Biomedical Signal Processing and Control 77 (2022) 103751

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Automatic sleep stage classification: From classical machine learning methods to deep learning

Rym Nihel Sekkal [a,*], Fethi Bereksi-Reguig [a,*], Daniel Ruiz-Fernandez [b], Nabil Dib [a], Samira Sekkal [c]

[a] Biomedical Engineering Research Laboratory, Department of Biomedical Engineering, Faculty of Technology, Tlemcen University, Tlemcen 13000, Algeria
[b] Department of Computer Technology, University of Alicante, Carretera San Vicente del Raspeig s/n, San Vicente del Raspeig, Alicante 03690, Spain
[c] Laboratory of Medical Toxicology (Toxicomed), Faculty of Medicine, Tlemcen University, Algeria

ARTICLE INFO

Keywords:
Sleep stage classification
EEG
Data preprocessing
Features selection
Machine learning
LSTM

ABSTRACT

Background and objectives: The classification of sleep stages is a preliminary exam that contributes to the diagnosis of possible sleep disorders. However, it is a tedious and time-consuming task when conducted manually by experts. Many studies explored ways of automating polysomnogram signals analysis. They are based on two main strategies: conventional machine learning and deep learning methods. The objective of this work is to carry out a comparative study on these two classes of models.

Method: A primary comparison of performance of these classifiers is carried out using eight conventional machine learning algorithms and a feed-forward neural networks to assess whether this latter method have definitely supplanted the first. As sleep epochs show inter-epochs correlation, a study of the distinctive influence of this temporal dependence on the classifiers performance is then conducted introducing for this purpose (uni- and bi-directional) long short-term memory networks. In a context of generalization of the use of wearable devices, a comparison of the classification methods examined is also carried out in their accuracy when dealing with a reduced number of channels. Finally, the robustness of the results obtained to the choice of features selection algorithms is discussed.

Results and conclusion: Our results show that support vector machine with radial basis function and random forest are just as valid for predicting sleep stages classification as feature-based neural networks with performance closed to the state of the art. This conclusion remains valid even after the introduction of inter-epochs temporal dependence, reduction of the number of channels or change in features selection method.

1. Introduction

People spend about a third of their life sleeping. During sleep, a person goes through several stages, the succession of which during the night is indicative of the quality of sleep. The consequences of poor sleep quality are numerous: drowsiness, lack of concentration, fatigue, irritability. More than the quality of sleep, analysis of the sequence of sleep stages can reveal sleep disorders. The International Classification of Sleep Disorder identifies seven major categories of sleep disorders [1]. For example, apneic patients (manifesting respiratory arrest at night) have a sleep fragmented by awakenings which allow them to resume their breathing but prevent the deepening of their sleep while narcoleptic patients (who are prone to daytime sleepiness) enter a phase of REM sleep as soon as they fall asleep. Therefore, the recording and classification of the sleep stages may be considered as a preliminary exam, which is important for clinical diagnosis and treatment of sleep disorders.

Sleep examinations are generally performed using a polysomnograph which records several signals including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), oxygen saturation and airflow. These signals are analyzed by human experts who assign a specific sleep stage to each sleep epochs. Epochs are generally evaluated with a resolution of 30 s.

In general, the sleep stages follow the traditional Rechtschaffen and Kales (R&K) rules [2] or the American Academy of Sleep Medicine (AASM) classification [3]. The R&K classification splits the sleep process into 7 discrete stages (Wakefulness state; Non Rapid Eye Movement (NREM) divided into Stage 1 (S1), Stage 2 (S2), Stage 3 (S3) and Stage 4

* Corresponding authors.
E-mail addresses: sekrym@yahoo.fr (R.N. Sekkal), fethi.bereksi-reguig@univ-tlemcen.dz (F. Bereksi-Reguig).

https://doi.org/10.1016/j.bspc.2022.103751
Received 26 October 2021; Received in revised form 30 March 2022; Accepted 27 April 2022
Available online 11 May 2022

(S4); Rapid Eye Movement (REM); and Movement time) [4]. The AASM modified the standard R&K guidelines for sleep classification and developed a new guideline whose major changes is a modification in terminology. The AASM classification divided the sleep process into 5 stages where sleep stages S1 to S4 in the R&K classification are referred to as N1, N2, and N3, with N3, which reflects slow wave sleep, replaces the R&K nomenclature stage 3 and stage 4 sleep.

These stages are characterized by distinct time, frequency and other nonlinear properties. They also differ in their length: stage N1 for example is rare in comparison with other stages. Moreover, the succession over time of the sleep stages is not random, some transitions between stages are allowed while others are not [5].

These characteristics of the sleep process make its classification a complex procedure and its scoring a challenging task. So, during the past two decades, statistical machine learning methods have been introduced in the study of the sleep process exploiting the availability of large amounts of data to learn their complex distributions.

Methods used in this perspective are based on two main strategies: (i) conventional machine learning (ML) methods and (ii) deep learning (DL) methods based on artificial neural networks.

According to their similarity, the classifiers of the first category may be grouped, without being exhaustive, in instance-based algorithms like the k-Nearest Neighbor (kNN) or the Support Vector Machines (SVM); in decision tree algorithms like the Random Forest (RF) or the Decision Tree and in Bayes rule-based classifiers like Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Linear and Quadratic Discriminant Analysis.

For their part, deep learning models are a subset of machine learning based on artificial neural networks. In the simple form of a feed forward neural networks, the deep learning architecture is a succession of layers where each layer computes an output vector using the output of the previous layer and a non-linear activation function. The crucial element however is that the weight parameters appearing in the different layers are considered as being adaptive, so that their values are updated during the training process through a backpropagation algorithm [6,7,8]. Deep learning methods include algorithms like the Multi-Layers Perceptron (MLP) and its modern refinements, including Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTM), Gated Recurrent Unit (GRU), Bi-directional LSTM, Convolutional Neural Networks (CNN).

The application of deep neural networks has grown significantly over the past few years and have recently been successfully applied for automatic sleep stage classification (ASSC) with promising performance. This context led Loh *et al.* [9] to argue that, in the foreseeable future, all studies on the topic of ASSC would either include deep learning or at least mention deep learning-based techniques as a 'referent point'. In the same vein, Qian *et al.* [10] predict that the deep neural networks will be the main research method of automatic arousal detection in the future.

Has deep learning effectively superseded traditional computer sleep stage classification?

In fact, few works have explicitly carried out a systematic comparison of the performance of these two classes of models in the context of automatic sleep stage classification. Among the authors who addressed this issue are Biswal *et al.* [11] who compared different learning methods trained on a very large sleep physiology database. They conduct evaluation using not only Logistic Regression and Tree Boosting but also MLP, CNN, RNN and a combination CNN-RNN method. These authors found that multilayer RNN with expert-defined features has the best performance with an accuracy of 85.8% and that in general, deep learning models outperform the traditional classification methods.

Evaluating data from 62 people, Dong *et al.* [12] used a mixed neural networks concatenating a multi-layer perceptron and a recurrent neural networks to address the sleep stage classification with a single electrode recording. The authors compared this mixed neural networks with that of SVM, Random Forest and MLP and found that the deep learning model proposed has better overall accuracy and macro F1-score than those of the classifiers control group.

Using data from 25 people in a six-state sleep stage (R&K classification), Şen *et al.* [13] find that a hybrid approach for selection features and random forest classifier provides the best accuracy (96.4%) compared to four other classifiers including feed-forward neural networks.

Boostani *et al.* [14] compared the most suitable methods in terms of preprocessing, feature extraction and selection of the sleep stage classifier. They conclude that entropy of wavelet coefficients along with random forest classifier are the best feature and classifier respectively, among a group of five classifiers, with an accuracy value of 88.7%.

A time–frequency spectra of several consecutive 30 s time points as an input to perform the sleep stage classification is used by Xu *et al.* [15] to carry out a comparison between LSTM networks and four classical convolutional neural networks (CNN). They observe that the LSTM model had a better classification performance than the CNNs and may reach a classification accuracy of 87.4%.

However, as can be noted, most of these prior studies either compare models while remaining within one of the two classes of machine learning methods mentioned above or compare one particular model with a control group of classifiers. Thus, although these studies contribute to improving automatic classification and prediction of sleep stages, there is still a lack of a comprehensive comparison between the performance of deep learning models and those based on conventional machine learning approach.

In order to fill this gap, the aim of this work is to see how conventional ML methods behave in relation to methods based on DL and to assess whether the second class of models has definitely supplanted the first for automatic sleep stage classification.

In this regard, the contribution of our paper is threefold. First and foremost, in addition to providing a review of the range of machine learning methods that have been used for the detection of sleep stages, the paper presents, in contrast to other related studies, a systematic comparison between conventional ML methods and featured based deep learning models used in the context of the ASSC.

Second, throughout this evaluation, this paper discusses the detail of the processing techniques required for the analysis of sleep stages, which includes pre-processing, features extraction, features selection and classification with attention to the time dependency of the polysomnographic signals.

Third, our paper complements the performance measures generally used for model comparison such as accuracy or F1-Score by the introduction of robustness tests that provides information about the sensitivity of these observed performances - and the ranking of the models they induce - to the introduction of the temporal correlations, the number of signals considered and the nature of the features selected.

The organization of this paper is structured around these challenges. After data preprocessing and features selection, we first compare the performance of these two categories of classifiers, namely, eight traditional ML algorithms and, as a preliminary matter, a feed forward neural networks using the same database and the same features set.

The successive sleep stages can be seen as a realization of a temporal process and, as such, may exhibit inter-epochs correlation. We then take into consideration this autocorrelation and study in a second experiment the distinctive influence of this temporal context on the performance classifiers, particularly for those based on neural networks.

Given the limitations of the polysomnography (PSG), wearable devices are increasingly being used with probably a loss of precision in the sleep stage classification. It is therefore important to know which of these alternative methods (conventional machine learning versus deep learning) can best mitigate this loss of information. Thus, a comparison of the two groups of classification methods is then carried out in their accuracy when dealing with a reduced number of channels.

Finally, as this work is limited to feature-based algorithms, the comparative robustness of these alternative classification methods with respect to the choice of the features selection model is examined.

2. Material and methods

The method used is mainly based on the following workflow: data acquisition; data denoising; data splitting, balancing and normalizing; feature extraction and selection and finally sleep stage classification and experimentations [14].

2.1. Sleep dataset description

The database used to conduct this study is the publicly available physionet Sleep-EDF Database Expanded (Sleep Cassette) [16 17]. This database contains 153 PSG recorded during two-night periods. The PSG of the first night of the first 19 subjects in the database were used because of the characteristics of the processing machine used in this study. This machine is equipped with an Intel Core i5-8250U CPU@1.60 GHz, 8 GB memory and an Intel UHD Graphics 620 graphics card. Each PSG contains two EEG signals detected at Fpz-Cz and Pz-Cz scalp electrodes locations, one EOG signal with an horizontal placement, one EMG signal detected at the submental area and one oro-nasal respiration signal, with a sampling rate of 100 Hz. The two EEG and the EOG signals of about 10 h each were used from PSG signals so that the raw database, given the sampling rate, is of (3, 19, 8 460 000) dimension.

Signals are segmented in epochs of 30 s duration. Each epoch is scored into one of the eight classes (W, N1, N2, N3, N4, REM, Movement and Unknown) by experts according to the R&K manual. Classes N3 and N4 were merged into a single stage (N3) in accordance with AASM recommendations while Movement and Unknown classes were excluded.

2.2. Data preprocessing

Several artefacts which result from sources other than neuronal activity and which lower the quality of electroencephalography frequently contaminate the EEG signal. Hence, as a first step, it is necessary to eliminate such artefacts to obtain more accurate and appropriate results.

Artefacts can be distinguished by their sources [18]:

- Non-physiological artefacts that may be environmental or experimental such as detachment of the electrode, interference from the electrical networks, patient movements, sweating [19]. Environmental artefacts can often be removed using a simple filter. This is either because they move around a narrow frequency band like the DC artefact or because their frequency band does not overlap with that of the useful signal [20].
- Physiological artefacts that result from the contamination of the EEG signals firstly by the movement of the eyes which appears especially in the frontal electrodes [21,22]. Fig. 1 illustrates an example of such EOG artefact and its incidence on the EEG signals. The EOG artefact is identified by the blue frequency band through a peak in the EOG followed by an abrupt slope.

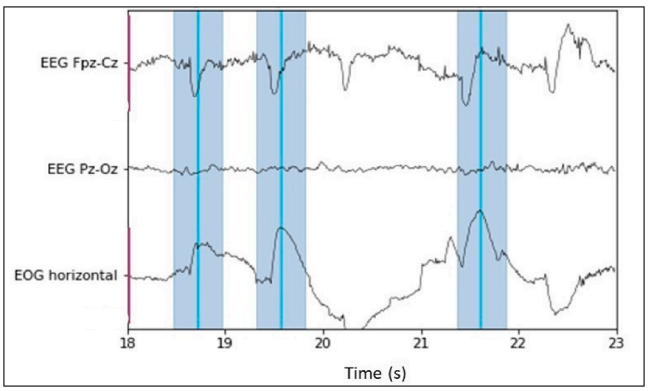

Fig. 1. EOG artefact and its incidence on the Fpz-Cz signal (1st subject).

However, physiological artefacts do not have a specific frequency band and often overlap with the frequency band of useful signals so that simple spectral filtering techniques are not applicable. Although several methods have been proposed to remove these artefacts [23,24], the research on physiological artefact removal continues to be an open problem [22]. In this study, a two-step approach was adopted. In a first step, the EEG signals were filtered in order to eliminate particularly the DC artefacts and very low frequency variations that appear on EEG signals and which are usually caused by breathing or sweating.

The second step is more specific to physiological artefacts, in particular to ocular artefacts. Due to the spectral overlap between neurological and artefactual events in that case, those artefacts are particularly challenging to eliminate. Statistical techniques used to remove this EOG contamination include the blind source separation, especially independent component analysis (ICA) [25] and principal component analysis (PCA) [26], but also, wavelet transform [27].

Among these methods, ICA is perceived to be a robust method for sources separation and is certainly the most commonly used algorithm [22]. However, this procedure requires the identification of the artefactual independent ICA-components. Moreover, it assumes the statistical independence of the sources, the linearity of the mixture and the non-gaussian distribution of the sources (or the gaussian distribution of one source only) [28]. These conditions are difficult to verify especially the third one [22]. If these assumptions are not fulfilled, the contaminated ICA-components do not necessarily contain only artefact data, but also contain underlying EEG data so that removing the contaminated components can lead to a loss of EEG data. At the same time, the other source components, considered as the meaningful part of the EEG data, are not artefact-free and may still be noisy.

In their extensive review on the algorithms used to remove physiological artefact encountered in the EEG, Uriguen *et al.* [18] and Jiang *et al.* [22] concluded that the optimal method for removing artefacts consists on combining more than one algorithm to correct the signals instead of using single methods like ICA. Thus, based on the seminal paper of Castellanos and Makarov [29], a hybrid method was used in this work: instead of eliminating the independent components (ICs) estimated to be contaminated, all the ICs were denoised using a wavelets transformation before reconstructing the EEG signals via the inverse ICA. Following [29], Fig. 2 summarize the process flow for this hybrid method ICA-Wavelet used for EEG artefacts removal.

Steps involved:

1. Get the EEG signals in the database;
2. Filter the EEG signals to eliminate the DC artefacts and the very low frequency variations;
3. Get the IC1 and IC2 sources by ICA separation;
4. Perform denoising of IC1 et IC2 using wavelet transform;
5. Get the denoised EEG signals by ICA reconstitution using the previously denoised components of IC1 and IC2.

2.3. Features extraction and selection method

The identification from the filtered PSG signals of a set of features capable of distinguishing the different sleep classes is a core challenge in classification problems. Two main approaches could potentially address this issue depending on whether features are extracted starting from the knowledge of the experts or directly through deep learning methods such as convolutional neural networks. This study is limited to the first process. In this context, a two-stage process of features representation is generally used: (i) the feature extraction that identifies a large set of candidate features and (ii) the feature selection that reduces the number of features according to a predetermined criterion to prevent overfitting.

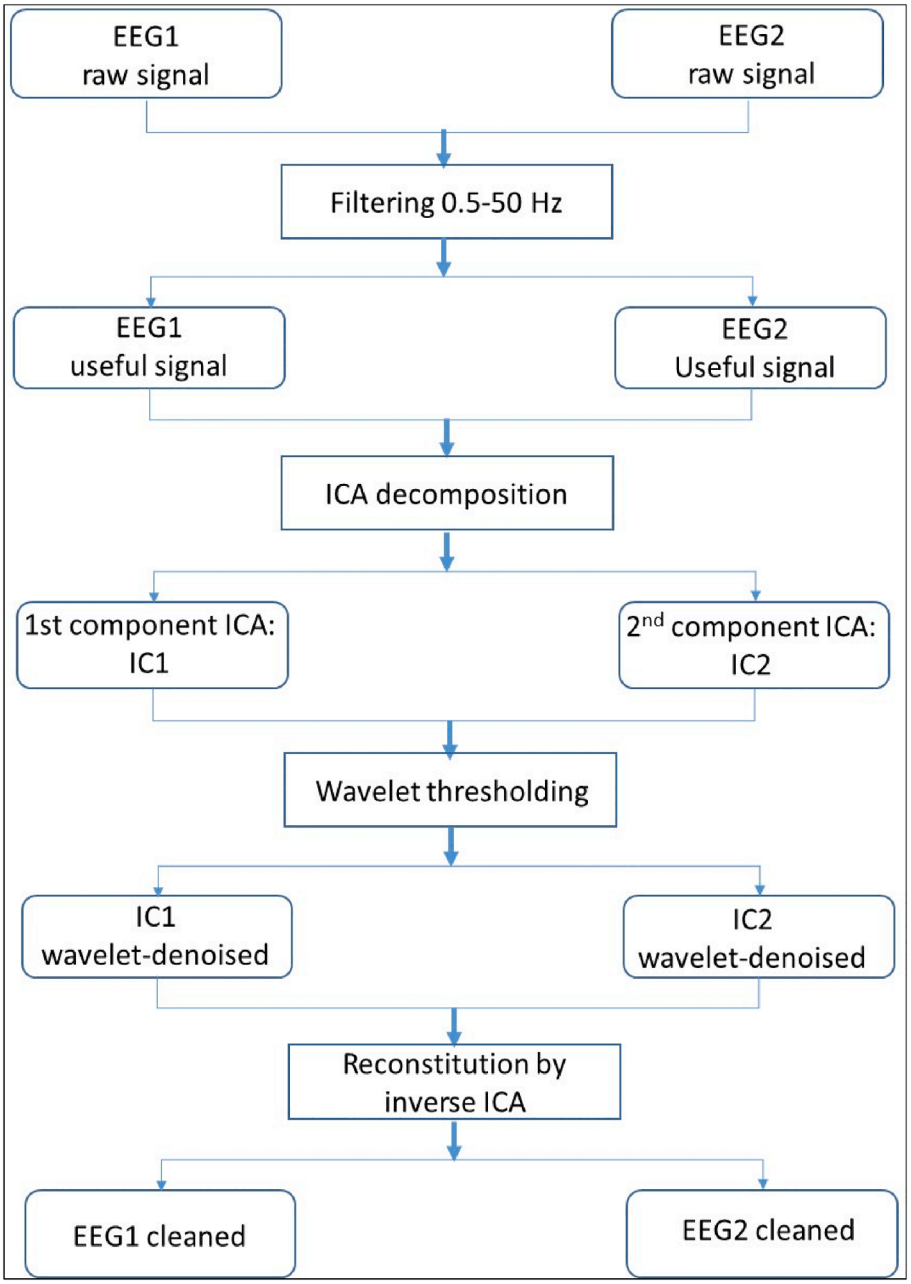

Fig. 2. Artefact removal workflow.

2.3.1. Features extraction

There is some arbitrary in choosing the set of the candidate features. The AASM identifies a set of rules that guide the practitioners. However, the visual process of sleep stage classification remains subjective making it difficult to implement. In any case, these features may be classified broadly into four categories, namely, time, frequency-based, non-linear and entropy categories [13,30]. Contribution of distance-based features to the sleep stage classification was also investigated.

Time domain features. In the time domain, the signals can be seen as a realization of a temporal stochastic process. Therefore, the features that were adopted are the empirical counterpart of:

- the moment of first order (mathematical expectation $\mu = E(X)$),
- the second moment (variance $\sigma^2 = E\left[(X-\mu)^2\right]$),
- the third moment (the skewness which provides information on the degree of asymmetry of the distribution over each epoch $E\left[((X-\mu)/\sigma)^3\right]$),
- the fourth moment (the kurtosis which indicates the degree of 'tailedness' of the distribution and is equal to $E\left[((X-\mu)/\sigma)^4\right]$). The kurtosis recognizes the existence of K-complex and, thus, helps to identify the S2 stage of sleep [31],
- Zero cross rate and zero cross mean which are simple and at the same time very effective features especially in sleep stage classification [30].

The amplitude of the distribution is also included as a candidate feature through the minimum and the maximum amplitude of the EEG signals for each epoch.

Frequency-based features. The power spectral density (PSD) of Welch

was adopted as a frequency attribute. It provides information on the energy distribution of the signal at the different frequencies calculated for each of the two EEGs, for the EOG, for each of the epochs and for each of the frequency bands, namely, delta (0.5 Hz, 4.5 Hz); theta (4.5 Hz, 8.5 Hz); alpha (8.5 Hz, 11.5 Hz); sigma (11.5 Hz, 15.5 Hz) and beta (15.5 Hz, 30 Hz).

Entropy-based features. Entropy indicates the randomness (and hence unpredictability) in the information contained in the signal. Therefore, entropy is an indicator of the system complexity [32]. If an epoch has a high entropy, it has a high probability of being categorized as N1 or REM stages which are phases of transition between awakening and sleep [31 33]. Three indicators of entropy were used: spectral, permutation and sample entropy.

Nonlinear-based features. A signal is less irregular as it is fractal. This is why a low fractal dimension (high irregularity) should be associated with high neuronal activity and should correspond to the S1 or REM sleep stages. The Petrosian Fractal Dimension algorithm [34] based in particular on the number of sign changes of the derivative (slope) of the EEG signal was retained. Two other indicators of signal irregularity were also considered, namely, the Higushi fractal dimension which is a nonlinear measure of waveform complexity in the time domain [35,36] and the Hjorth's mobility and complexity indicators based on the variance of the original signal and its first and second order derivatives [37,38].

Distances-based features. As observed by Ebrahimi *et al.* [39], distance between EEG and EOG may contribute positively to the sleep stage classification. Several distance measures of signal proximity have been proposed. In [40], Gharbali *et al.*, using different ranking features methods, note that Itakura distance appears constantly in the top of the selected features set. This type of distance between EEG and EOG were therefore adopted as a candidate feature.

2.3.2. Features selection

The set of features extracted allows us to compress the information contained in the PSG signals. However, there may be a correlation between these candidate features that leads to information redundancy and increase in the model complexity. This is why a regularization term must be introduced in the loss function and whose role is to cancel the coefficient of features that are not relevant (Lasso method) or penalize those that are too far from zero (Ridge method).

In doing so, the risk of overfitting which results in poor performance is reduced when the estimated model is generalized to predict sleep stages with new data [8]. However, at the same time, the bias of the estimated parameters, that is, the difference between the expectation of the estimated and the true value of the parameters, may increase. This trade-off 'bias-variance' can be illustrated by the Ridge regression and a loss function that adds to the mean square error $\|y - X\beta\|^2$ a regularization term $\|\beta\|^2$ to penalize the features coefficient of the regression distant from zero. The Ridge loss function is then $\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$ where the notation $\|.\|_2$ is for the norm associated to the Hilbert space of square summable sequences and where X is the matrix of the features and y the sleep stages vector. The larger the value of λ, the stronger is the coefficients' size penalized. The minimisation of this function leads to the Ridge regression estimates [41,42]:. $\widehat{\beta} = (X'X + \lambda I)^{-1}(X'y)$

where X' is the transposed matrix of X and I the unit matrix. Incorporating this expression into the bias and the variance of $\widehat{\beta}$, leads to:

$$Bias(\widehat{\beta}) = \mathrm{E}(\widehat{\beta}) - \beta = -\lambda(X'X + \lambda I)^{-1}\beta$$

$$Variance(\beta) = \sigma^2(X'X + \lambda I)^{-1}X'X\left[(X'X + \lambda I)^{-1}\right]'$$

As the penalty parameter becomes larger, the bias (in absolute value) of the Ridge regression coefficient estimates increases while the variance decreases towards zero and vice-versa when λ decreases. These characteristics of the Ridge regression illustrate the well know trade-off bias-variance: the bias and the variance cannot be minimized at the same time and an increase in the bias in order to reduce the variance and the risk of overfitting must be accepted.

The Lasso method (Least Absolute Shrinkage and Selection Operator) differs from the Ridge regression in that it incorporates in the penalty term the l_1-norm instead of the l_2-norm as in Ridge regression. In doing so, for high values of λ, many coefficients of features in the Lasso method are zeroed which cannot be the case in Ridge regression [43,44,45].

A compromise between the Lasso and Ridge regressions is the Elastic Net regression whose regularization term is a convex combination of the penalties from the Ridge and Lasso regressions so that the loss function to minimize is:

$$L(\widehat{\beta}) = (1/2n)\|y - X\beta\|_2^2 + \lambda\left[\alpha\|\widehat{\beta}\|_1^2 + ((1-\alpha)/2)\|\widehat{\beta}\|_2^2\right]$$

where $\|.\|_1$ is the norm associated to the Hilbert space of absolute summable sequences.

It is this hybrid regression that were used to select the PSG features because it generalizes the Ridge and Lasso regressions and is more flexible since it has two hyperparameters λ and α to tune.

However, a limit of these methods is that they apply a regression method with a continuous dependant variable for a classification problem involving a discrete dependant variable, namely the hypnogram. This is why a regularized multinomial logistic regression was also run with a l_2 penalty and which may be a more proper way to handle the classification problems.

The feature selection process can also be linked to an explicit classifier so that the relevance of each features subset is evaluated with its correspondent classification accuracy. These kind of wrapper methods base the 'optimality' of the features subset on the classifier performance. For this reason, an implementation of this method, although computationally expensive compared to other methods, was undertaken with a multi-layer perceptron classifier with two hidden layers.

The performance of these three feature selection methods were compared in the last part of this work.

2.4. Comparison of the classification methods

With the selected features in hand, the sleep stages classification can be carried out as well as experiments for comparing some classical machine learning and deep learning methods. Eight classifiers were used from conventional machine learning methods. On the other hand, deep learning methods included a feed forward neural networks and algorithms with feedback loop, namely unidirectional and bi-directional Long Short-Term Memory Networks (LSTMs).

In the study of automatic sleep stage classification in the deep learning perspective, the LSTM model is of particular importance since it takes into account the temporal correlation between the sleep stages. This model, which was initially describes by Graves and Schmidhuber [46], includes three gates (input, forget, output) and a memory cell in addition to a block input and activation functions.

Following Greff *and al.* [47] the structure of a basic LSTM (vanilla without peephole connections) may be described by the following equations:

$z_t = tanh(W_z x_t + R_z y_{t-1} + b_z)$ bloc input.
$i_t = \sigma(W_i x_t + R_i y_{t-1} + b_i)$ input gate.
$f_t = \sigma(W_f x_t + R_f y_{t-1} + b_f)$ forget gate.
$o_t = \sigma(W_o x_t + R_o y_{t-1} + b_o)$ output gate.
$c_t = z_t \odot i_t + c_{t-1} \odot f_t$ cell memory.
$y_t = \tanh(c_t) \odot o_t$ block output.

where:

- the logistic sigmoid $\sigma(z) = 1/(1+e^{-z})$ and the hyperbolic tangent $\tanh(z) = (e^z - e^{-z})/(e^z + e^{-z})$ are activation functions,

- W and R the weight matrix associated to the input and the lagged output,
- b the bias weight
- x_t the input at current time step, y_t the output at previous step, c_t the cell memory and
- $\odot$ the Hadamard product.

In terms of time dependency, it is worth noting that ouput y_t depends not only on x_{t-1} via c_{t-1} but also on y_{t-1}. We shall return to this point later to classify the sleep stages using an LSTM approach.

2.5. Data splitting and cross-validation features selection

To prevent overfitting that frequently occurs in machine learning, in which case the models may become unable to generalize properly to new data, the test dataset should be independent not only of the training but also of the validation data used to select features or to tune hyper-parameters. This can be accomplished by randomly splitting the entire dataset into two parts using the first subset of epochs for testing and the remaining epochs for training and cross-validation. By this procedure, the testing data is a new dataset that the model had not used nor seen before. Since the dataset is relatively large containing 21,265 samples (Table 1 below), this splitting will not lead to a significant reduction of the number of observations, neither in the train nor in the test data.

The features selection was performed by cross-validation using only the train data subset and, therefore, after the data splitting step. Otherwise, the features selected or the optimized hyperparameters would have already seen the test dataset [48].

3. Results and discussion

3.1. Data preprocessing and features selection results

3.1.1. Denoising

As mentioned above in the workflow (Fig. 2), the first stage is filtering EEG signals which is recommended not only to reduce non physiological artefacts but also to help ICA estimation by increasing the independence between sources [49]. Filtering was performed with a zero-phase finite impulse response filter (FIR) with respectively a lower cut-off frequency and transition bandwidth of 0.5 and 1 Hz and respectively an upper cut-off frequency and transition bandwidth of 47.5 Hz and 5 Hz.

Next, the ICA-Wavelet decomposition step was carried out. After some experiments, the Daubechies wavelet DB4 was chosen for the ICs denoising with a five-level decomposition and a soft thresholding.

Fig. 3 shows the two EEG signals of the first subject corrected for artefacts by the proposed procedure. As expected, removing noise by cancelling the small wavelets coefficients in the ICA components smooth the original signals without affecting their quality.

3.1.2. Balancing datasets

The structure of a data, when it is characterized by an imbalanced classification, is an important learning challenge since it has significant implications for the sleep stage classification [5]. In fact, examination of the PSG signals of the 19 subjects shows that there is a marked predominance of the waking stage at the start and at the end of each recording. For the whole subjects of our database, 68.9% of the epochs correspond to the waking stage. This leads the predictive models to be biased towards the waking stage (the most frequent) while the rarest class (stage N1) tends to be never predicted [50]. For this reason, the majority class was resampled by reducing the wake-up stage before and after sleep to sixty minutes corresponding to 120 periods of 30 s each. This procedure reduces the imbalance in the sleep stage structure of the different PSG signals even though it does not completely remove it (Table 1).

Table 1
Number of 30-s epochs for each sleep stage.

	Sleep Stages					Total Samples
	1	2	3/4	R	W	
Number epochs	1122	8367	2871	3470	5435	21,265
% epochs	5.3%	39.3%	13.5%	16.3%	25.6%	100%

3.1.3. Features selection and standardisation

To prevent overfitting, the entire dataset was split randomly into two parts using 1/4 of all the epochs for testing and the 3/4 remaining dataset for training and cross-validation.

Using only the train dataset, the features were extracted from each of the 30-second segments of the PSG signals. The classification of features into the categories above (time, frequency-based, non-linear, entropy and distance-based) enabled us to extract 62 candidate features from the EEG and EOG signals.

The features were also scaled such that each of them has the same importance with zero mean and unit variance.

Following feature extraction and standardization, the subset of features that avoids the overfitting problem and leads to the most accurate sleep stage classification shall be selected. Regularized regression methods (Elastic Net and multinomial logistic regression) were used with a combination of l_1 and l_2 penalties, a stratified 5-fold cross-validation and a feature selection with recursive feature elimination using the (negative) mean squared error as the performance criterion. As shown in Fig. 4, these regressions select 46 features for the hybrid regression and a subset of 49 features for the multinomial logistic regression since beyond that, there is no improvement in the cross-validation score. The results are not too far apart since 40 features are common to both approaches.

The other wrapper method applied used an artificial neural networks as classifier, namely a MLP with two hidden layers with the training dataset segmented in only 3-folds cross-validation given the computing intensive nature of the method. This classifier was coupled with a search algorithm which, starting with zero feature, selects, at each iteration, the best new feature (the one with the highest cross-validation score). This method -which relies on MLP and a sequentially features selection - Wrapper (MLP-SFS) - provides us, at each iteration, with a series of feature subsets with their respective score from which the one with the highest score can be selected. The optimal features subset obtained using this method includes 43 features (16 attached to the EEG1 signal (Fpz-Cz), 12 to EEG2 (Pz-Oz) and 15 to EOG).

Table 2 presents the three subsets of features resulting from the three considered selection methods (Elastic Net, logistic regression, MLP-SFS) with the indication of the signal to which these features are specifically attached (EEG1, EEG2, EOG). It can be observed that 30 features belong simultaneously to these three subsets of features whereas 48 features belong simultaneously, at least, to two of them. Furthermore, the variance, the zero cross rate and the Hjort mobility appear as robust features since they are selected by the three methods and are attached to the three signals.

Although the MLP-SFS method was the one we mainly used, an experiment using the features set selected by the regression methods was carried out in the last part of this work in order to compare the classification results.

3.2. Evaluating studies and discussion

This section describes the experiments that have been carried out. The aim of these experiments is to:

(i) compare the performance of a selection of conventional classification methods with a deep learning approach;

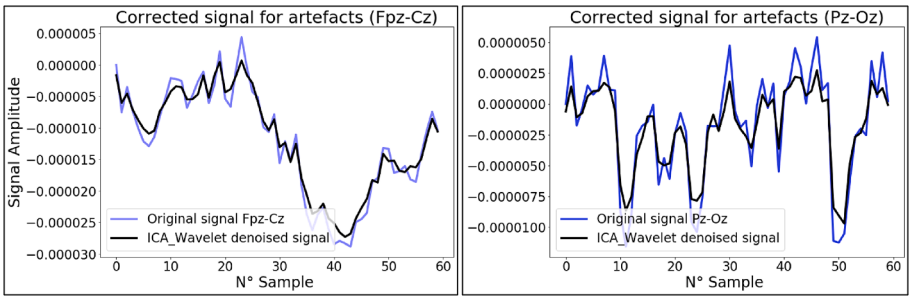

Fig. 3. Correcting the two EEG signals for artefacts (1st subject).

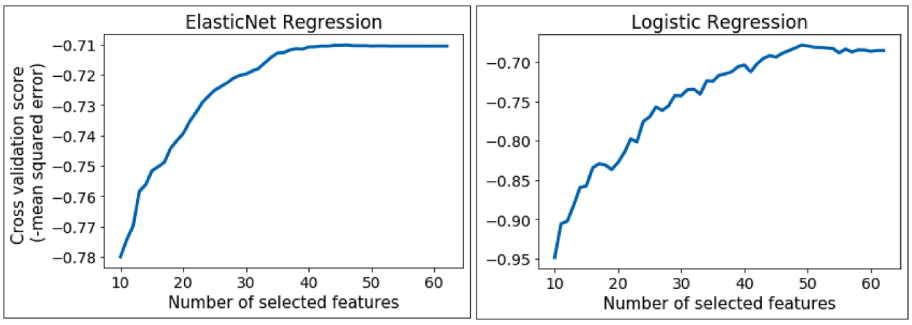

Fig. 4. Cross validation score for regression features-selection methods.

Table 2
Sleep data selected features.

Selection Methods	Elastic Net			Multi. Logistic			Wrapper (MLP-SFS)		
Sources	EEG1	EEG2	EOG	EEG1	EEG2	EOG	EEG1	EEG2	EOG
Minimum value		x	x	x	x	x	x	x	x
Maximum value	x	x	x	x	x	x	x	x	
Arithmetic mean	x		x			x			x
Variance	x	x	x	x	x	x	x	x	x
Skewness	x	x							
Kurtosis	x			x	x		x	x	x
Zero cross rate	x	x	x	x	x	x	x	x	x
Zero cross mean			x	x		x		x	x
Delta band Power spectral density	x	x		x	x		x		x
Theta band Power spectral density	x		x	x	x			x	x
Alpha band Power spectral density		x	x		x	x	x	x	x
Sigma band spectral density	x	x	x	x	x	x	x		
Beta band spectral density		x		x	x	x	x	x	x
Permutation entropy	x		x	x	x	x	x		x
Sample entropy	x	x	x	x	x	x	x		
Spectral entropy	x	x	x	x	x	x	x		x
Petrosian fractal dimension	x	x	x	x	x	x		x	x
Higushi fractal dimension	x		x				x	x	
Hjort mobility	x	x	x	x	x	x	x	x	x
Hjort complexity	x	x	x	x	x	x	x		x
Itakura distance from EOG				x	x		x		
Total	**16**	**14**	**16**	**17**	**17**	**15**	**16**	**12**	**15**

(ii) introduce the temporal structure of the sleep stage process in deep learning models and assess the gain obtained in this context;
(iii) study, for both approaches, the dependence of the classification accuracy on the channel's selection and
(iv) assess the impact of the feature selection model on classification performance.

3.2.1. Evaluating study n°1: Comparison of classical methods classification with a baseline sleep stager networks

In this experiment, using the same set of features, the predictions of an Artificial Neural Networks (ANN) were compared to the results of some selected set of classifiers mentioned above, namely, SVM, Decision Tree, Random Forest, Logistic Regression, kNN, QDA and Naïve Bayes.

The Random Forest is a generalisation of the decision trees model and correct for the tendency of decision trees to overfit the training set [51]. In this experiment, the number of trees of the Random Forest was set to 1000 which is a usual value given the large dimension of the dataset while the quality of a particular split was measured by the Gini impurity criterion.

For the kNN method, the euclidian distance was used and, after a grid search with 5-fold cross validation, the number of neighbors was fixed to $k = 15$.

Using an SVM classifier, it is possible, by the choice of the kernel

function, to conduct a linear or a nonlinear space discrimination. So, first, an SVM with a linear kernel was tried and then with a radial basis function (RBF) as a kernel in a one-versus-all multiclass approach [52]. In this latter case, the regularization parameter C and the parameter gamma for the kernel function were fine-tuned with a gridsearch with 5-folds cross validation and set respectively equal to 15 and to 0.01.

For the ANN, a baseline sleep stager network was used in this experiment, namely a feed forward neural networks, as our first choice, which consists of two hidden layers of respectively 64 and 128 neurons followed by a softmax activation.

To facilitate comparison between the different methods, all classifiers used the same features selected by the Wrapper (MLP-SFS) method described above. Accuracy, precision, recall and F1-Score are the performance metrics used to analyze the results (Tables 3 and 4).

Results shown in Table 3 indicate that MLP on one side and SVM with RBF kernel and Random Forest in the other side, exhibit substantially similar performances in the global accuracy metric with, as expected, overfitting for the Decision Tree and to a lesser extend the Random Forest classifier.

At this stage, the SVM_RBF and the Random Forest methods are effective and their performance are at the level of the basic artificial neural networks (MLP). In particular, the SVM appears slightly superior to the MLP in the global accuracy metric despite its simplicity compared to deep learning methods.

The confusion matrix of the three best performing models was also generated (Table 4) to illustrate the performance of these methods at the stage level.

As expected, of the five different stages of sleep, N1 is the hardest to classify particularly with the Random Forest. This is due to the low number of occurrences of this stage which makes it a relatively rare event. As reported in literature, the F1-score in this sleep stage can be as low as 0.30 [50,53]. In our case, the F1 score is greater than 0.50 with a value of 0.56 for the SVM classifier. We also note that misclassifications of N1 are distributed between the three stages N2, REM and Wake, except N3, with a value around 20% for the Random Forest classifier.

As shown in the confusion matrices above, the four others stages are all correctly classified by the three methods with a recall equal at least to 0.85 with 0.95 for the wake stage which is the easiest to classify. It may also be noted that SVM performs better than Neural Networks classifier in three sleep stages (N2, REM and Wake).

3.2.2. *Evaluating study n°2: Taking temporal dependency into account in the ASSC: LSTM model*

This experiment focused on time dependency of the sleep stage process. The non-random nature of the sleep stages is highlighted by the AASM in its rules of start, continuation and end of sleep stages as well as in the transition particularly from N2 to REM. The 'Time Distributed Multivariate Network' proposed by Chambon *et al.* [5] takes this time dependency into account by aggregating the features of a sequence of 2 k + 1 adjacent epochs centred on the present epoch. The resulting aggregation of features is then fed into a classifier to predict the label y_t associated to the present epoch. Similarly, Dong *et al.* [12] use lagged features to train SVM and Random Forest classifiers while Sadr *et al.* [54] use feature combining set at ± 4 epochs. As can be seen, the predictive model in these approaches takes the form $f : (X_{t-k}, \cdots, X_t, \cdots, X_{t+k}) \rightarrow y_t$ where X_t is the set of features at time t and y_t the correspondant sleep stage.

However, as documented by AASM, the temporal dependence in the sleep stages process is in terms of autocorrelation of the sleep stage variable. In a time-series dimension, the predictive model will be, in its general form, an autoregressive one with external inputs X_t:

$$f : (X_{t-k}, \cdots, X_t, y_{t-k}, \cdots, y_{t-1}) \rightarrow y_t \quad (1)$$

This specification means that the output y_t depends on *all* past features and not just on a finite lagged or forward features as it appears in [5] or [54]. This dynamic property can be seen by successive backward substitutions of y_{t-1} in (1).

With reference to the equations defining the LSTM algorithm mentioned above, one can see that it is this latter modelling that takes correctly into account the autoregressive specification (1) of the sleep stage process. This brings up the following question: does temporal dependency, as formalized in LSTM modelling, help to improve sleep stage classification using neural networks compared to conventional machine learning classifiers?

This temporal correlation was taken into account in an LSTM based-framework with a sequence-to-label architecture and a time-step of 2 epochs which means that the classifier is trained by using sequences of length 2.

A time step of 3 epochs was also considered to see if the length of the temporal lag affects the classification performance. A bidirectional LSTM was finally applied knowing that transition between stages sleep is affected not only by the nature of previous sleep stages but also by forward sleep stages.

A dedicated fully connected layer to the output of the LSTM was added. For the purpose of sleep stage classification, a final dense layer with a softmax activation and neurons equal to the number of sleep stages is needed to predict the probability of each sleep stage. Dropout as regularization technique was also used to address the overfitting problem [55]. Specifically, each unit of gate outputs of LSTM cells along with its connections are retained with a probability p independently of other units, which was set to 0.7 after a gridsearch.

Adaptive moment estimation (Adam) algorithm was applied to minimize a sparse categorical crossentropy loss function. Learning rate lr and exponential decay rate for the first and second moment estimates β_1 and β_2 were fixed at their default value in Keras, respectively $lr = 0.001$; $\beta_1 = 0.9$ and $\beta_2 = 0.999$. An early stopping callback on the validation accuracy was used to stop the training process when no improvements in the global accuracy were detected. Accordingly, the patience parameter was fine-tuned and fixed finally to 4 epochs.

The implementation code was written in Keras [56] with the back-propagation procedure completed by Tensorflow on the platform of Python [57].

In Table 5, the accuracy, the macro average F1-Score and the Cohen Kappa interrater agreement metrics are listed. Comparing these results with those of Table 3, it can be seen that the LSTM or the bidirectional LSTM do not improve the classification of sleep stages although they still deliver very acceptable accuracy. These results seem somewhat disappointing. However, as documented by AASM and mentioned in [50], the temporal stage transitions that the LSTM must capture are not explained only by the EEG or EOG signals. For example, experts measure muscle relaxation using EMG to identify N1 to R or N2 to R transitions. Similarly, as recommended by AASM, experts consider body movements and slow eye movements to identify the transition from N1 to N2. These factors cannot be taken into account by observing EEG or horizontal EOG signals alone. In other words, the features sets selected from the channels considered in this work may suffer from a bias of explanatory

Table 3
Comparison between classifiers for global accuracy (in percent).

Classifier	MLP	SVM_RBF	Random Forest	Linear SVM	Logistic Regres.	k-NN	QDA	Decision Tree	Naive Bayes
Training_score	93.5	90.5	97.0	87.5	86.8	90.3	80.7	100.0	74.6
Test_score	88.1	89.1	87.9	86.5	85.6	85.2	81.1	80.6	73.9

Table 4
Confusion matrix of the three best classifiers.

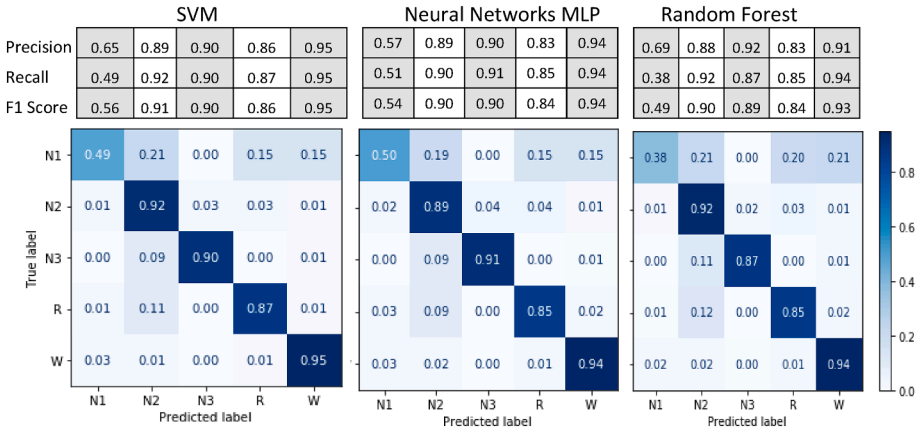

Table 5
Classification scores of LSTM (in percent).

Classifiers	Time Step = 2			Time Step = 3		
	Accuracy	Average F1 Score	Cohen Kappa	Accuracy	Average F1 Score	Cohen Kappa
LSTM	87.6	80	82.9	87.1	81	82.0
Bidirectional LSTM	87.8	81	82.4	87.1	80	82.0

factors omission. The dynamic incorporated in LSTM that links, as mentioned above in equation (1), the prediction of a sleep stage to a number of predictions of previous epochs may amplify this bias by being multiplied many times.

3.2.3. *Evaluating study n°3: Single signal-EEG versus polysomnography*

This subsection examines the influence of a reduction of the number of channels on the performance accuracy of the different classifiers.

PSG signals recording requires the patient to spend a whole night in hospital. In addition, PSG recording procedures and equipment are cumbersome and can disrupt sleep [58]. In this context, significant progress has been made in recent decades in wearable physiological monitoring devices that overcome these limitations [50,59,60,61]. The wearable EEG can record one EEG signal, and possibly, an EEG and the right and left EOG [62]. Therefore, in this experiment, only one EEG channel was first used (Fpz-Cz and Pz-Oz, respectively) and then a single EEG coupled with the horizontal EOG signal, in order to compare the results with the performance of the PSG.

Table 6
Influence of channel selection on the classification accuracy (in percent).

Classifiers	Fpz-Cz	Pz-Oz	Fpz-Cz + EOG	Pz-Oz + EOG	PSG (Fpz-Cz + Pz-Oz + EOG)
BiLSTM (time Step = 1)	84.4	82.8	86.3	86.6	87.8
BiLSTM (time step = 2)	83.7	81.8	85.1	84.9	87.1
LSTM (time step = 1)	84.2	82.3	86.7	86.6	87.6
LSTM (time step = 2)	83.5	81.6	84.5	84.2	87.1
MLP	85.0	84.1	86.3	86.3	88.1
SVM _RBF	85.0	83.1	86.7	87.3	89.1
RandomForest	85.7	84.1	85.7	86.4	87.9
LinearSVM	82.4	81.0	83.6	84.5	86.5
LogisticRegression	81.2	80.1	82.5	83.5	85.6
k-NearestNeighbors	82.3	80.8	83.2	84.6	85.2
QuadraticDiscriminant Analysis	75.8	72.5	75.9	78.0	81.1
DecisionTree	77.0	75.2	77.6	77.4	80.6
NaiveBayes	69.2	65.5	70.3	71.5	73.9

As shown in the Table 6 above, the best single EEG signal for sleep stages detection is the Fpz-Cz signal for all classifiers with an accuracy varying between 69.2% for the Naïve Bayes classifier up to 85.7% for the Random Forest classifier. This result generalizes that established by Michielli *et al.* [63] for the case of the recurrent neural networks to a set of conventional machine learning methods. As explained in [63], this result may be due to the fact that K-complexes and sleep-spindles during the stage N2 are recorded in central/frontal regions. The same remark applies for vertex sharp waves, representative of stage N1 that often occurs in central/frontal areas of the brain. It can also be noted that the combination of the EOG signal with an EEG signal (Fpz-Cz or Pz-Oz) improves the accuracy of the classifiers.

Another result that is more related to our objectives is that the ranking of methods in terms of global accuracy is not fundamentally disturbed by the use of a single channel. Even in a single EEG setting, conventional methods are still comparatively relevant with SVM-RBF and Random Forest methods exhibiting respective accuracies of 85.0% and 85.7% with the Fpz-Cz setting.

Finally, the comparison between scores based on PSG signals and those from a single-channel EEG sleep staging (or even better from an EEG/EOG combination) does not reveal excessive loss of accuracy in particular for the deep learning methods or the Random Forest and SVM-RBF classifiers. The loss of accuracy for the LSTM or the Random Forest for example is less than 1.5 percentage point with the Pz-Oz + EOG setting. This makes deep learning methods but also conventional classifiers sufficiently suitable for longitudinal monitoring using portable EEG without causing obstructions to natural sleep.

3.2.4. *Evaluating study n°4: Robustness of performance classifiers to the change of the features selection.*

Feature selection is an important and a necessary step intended to remove irrelevant and redundant features from the dataset [64]. However, the prediction accuracies are linked to the selected features as these particular features constitute the input of the classifiers considered. Therefore, in this evaluating study, the robustness of classification accuracy of different classifiers to different features selection models was tested. The classifiers concerned by this evaluation were the classical machine learning and the deep learning classifiers already used

above whereas the features selection models are the Elastic Net regression, the regularized Multinomial logistic regression and the Wrapper (MLP-SFS) method. All these selected features are listed in the Table 2 above.

Table 7 below resumes the results of classification accuracy obtained from the different classifiers according to the different feature selection models. As can be noted, whatever the classifier, the multinomial logistic regression method of features selection produced slightly better performances (classification accuracy ranging from 74.2% up to 89.6%) compared to the other features selection methods. Similarly, the Elastic Net and Wrapper (MLP-SFS) features selection methods lead to a very similar classification accuracy (ranging from 73.9% up to 88.9% or 89.1%) whatever the classifier.

It can also be noted that, the same SVM_RBF, MLP and Random Forest classifiers continue to produce the highest percentage of classification accuracy (ranging from 87.9% up to 89.6%) whatever the used features selection method whereas the Naïve Bayes classifier produce the lowest percentage of classification accuracy (ranging from 73.9% up to 74.2%) whatever the used features selection model.

3.2.5. Comparison with other studies results.

Although the focus of this article is to compare groups of classifiers and not to obtain the highest possible accuracy values, our prediction results obtained in Tables 6 And 7 can be used for comparison with other recent works in the literature. For this purpose, Table 8 resumes the performance of some of the existing sleep stage classification systems using the same physionet database.

As shown in this table, among the different studies, Fraiman and Alkhodari [65] achieved the best accuracy of 97.1% with a bi-directional LSTM and a single channel. However, by observing the training set of Fraiman and Alkhodari, it can be found that the classes are strongly unbalanced, the stage of awakening representing 68% of the total epochs. Due to this large number of samples of the awakening class in the training set, the classifier is biased towards this class and leads to skewed performance in favor of this most represented sleep stage. In the same vein, Memar and Faradji [66] achieved an accuracy value of 95%. However, their study also suffers from the same class imbalance across sleep stages. Furthermore, the nested k-fold cross validation employed by these authors to evaluate the system performance may be, according to Zhou *et al.* [67], overly-optimistic for sleep staging. Rahmani *et al.* [68], using three machine learning algorithms including RF and SVM, and a single channel EOG have also achieved an accuracy higher than 90% and assert that EOG can be a viable alternative to single channel EEG in sleep stage classification. However, when a more balanced training set was used in a second part of their study, Rahmani *et al.* observed that RF and SVM performance drops to accuracy values of 84.3% and 80.5% respectively, much less than previously obtained.

Conversely, Tsinalis *et al.* [53], Sokolovsky *et al.* [69] and Wang *et al.* [70] tackle the problem of misclassification due to class imbalance in training data. Sokolovsky *et al.* keep only the data between sleep onset and the last epoch of sleep and get an accuracy of 81%. For their part, Wang *et al.,* after undersampling the wake classes and balancing the dataset before combining EEGNet and an LSTM model, obtained, in a multichannel setting (EEG and EOG), a 90% accuracy rate slightly higher than our results.

Compared with the rest of the sleep stage classification systems in Table 8, it may be seen that the accuracies involved by our results are globally competitive with the state-of-the-art results.

Table 7
Results of classification accuracy using different features sets (in percent).

Classifiers	Elastic Net Features	Multinomial Logistic_Features	Wrapper (MLP-SFS)_Features
BiLSTM	87.5	87.9	87.8
LSTM	87.3	87.9	87.6
MLP	88.4	89.0	88.1
SVM _RBF	88.9	89.6	89.1
RandomForest	87.9	88.2	87.9
LinearSVM	86.5	86.8	86.5
LogisticRegression	85.6	86.1	85.6
k-Nearest Neighbors	85.1	86.5	85.2
Quadratic Discriminant Analysis	81.1	80.8	80.4
DecisionTree	80.6	81.1	80.6
NaiveBayes	73.9	74.2	73.9

Table 8
Comparison with results from other studies.

Authors	Models	Channels	Accuracy rate
Tsinalis *et al.* (2016) [53]	CNN	Fpz-Cz	81%
Memar & Faradji (2017) [66].	RF with nested 5-fold cross validation	Pz-Oz	95%
Rahman, M. *et al.* (2018) [69]	RUSBoost, RF and SVM	EOG	90.0%; 91.0% and 91.7%
Mousavi *et al.* (2019) [71]	CNN and BiLSTM with Attention	Fpz_Cz	84.3%
Sokolovsky *et al.* (2019) [69]	CNN	Fpz-Cz; Pz-Oz; EOG	81%
Zhou *et al.* (2020) [67]	RF combined with LightGBM	Pz-Oz	85.3%
Wang *et al.* (2020) [70]	Combined EEGNet and Bi-LSTM	EEG + EOG	90.0%
Fraiwan & Alkhodari (2020) [65]	Bi-directional LSTM	Fpz-Cz	97.1%
Fu, M. *et al.* (2021) [58]	Combination of Attention with BiLSTM	Fpz_Cz	83.8%
Duan *et al.* (2021) [72]	Deep learning fusion network	Fpz-Cz; Pz-Oz; EOG	87.5%

4. Conclusion

The objective of this study was to carry out a comparative study on conventional machine learning and feature-based deep learning approaches in the context of automatic sleep stage classification.

Since it is difficult to apply machine learning methods to raw EEG data as they are likely to be affected by noise from different sources, the data were first pre-processed using a hybrid denoising approach. In addition, in order to avoid spurious relationships, the dataset has been randomly split into training data and test data making it completely independent of the training and validation process. A set of representative features was then extracted and selected by cross validation from the training data to compress the information to be used.

Well-known regularization techniques were used to reduce overfitting, namely the addition of l_1 and/or l_2 penalties to the loss functions and the use of drop out layers in the deep learning architecture.

Based on this methodology and on models whose accuracy is close to the state of the art, the following results were obtained:

- After testing various machine learning algorithms, conventional machine learning models proved to be as valid as feature-based neural networks and continue to provide significant performance to predict the sleep stages classification.
- SVM_RBF and Random Forest were systematically ranked among the best conventional models in ASSC.
- These conclusions are robust: they remains valid even after (i) the introduction of inter-epochs temporal dependence in neural networks, (ii) a reduction in the number of channels or (iii) a change in feature selection model.

As implication of these findings, we suggest that conventional machine learning models should continue to have a key role in the

classification and prediction of sleep stages with regard to their performance compared to deep learning models.

Declaration of Competing Interest

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

References:

[1] American Academy of Sleep Medicine, International Classification of Sleep Disorders, 3rd ed., American Academy of Sleep Medicine, Darien, IL, 2014.
[2] A. Rechtschaffen, A. Kales, A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects, National Institutes of Health, Bethseda, MD, 1968.
[3] [2] Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., ... & Vaughn, B. V. (2017). AASM scoring manual updates for 2017 (version 2.4).
[4] D. Moser, P. Anderer, G. Gruber, S. Parapatics, E. Loretz, M. Boeck, G. Dorffner, Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters, Sleep 32 (2) (2009) 139–149.
[5] S. Chambon, M.N. Galtier, P.J. Arnal, G. Wainrib, A. Gramfort, A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (4) (2018) 758–769.
[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[7] Bengio, Y. (2007). Learning deep architectures for Al. Foundations and Trends in Machine Learning.-2009.—2 (1).-pp, 1-127.
[8] C.M. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995.
[9] H.W. Loh, C.P. Ooi, J. Vicnesh, S.L. Oh, O. Faust, A. Gertych, U.R. Acharya, Automated Detection of Sleep Stages Using Deep Learning Techniques: A Systematic Review of the Last Decade (2010–2020), Appl. Sci. 10 (24) (2020) 8963.
[10] X. Qian, Y. Qiu, Q. He, Y. Lu, H. Lin, F. Xu, J. Shuai, A Review of Methods for Sleep Arousal Detection Using Polysomnographic Signals, Brain Sciences 11 (10) (2021) 1274.
[11] [10] Biswal, S., Kulas, J., Sun, H., Goparaju, B., Westover, M. B., Bianchi, M. T., & Sun, J. (2017). SLEEPNET: automated sleep staging system via deep learning. arXiv preprint arXiv:1707.08262.
[12] H. Dong, A. Supratak, W. Pan, C. Wu, P.M. Matthews, Y. Guo, Mixed neural network approach for temporal sleep stage classification, IEEE Trans. Neural Syst. Rehabilitation Eng. 26 (2) (2017) 324–333.
[13] B. Şen, M. Peker, A. Çavuşoğlu, F.V. Çelebi, A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms, J. Med. Syst. 38 (3) (2014) 1–21.
[14] R. Boostani, F. Karimzadeh, M. Nami, A comparative review on sleep stage classification methods in patients and healthy individuals, Comput. Methods Programs Biomed. 140 (2017) 77–91.
[15] Z. Xu, X. Yang, J. Sun, P. Liu, W. Qin, Sleep stage classification using time-frequency spectra from consecutive multi-time points, Front. Neurosci. 14 (2020) 14.
[16] [15] PhysioNet, The Sleep-Edf Database. Available online: https://www.physionet.org/content/sleep-edfx/1.0.0/ (accessed on 29 avril 2021).
[17] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation [Online]. 101 (23) (2000) e215–e220.
[18] J.A. Urigüen, B. Garcia-Zapirain, EEG artefact removal—state-of-the-art and guidelines, J. Neural Eng. 12 (3) (2015).
[19] Dora, C., & Biswal, P. K. (2017). Automated detection of nonphysiological artefacts in polysomnographic EEG using conventional signal processing techniques. In TENCON 2017-2017 IEEE Region 10 Conference (pp. 1568-1572). IEEE.
[20] K.T. Sweeney, T.E. Ward, S.F. McLoone, Artefact removal in physiological signals—Practices and possibilities, IEEE Trans. Inf Technol. Biomed. 16 (3) (2012) 488–500.
[21] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, G. Pfurtscheller, A fully automated correction method of EOG artefacts in EEG recordings, Clin. Neurophysiol. 118 (1) (2007) 98–104.
[22] X. Jiang, G.B. Bian, Z. Tian, Removal of artefacts from EEG signals: a review, Sensors 19 (5) (2019) 987.
[23] M. Dursun, S. Özşen, C. Yücelbaş, Ş. Yücelbaş, G. Tezel, S. Küççüktürk, Ş. Yosunkaya, A new approach to eliminating EOG artifacts from the sleep EEG signals for the automatic sleep stage classification, Neural Comput. Appl. 28 (10) (2017) 3095–3112.
[24] M. Tavakoli, H. Ahani, Removing EOG Artifacts from EEG Signals Using a Modified Wavelet-RLS Method, J. Bioeng. Res. 2 (2) (2020).
[25] Li, P., Chen, Z., & Hu, Y. (2017). A method for automatic removal of EOG artifacts from EEG based on ICA-EMD. In 2017 Chinese Automation Congress (CAC) (pp. 1860-1863). IEEE.
[26] T.D. Lagerlund, F.W. Sharbrough, N.E. Busacker, Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition, J. Clin. Neurophysiol. 14 (1) (1997) 73–82.
[27] P.S. Kumar, R. Arumuganathan, K. Sivakumar, C. Vimal, Removal of ocular artifacts in the EEG through wavelet transform without using an EOG reference channel, Int. J. Open Problems Compt. Math 1 (3) (2008) 188–200.
[28] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural networks 13 (4–5) (2000) 411–430.
[29] N.P. Castellanos, V.A. Makarov, Recovering EEG brain signals: artifact suppression with wavelet enhanced independent component analysis, J. Neurosci. Methods 158 (2) (2006) 300–312.
[30] S. Najdi, A.A. Gharbali, J.M. Fonseca, Feature ranking and rank aggregation for automatic sleep stage classification: a comparative study, Biomed. Eng. Online 16 (1) (2017) 1–19.
[31] Yulita, I. N., Fanany, M. I., & Arymurthy, A. M. (2018). Fast convolutional method for automatic sleep stage classification. Healthcare informatics research, 24(3), 170.
[32] Dib, N. (2015). Analyse non linéaire des différents intervalles du signal ECG en vue d'une reconnaissance de signatures de pathologies cardiaques. Thèse de doctorat. Département de génie biomédical. Laboratoire de recherche en Génie Biomédical-Université de Tlemcen, Algéria.
[33] Y. Ma, W. Shi, C.K. Peng, A.C. Yang, Nonlinear dynamical analysis of sleep electroencephalography using fractal and entropy approaches, Sleep Med. Rev. 37 (2018) 85–93.
[34] Petrosian, A. (1995, June). Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns. In Proceedings eighth IEEE symposium on computer-based medical systems (pp. 212-217). IEEE.
[35] Klonowski, W., Olejarczyk, E., & Stepien, R. (2005, October). Sleep-EEG analysis using Higuchi's fractal dimension. In International Symposium on Nonlinear Theory and its Applications (pp. 18-21).
[36] Cusenza, M., Accardo, A., D'Addio, G., & Corbi, G. (2010, September). Relationship between fractal dimension and power-law exponent of heart rate variability in normal and heart failure subjects. In 2010 Computing in Cardiology (pp. 935-938). IEEE.
[37] A. Rizal, R. Hidayat, H.A. Nugroho, Hjorth descriptor measurement on multidistance signal level difference for lung sound classification, J. Telecommun. Electr. Comput. Eng. (JTEC) 9 (2) (2017) 23–27.
[38] R. Jenke, A. Peer, M. Buss, Feature extraction and selection for emotion recognition from EEG, IEEE Trans. Affective Comput. 5 (3) (2014) 327–339.
[39] Ebrahimi, F., Mikaili, M., Estrada, E., & Nazeran, H. (2007, August). Assessment of Itakura distance as a valuable feature for computer-aided classification of sleep stages. In 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 3300-3303). IEEE.
[40] A.A. Gharbali, S. Najdi, J.M. Fonseca, Investigating the contribution of distance-based features to automatic sleep stage classification, Comput. Biol. Med. 96 (2018) 8–23.
[41] McDonald, G.C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics. 1(1), 93-100.
[42] van Wieringen, W. N. (2015). Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169.
[43] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc.: Ser. B (Methodological) 58 (1) (1996) 267–288.
[44] Dehkordi, P., Garde, A., Dumont, G. A., & Ansermino, J. M. (2016). Sleep/wake classification using cardiorespiratory features extracted from photoplethysmogram. In 2016 Computing in Cardiology Conference (CinC) (pp. 1021-1024). IEEE.
[45] Azimi, H., Gunnarsdottir, K. M., Sarma, S. V., Gamaldo, A. A., Salas, R. M., & Gamaldo, C. E. (2020). Identifying Sleep Biomarkers to Evaluate Cognition in HIV. 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 2332-2336). IEEE.
[46] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks 18 (5–6) (2005) 602–610.
[47] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.
[48] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Second edition 2017.
[49] Winkler, I., Debener, S., Müller, K. R., & Tangermann, M. (2015, August). On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4101-4105). IEEE.
[50] O. Tsinalis, P.M. Matthews, Y. Guo, Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders, Ann. Biomed. Eng. 44 (5) (2016) 1587–1597.
[51] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
[52] 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.
[53] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, Y.-H. Pan, Y.-H. Wang, Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models, IEEE Trans. Instrum. Meas. 61 (6) (2012) 1649–1657.
[54] Sadr, N., & de Chazal, P. (2018, September). Automatic scoring of non-apnoea arousals using the polysomnogram. In 2018 Computing in Cardiology Conference (CinC) (Vol. 45, pp. 1-4). IEEE.
[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[56] Chollet, F. (2015). 'Keras.' https://github.com/fchollet/keras.
[57] [56] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.
[58] M. Fu, Y. Wang, Z. Chen, J. Li, F. Xu, X. Liu, F. Hou, Deep Learning in Automatic Sleep Staging With a Single Channel Electroencephalography, Front. Physiol. 12 (2021) 179.
[59] A.J. Casson, Wearable EEG and beyond, Biomed. Eng. Lett. 9 (1) (2019) 53–71.
[60] Nakamura, T., Goverdovsky, V., Morrell, M. J., & Mandic, D. P. (2017). Automatic sleep monitoring using ear-EEG. IEEE journal of translational engineering in health and medicine, 5, 1-8.
[61] Zhang, Y., Yang, Z., Lan, K., Liu, X., Zhang, Z., Li, P., …& Pan, J. (2019, April). Sleep stage classification using bidirectional LSTM in wearable multi-sensor systems. In IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (pp. 443-448). IEEE.
[62] E. Bresch, U. Großekathöfer, G. Garcia-Molina, Recurrent deep neural networks for real-time sleep stage classification from single channel EEG, Front. Comput. Neurosci. 12 (2018) 85.
[63] N. Michielli, U.R. Acharya, F. Molinari, Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals, Comput. Biol. Med. 106 (2019) 71–81.
[64] E.M. Karabulut, S.A. Özel, T. Ibrikci, A comparative study on the effect of feature selection on classification accuracy, Procedia Technol. 1 (2012) 323–327.
[65] L. Fraiwan, M. Alkhodari, Investigating the use of uni-directional and bi-directional long short-term memory models for automatic sleep stage scoring, Inf. Med. Unlocked 20 (2020).
[66] P. Memar, F. Faradji, A novel multi-class EEG-based sleep stage classification system, IEEE Trans. Neural Syst. Rehabilitation Eng. 26 (1) (2017) 84–89.
[67] J. Zhou, G. Wang, J. Liu, D. Wu, W. Xu, Z. Wang, Y. Tian, Automatic sleep stage classification with single channel EEG signal based on two-layer stacked ensemble model, IEEE Access 8 (2020) 57283–57297.
[68] M.M. Rahman, M.I.H. Bhuiyan, A.R. Hassan, Sleep stage classification using single-channel EOG, Comput. Biol. Med. 102 (2018) 211–220.
[69] M. Sokolovsky, F. Guerrero, S. Paisarnsrisomsuk, C. Ruiz, S.A. Alvarez, Deep learning for automated feature discovery and classification of sleep stages, IEEE/ACM Trans. Comput. Biol. Bioinf. 17 (6) (2019) 1835–1845.
[70] Wang, I. N., Lee, C. H., Kim, H. J., Kim, H., & Kim, D. J. (2020, October). An Ensemble Deep Learning Approach for Sleep Stage Classification via Single-channel EEG and EOG. In 2020 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 394-398). IEEE.
[71] S. Mousavi, F. Afghah, U.R. Acharya, SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach, PLoS One 14 (5) (2019).
[72] L. Duan, M. Li, C. Wang, Y. Qiao, Z. Wang, S. Sha, M. Li, A Novel Sleep Staging Network Based on Data Adaptation and Multimodal Fusion, Front. Hum. Neurosci. 600 (2021).