorithms were executed only on the training and validation data and subsequently the relationships between new data and original features were applied to the examples in the test set in each of ten folds. The processing was performed on a workstation with a 2.5 GHz quad-core CPU, 16 GB of memory RAM and 64-bit version of Windows. The final confusion matrices are the sum of all confusion matrices of each fold, for the sleep stage classification using the two LSTM RNNs (shown in Tables 5–6).

It can be seen from Tables 5–6, that the first RNN with LSTM units was able to classify 4 classes with a PCC of 90.80%. The highest classification rate was obtained in stage W. The second RNN with two LSTM layers, performing 2-class classification, obtained a PCC of 83.56%. In Table 5, some epochs of the fused stage N1-REM and N3 were misclassified as N2 epochs, while some N2 epochs were misclassified as N1-REM and N3 stages. The other misclassification values were negligible for the 4-class classification. Finally, the overall performances, for each sleep stage, obtained by the cascaded RNN architecture are: 95.29%, 61.09%, 89.48%, 91.66% and 83.76% for stage W, N1, N2, N3 and REM respectively. The overall PCC of the proposed method is 86.74%.

4. Discussion

Sleep scoring is a difficult and time-consuming task performed manually by sleep experts. The objective of this work is to propose a novel automated sleep stage classification method. There is an extensive literature on automated scoring of sleep stages. Table 7 presents the comparison of our proposed approach with several research studies, both multi-channel and single-channel signal based, performed on five-class classification following the AASM rules. The results for each sleep stage and the overall percentage of correct classification (PCC) are reported.

In few of these methods [8,18,20–22,27,30], the public Sleep-EDF database [34] was used. In addition, few authors employed more than one EEG channel [16,19,22] and other PSG signals [17,24], like the EOG and EMG signals for sleep stage classification. Hsu et al. [27] used

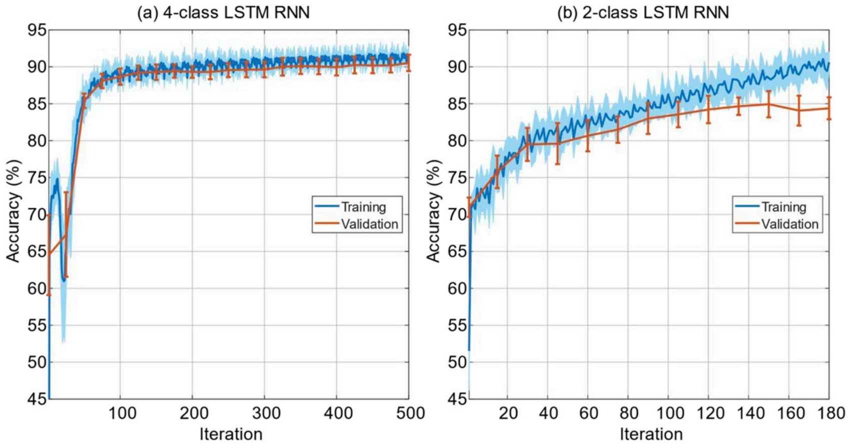


Fig. 6. a) Training and validation accuracy over iterations for the 4-class RNN. b) Training and validation accuracy over iterations for the 2-class RNN.

Table 5
Confusion matrix for 4-class classification using 10-fold cross-validation.

	Predicted				Per-class [%]		
True	W	N1-REM	N2	N3	Se	Sp	Acc
W	810	28	3	9	95.29	99.77	99.40
N1-REM	10	2537	231	2	91.26	95.21	94.14
N2	4	331	4438	187	89.48	93.10	91.35
N3	8	0	133	1549	91.66	97.69	96.70

Percentage of correct classification (PCC) = 90.80%.

Table 6
Confusion matrix for 2-class classification using 10-fold cross-validation.

	Predicted		Per-class [%]		
True	N1	REM	Se	Sp	Acc
N1	562	202	73.56	87.87	83.56
REM	215	1558	87.87	73.56	83.56

Percentage of correct classification (PCC) = 83.56%.

Elman RNN to automatically classify sleep stages employing energy features extracted from the EEG signal of the Fpz-Oz channel and reported the classification rate of 87.20%. Haung et al. [16] proposed a sleep classification system using two EEG channels (Fp1 and Fp2), based on fuzzy *C*-means (FCM) for dimension reduction and multi-class SVM for classification. They obtained an average accuracy of 70.92%. Hassan et al. proposed two methods for sleep stage classification: in the first work [20], they extracted statistical and spectral features from a single EEG channel (Pz-Oz) and applied statistical analysis to validate the selection of features. Subsequently, they used bootstrap aggregating (Bagging) to perform classification of different sleep stages and reported an accuracy of 86.53%. In the second work [21], they applied complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) to decompose the single-channel (Pz-Oz) EEG signal into intrinsic mode functions (IMFs). Subsequently they extracted higher order statistical moments and employed bagging to classify sleep stages and reported a classification rate of 90.69%. A novel method [44] used a multiresolution approach to decompose the RR-time series into IMFs with the aim of classifying “sleep vs wake”, “light sleep vs deep sleep” and “REM vs NREM” sleep stages. Rodriguez-Sotelo et al. [22] proposed a novel automated sleep classification method using entropy features fed to an unsupervised feature classification algorithm (J-means clustering). They used two EEG channels (Fpz-Cz and Pz-Oz) and implemented a new feature selection method called Q-α relevance

Table 7
Comparison of multi-channel and single-channel state-of-the-art studies.

Multi-channel methods		Results [%]					
Authors	Classifier	W	N1	N2	N3	REM	PCC
Huang et al. [16]	Multi-class SVM	76.85	34.24	72.39	88.67	69.07	70.92
Rodriguez-Sotelo et al. [22]	J-means clustering	84.00	15.00	91.00	59.00	38.00	81.00
Shuyuan et al. [24]	K-means clustering	76.14	11.76	69.94	97.12	94.44	74.70
Lajnef et al. [17]	Dendrogram SVM	90.00	41.00	70.00	76.00	97.00	76.20
Acharya et al. [19]	GMM	87.13	94.02	85.24	82.83	98.34	88.71
Single-channel methods		Results [%]					
Authors	Classifier	W	N1	N2	N3	REM	PCC
Hsu et al. [27]	Elman RNN	70.80	36.70	97.30	89.70	89.50	87.20
Hassan et al. [20]	Bagging	96.60	27.48	82.93	76.92	69.57	86.53
Hassan et al. [21]	Bagging	95.28	47.02	92.38	90.00	80.87	90.69
Fraiwan et al. [23]	Random forest	93.33	43.22	84.76	68.37	76.41	82.57
Seifpour et al. [18]	Multi-class SVM	98.76	40.07	90.94	85.08	83.98	91.82
Sors et al. [29]	CNN	91.40	34.92	89.24	85.08	85.82	86.79
Wei et al. [30]	CNN	92.70	26.66	87.40	87.05	82.74	83.93
Sharma et al. [8]	Multi-class SVM	95.41	17.39	76.38	57.11	36.46	83.92
Proposed method	LSTM RNN	95.29	61.09	89.48	91.66	83.76	86.74

analysis. They obtained an average accuracy of 81.00%. Fraiwan et al. [23] employed time-frequency analysis and Renyi entropy for feature extraction from a single EEG channel (C3-A1). They obtained a classification rate of 82.57% using random forest classifier. Shuyuan et al. [24] extracted both time- and frequency-domain features from PSG signals (4 EEG channels, 2 EOG channels and EMG) and employed an improved K-means clustering algorithm for the classification of sleep stages with a classification rate of 74.70%. Lajnef et al. [17] proposed a multi-class SVM classifier based on a decision tree approach. They extracted features from PSG signals (2 EEG channels, 2 EOG channels and EMG) and used forward sequential selection technique for feature selection, obtaining a classification rate of 76.20%. Seifpour et al. [18] extracted a novel variable called statistical behavior of local extrema (SBLE), using a single EEG channel (Fpz-Cz). A multi-class feature selection method was employed, and the selected features were used as input to the SVM classifier, obtaining a classification rate of 91.82%. Sors et el [29]. presented a deep CNN using single-channel raw EEG signals for supervised learning of sleep stage prediction and obtained a classification rate of 86.79%. Wei et al. [30] computed the Hilbert-Huang transform and temporal feature matrix from single-channel (Fpz-Cz) EEG time series. They fed these features to the CNN and reported an average accuracy of 83.93%. Acharya et al. [19] used higher order spectra (HOS) to extract hidden information in the EEG signals using two channels (C4-A1 and C3-A2). These nonlinear features were fed to a GMM classifier for automatic identification of sleep stages. They reported a classification rate of 88.71%. Sharma et al. [8] extracted a novel set of wavelet-based features which were fed to the SVM classifier and they obtained an average accuracy of 83.92%.

In this study, we applied EEG signals to the cascaded RNN architecture with LSTM units with the aim of classifying sleep stages automatically. To the best of our knowledge, it is the first study that uses LSTM units, applied to the sleep stage classification. In this process was involved only a single EEG channel, which solves the problems typical of multi-channel signal base methods, such as the limitations on subject's movements and the low signal-to-noise ratio due to multi-electrode interference. The sleep EEG signals are nonlinear and non-stationary in nature, thus we extracted time and frequency-domain features from 1 s EEG time-segment, repeating this process 30 times in order to cover the whole 30 s epoch and we exploited the nonlinear modelling capability of the RNN. We extracted hand-crafted features in order to keep a strict physiological meaning with the nature of EEG signals. In a future work, CNNs can be employed to extract deep features from the EEG time-frequency distributions and the advantages and limitations of these two different approaches can be investigated. After the feature extraction process, a subset of feature was selected using the mRMR algorithm. This subset included only features that had low correlation with each other and were highly correlated with the target. These features were amplitude, energy and relative power of beta sub-band, which was typical for stage W, the amplitude of sigma sub-band which belonged to stage N2 sleep and the amplitude and relative power of theta sub-band were typical for stage N1, N2 and REM. In addition, the Hjorth mobility can be considered to discriminate stage N3 due to the low-frequency waves (delta waves) present in this stage. The first-time derivative acts as a high-pass filter, attenuating these frequency components. Hence the variance of the derivative of the EEG signal in stage N3 is lower than in other sleep stages. For the RNN performing 2-class (N1 vs REM) classification, the dimensionality reduction process selected relatively high number of relevant principal components due to the similarity of these two stages in terms of amplitude and spectrum variations. It is evident from Fig. 3 that, the three principal components (that are more correlated with the target) show the highest coefficients of the linear combination in correspondence of feature 8, 9 and 13 for the first component, feature 54 for the second component and feature 43 for the third component. These features, as can be seen in Table 2, are the RMS and ARV, the amplitude of delta sub-band, the product of relative powers in low-frequency sub-bands (i.e. delta and theta) and relative power in delta sub-band normalized to the power in gamma sub-band. This relation between principal components and features computed in delta sub-band can be justified by the presence of low-frequency waves, like the vertex sharp waves, in stage N1, instead of sawtooth waves, typical of REM sleep, because they are rarely present in our dataset. As shown in the results reported in Tables 5–6, the algorithm presents the difficulty in discriminating between N1 and REM sleep stages. This is understandable because these stages are characterized by similar EEG activity and the recognition is made even harder by lack of EOG and EMG signals. Charbonnier et al. [25] showed an increase of 20% classification accuracy in stage N1, when EOG signal was added to the analysis, since during REM sleep, rapid eye movements were observed. In addition, the classification accuracy of REM sleep epochs was highly improved when EMG signal was added to EEG, because the muscle activity is lower in this stage with respect to stage W and N1. In stage N1, there are no regularly repeating patterns, in stage N2 the aim is to detect the presence of specific waveforms by using time or frequency-domain features. Other possible explanation for the misclassification of stage N1 sleep, could be the inconsistency between sleep experts in classifying this stage which is a transition from stage W to deep sleep. In fact, stage N1 sleep lasts only 2–5% of the total duration in a standard sleep cycle [45]. The percentages of the other sleep stages with respect to the total sleep episode are 45–55% for stage N2, 15–25% for stage N3 and 20–25% for REM sleep. In this work we used many epochs for each class to reproduce these sleep cycle percentages. But the number of N1 epochs can be increased to improve the results of our approach. In future, time-frequency and/or complexity-domain features [46] can be used.

Table 7 indicates that our method outperforms the state-of-the-art methods, which used single-channel EEG signals, in the N1 stage detection. In addition, this work reports the highest number of correctly classified values for stage N3 (compared to other single-channel methods) and obtains a percentage greater than 83% in the other sleep stages. In addition, impressive results for stage W are obtained, using a small number of training epochs compared to other studies. The performance for stage N1 are higher with respect to multi-channel signal based methods except for the work of Acharya et al. [19] which showed excellent results using a private clinical database.

The limitation of this work is the error propagation in the cascaded RNN architecture, which becomes relevant especially in the REM stage detection. The disadvantage of using deep neural networks, respect to other classifiers is the computation cost due to the training process. This study proved that, the RNN is capable to handle EEG time series efficiently and hence this model can be used to evaluate other biomedical time signals. Recently, few authors [47–49] used LSTM RNN to classify the cardiac abnormalities using electrocardiographic (ECG) signals with high classification accuracy.

The next step towards the automated assessment of neurocognitive performance (NCP) should be the application of the RNNs to the classification of the hypnagogic EEG stages. In this context, the variable called sleep onset period (SOP), i.e. the period interposed between weak wakefulness and drowsiness, becomes very important. According to this definition, the SOP is centered around stage N1, but it clearly overlaps into sleep stage W and N2 [5]. The most of the hypnagogic EEG stages are classified as stage N1 in the AASM standard, hence the results obtained with the proposed method in the first two sleep stages (stage W and stage N1) encourage us to explore the possibility of NCP assessment.

5. Conclusion

In this work a novel approach based on LSTM networks, is developed for automated sleep stage classification using single-channel EEG signals. The EEG signal analysis can be divided into *five* essential parts: data acquisition, signal pre-processing, feature extraction, feature selection or dimensionality reduction and classification. The most relevant features are extracted from the signals and are used as input for the recurrent neural classifier. Two RNN models are proposed; the first performed multi-class classification by merging into a single class the stage N1 and REM, while the second performed the binary classification (N1 vs REM). The results obtained with the cascaded RNN architecture, proposed in this paper, are encouraging for automated NCP assessment.

In future works, the raw EEG signals can be fed to a cascaded convolutional-recurrent neural network architecture. The CNN model helps to extract deep features without the feature selection and the RNN performs classification. In addition, spatial and temporal feature extraction can be achieved by the CNN and RNN layers respectively.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

The Authors would like to express their gratitude to Prof. Alessandro Cicolin, Head of the Sleep Clinic of the “Città della Salute e della Scienza di Torino” hospital (Torino, Italy) for the discussion about the neurocognitive performance. Also, the Authors thank Dr. Silvia Seoni, Dept. of Electronics and Telecommunications, Politecnico di Torino (Torino, Italy) for having helped in the interpretation of the results.

References

[1] A. Gevins, Non-invasive human neurocognitive performance capability testing method and system, U.S. Pat. (1994). https://patents.google.com/patent/US5295491A/en (accessed February 24, 2018).

[2] C.F. Reynolds, R. O'Hara, DSM-5 sleep-wake disorders classification: overview for use in clinical practice, Am. J. Psychiatry 170 (2013) 1099–1101, https://doi.org/10.1176/appi.ajp.2013.13010058.

[3] N. Goel, H. Rao, J.S. Durmer, D.F. Dinges, Neurocognitive consequences of sleep deprivation, Semin. Neurol. 29 (2009) 320–339, https://doi.org/10.1055/s-0029-1237117.

[4] A. Garcés Correa, L. Orosco, E. Laciar, Automatic detection of drowsiness in EEG records based on multimodal analysis, Med. Eng. Phys. 36 (2014) 244–249 https://doi.org/10.1016/j.medengphy.2013.07.011.

[5] H. Tanaka, M. Hayashi, T. Hori, Topographical characteristics and principal component structure of the hypnagogic EEG, Sleep 20 (1997) 523–534, https://doi.org/10.1093/sleep/20.7.523.

[6] T. Hori, Y. Sugita, E. Koga, S. Shirakawa, K. Inoue, S. Uchida, H. Kuwahara, M. Kousaka, T. Kobayashi, Y. Tsuji, M. Terashima, K. Fukuda, N. Fukuda, Proposed supplements and amendments to “A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects”, the rechtschaffen & Kales (1968) standard, Psychiatr. Clin. Neurosci. 55 (2001) 305–310, https://doi.org/10.1046/j.1440-1819.2001.00810.x.

[7] R.S. Rosenberg, S. Van Hout, The American Academy of sleep medicine inter-scorer reliability program: sleep stage scoring, J. Clin. Sleep Med. 9 (2013) 81–87, https://doi.org/10.5664/jcsm.2350.

[8] M. Sharma, D. Goyal, P. V Achuth, U.R. Acharya, An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank, Comput. Biol. Med. 98 (2018) 58–75 https://doi.org/10.1016/j.compbiomed.2018.04.025.

[9] U.R. Acharya, O. Faust, N. Kannathal, T. Chua, S. Laxminarayan, Non-linear analysis of EEG signals at various sleep stages, Comput. Methods Progr. Biomed. 80 (2005) 37–45 https://doi.org/10.1016/j.cmpb.2005.06.011.

[10] R. Boostani, F. Karimzadeh, M. Nami, A comparative review on sleep stage classification methods in patients and healthy individuals, Comput. Methods Progr. Biomed. 140 (2017) 77–91 https://doi.org/10.1016/j.cmpb.2016.12.004.

[11] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recogn. Lett. 15 (1994) 1119–1125 https://doi.org/10.1016/0167-8655(94)90127-9.

[12] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238, https://doi.org/10.1109/TPAMI.2005.159.

[13] I. Kononenko, Estimating attributes: analysis and extensions of RELIEF, Lect. Notes Comput. Sci, Springer Berlin Heidelberg, 1994, pp. 171–182 https://doi.org/10.1007/3-540-57868-4_57.

[14] I. Jolliffe, M. Lovric (Ed.), Principal Component Analysis BT - International Encyclopedia of Statistical Science, Springer Berlin Heidelberg, 2011, pp. 1094–1096, , https://doi.org/10.1007/978-3-642-04898-2_455.

[15] Q. Du, N.H. Younan, I. Lovrek, R.J. Howlett, L.C. Jain (Eds.), Dimensionality Reduction and Linear Discriminant Analysis for Hyperspectral Image Classification BT - Knowledge-Based Intelligent Information and Engineering Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 392–399.

[16] C.S. Huang, C.L. Lin, W.Y. Yang, L.W. Ko, S.Y. Liu, C.T. Lin, Applying the fuzzy c-means based dimension reduction to improve the sleep classification system, 2013 IEEE Int. Conf. Fuzzy Syst, 2013, pp. 1–5, , https://doi.org/10.1109/FUZZ-IEEE.2013.6622495.

[17] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, K. Jerbi, Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines, J. Neurosci. Methods 250 (2015) 94–105 https://doi.org/10.1016/j.jneumeth.2015.01.022.

[18] S. Seifpour, H. Niknazar, M. Mikaeili, A.M. Nasrabadi, A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal, Expert Syst. Appl. 104 (2018) 277–293 https://doi.org/10.1016/j.eswa.2018.03.020.

[19] u.r. Acharya, e.c.-p. Chua, k.c. Chua, l.i.m.c. Min, t. Tamura, Analysis and automatic identification of sleep stages using higher order spectra, Int. J. Neural Syst. 20 (2010) 509–521, https://doi.org/10.1142/s0129065710002589.

[20] A.R. Hassan, S.K. Bashar, M.I.H. Bhuiyan, On the classification of sleep states by means of statistical and spectral features from single channel Electroencephalogram, 2015 Int. Conf. Adv. Comput. Commun. Informatics, 2015, pp. 2238–2243, , https://doi.org/10.1109/ICACCI.2015.7275950.

[21] A.R. Hassan, M.I.H. Bhuiyan, Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating, Biomed. Signal Process. Control 24 (2016) 1–10 https://doi.org/10.1016/j.bspc.2015.09.002.

[22] J.L. Rodríguez-Sotelo, A. Osorio-Forero, A. Jiménez-Rodríguez, D. Cuesta-Frau, E. Cirugeda-Roldán, D. Peluffo, Automatic sleep stages classification using EEG entropy features and unsupervised pattern analysis techniques, Entropy 16 (2014) 6573–6589, https://doi.org/10.3390/e16126573.

[23] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, H. Dickhaus, Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier, Comput. Methods Progr. Biomed. 108 (2012) 10–19 https://doi.org/10.1016/j.cmpb.2011.11.005.

[24] X. Shuyuan, W. Bei, Z. Jian, Z. Qunfeng, Z. Junzhong, M. Nakamura, An improved K-means clustering algorithm for sleep stages classification, 2015 54th Annu. Conf. Soc. Instrum. Control Eng. Japan, 2015, pp. 1222–1227, , https://doi.org/10.1109/SICE.2015.7285326.

[25] S. Charbonnier, L. Zoubek, S. Lesecq, F. Chapotot, Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging, Comput. Biol. Med. 41 (2011) 380–389 https://doi.org/10.1016/j.compbiomed.2011.04.001.

[26] M.E. Tagluk, N. Sezgin, M. Akin, Estimation of sleep stages by an artificial neural network employing EEG, EMG and EOG, J. Med. Syst. 34 (2010) 717–725, https://doi.org/10.1007/s10916-009-9286-5.

[27] Y.-L. Hsu, Y.-T. Yang, J.-S. Wang, C.-Y. Hsu, Automatic sleep stage recurrent neural classifier using energy features of EEG signals, Neurocomputing 104 (2013) 105–114 https://doi.org/10.1016/j.neucom.2012.11.003.

[28] N.F. Güler, E.D. Übeyli, İ. Güler, Recurrent neural networks employing Lyapunov exponents for EEG signals classification, Expert Syst. Appl. 29 (2005) 506–514 https://doi.org/10.1016/j.eswa.2005.04.011.

[29] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, J.-F. Payen, A convolutional neural network for sleep stage scoring from raw single-channel EEG, Biomed. Signal Process. Control 42 (2018) 107–114 https://doi.org/10.1016/j.bspc.2017.12.001.

[30] L. Wei, Y. Lin, J. Wang, Y. Ma, Time-frequency convolutional neural network for automatic sleep stage classification based on single-channel EEG, 2017 IEEE 29th Int. Conf. Tools with Artif. Intell, 2017, pp. 88–95, , https://doi.org/10.1109/ICTAI.2017.00025.

[31] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–80 http://www.ncbi.nlm.nih.gov/pubmed/9377276 , Accessed date: 18 July 2018.

[32] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, ArXiv Prepr. ArXiv1412.3555 http://arxiv.org/abs/1412.3555, (2014) , Accessed date: 18 July 2018.

[33] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (1997) 2673–2681, https://doi.org/10.1109/78.650093.

[34] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet, Circulation 101 (2000) e215–e220, https://doi.org/10.1161/01.CIR.101.23.e215.

[35] B. Hjorth, EEG analysis based on time domain properties, Electroencephalogr. Clin. Neurophysiol. 29 (1970) 306–310 https://doi.org/10.1016/0013-4694(70)90143-4.

[36] 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 Electroacoust. 15 (1967) 70–73, https://doi.org/10.1109/TAU.1967.1161901.

[37] N. Kannathal, M.L. Choo, U.R. Acharya, P.K. Sadasivan, Entropies for detection of epilepsy in EEG, Comput. Methods Progr. Biomed. 80 (2005) 187–194 https://doi.org/10.1016/j.cmpb.2005.06.012.

[38] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley-Interscience, 2006, https://www.wiley.com/en-it/Elements+of+Information+Theory,+2nd+Edition-p-9780471241959 , Accessed date: 22 July 2018.

[39] Y. Goldberg, Neural network methods for natural language processing, Synth. Lect. Hum. Lang. Technol. 10 (2017) 1–309, https://doi.org/10.2200/S00762ED1V01Y201703HLT037.

[40] A. Graves, Generating sequences with recurrent neural networks, Eprint ArXiv:1308.0850 http://arxiv.org/abs/1308.0850, (2013) , Accessed date: 25 July 2018.

[41] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, Int. J. Uncertain. Fuzziness Knowledge-Based Syst. 06 (1998) 107–116, https://doi.org/10.1142/S0218488598000094.

[42] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, ArXiv Prepr. ArXiv1412.6980 http://arxiv.org/abs/1412.6980, (2014) , Accessed date: 26 July 2018.

[43] I.A. Basheer, M. Hajmeer, Artificial neural networks: fundamentals, computing, design, and application, J. Microbiol. Methods 43 (2000) 3–31 https://doi.org/10.1016/S0167-7012(00)00201-3.

[44] R.K. Tripathy, U.R. Acharya, Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework, Biocybern. Biomed. Eng. 38 (2018) 890–902 https://doi.org/10.1016/j.bbe.2018.05.005.

[45] H.R. Colten, B.M. Altevogt, Sleep Disorders and Sleep Deprivation: an Unmet Public Health Problem, National Academies Press (US), 2006, https://doi.org/10.17226/11617.

[46] U.R. Acharya, S. Bhat, O. Faust, H. Adeli, E.C.-P. Chua, W.J.E. Lim, J.E.W. Koh, Nonlinear dynamics measures for automated EEG-based sleep stage detection, Eur. Neurol. 74 (2015) 268–287, https://doi.org/10.1159/000441975.

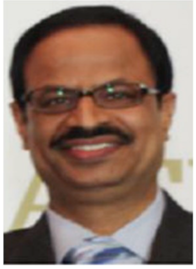
[47] Ö. Yildirim, A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification, Comput. Biol. Med. 96 (2018) 189–202 https://doi.org/10.1016/j.compbiomed.2018.03.016.

[48] J.H. Tan, Y. Hagiwara, W. Pang, I. Lim, S.L. Oh, M. Adam, R.S. Tan, M. Chen, U.R. Acharya, Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals, Comput. Biol. Med. 94 (2018) 19–26 https://doi.org/10.1016/j.compbiomed.2017.12.023.

[49] O. Faust, A. Shenfield, M. Kareem, T.R. San, H. Fujita, U.R. Acharya, Automated detection of atrial fibrillation using long short-term memory network with RR interval signals, Comput. Biol. Med. 102 (2018) 327–335 https://doi.org/10.1016/j.compbiomed.2018.07.001.

Nicola Michielli, PhD student, received his Master's Degree in Biomedical Engineering from the Politecnico di Torino, Torino, Italy, with the thesis “Microwave imaging of biological tissues: a multiresolution approach”. The activity was conducted at the Advanced Computing and Electromagnetics research area, Istituto Superiore Mario Boella (ISMB), Torino, Italy. He is currently a PhD student in Bioengineering and Medical-Surgical sciences at the Department of Electronics and Telecommunications of Politecnico di Torino, Italy. His research is mainly focused on mathematical models for applications to medical imaging and biomedical signal processing.

Filippo Molinari, PhD, DEng is Full Professor in Biomedical Engineering on faculty of the Dept. of Electronics and Telecommunications of the Politecnico di Torino, Torino, Italy. His main research interests include biomedical signal processing, medical imaging, ultrasound technologies, and non-invasive assessment of cerebral functions and autoregulation. Prof. Molinari is on the Editorial Board of several Journals in the field of bioengineering and currently Editor-in-Chief of the Journal “Computer Methods and Programs in Biomedicine”. Complete profile available at: https://scholar.google.it/citations?user=ttbUYiQAAAAJ&hl=it



U. R. Acharya, Ph.D., DEng is a senior faculty member at Ngee Ann Polytechnic, Singapore. He is also (i) Adjunct Professor at Taylor's University, Malaysia, (ii) Adjunct Faculty at Singapore Institute of Technology- University of Glasgow, Singapore, and (iii) Associate faculty at Singapore University of Social Sciences, Singapore. He received his Ph.D. from National Institute of Technology Karnataka (Surathkal, India) and DEng from Chiba University (Japan). He has published more than 400 papers, in refereed international SCI-IF journals (345), international conference proceedings (42), books (17) with more than 20,000 citations in Google Scholar (with h-index of 73), and ResearchGate RG Score of 47.05. He is ranked in the top 1% of the Highly Cited Researchers for the last *three* consecutive years (2016,2017, and 2018) in Computer Science according to the Essential Science Indicators of Thomson. He has worked on various funded projects, with grants worth more than 2 million SGD. He has *three* patents and in the editorial board of many journals. He has served as guest editor for many journals. His major academic interests are in biomedical signal processing, biomedical imaging, data mining, visualization and biophysics for better healthcare design, delivery and therapy. Please visit https://scholar.google.com.sg/citations?user=8FjY99sAAAAJ&hl=en for more details.