Alexandra-Maria Tăuţan*, Alessandro C. Rossi, Ruben de Francisco and Bogdan Ionescu

Dimensionality reduction for EEG-based sleep stage detection: comparison of autoencoders, principal component analysis and factor analysis

https://doi.org/10.1515/bmt-2020-0139
Received May 25, 2020; accepted August 19, 2020;
published online October 12, 2020

Abstract: Methods developed for automatic sleep stage detection make use of large amounts of data in the form of polysomnographic (PSG) recordings to build predictive models. In this study, we investigate the effect of several dimensionality reduction techniques, i.e., principal component analysis (PCA), factor analysis (FA), and autoencoders (AE) on common classifiers, e.g., random forests (RF), multilayer perceptron (MLP), long-short term memory (LSTM) networks, for automated sleep stage detection. Experimental testing is carried out on the MGH Dataset provided in the "*You Snooze, You Win: The PhysioNet/Computing in Cardiology Challenge 2018*". The signals used as input are the six available (EEG) electoencephalographic channels and combinations with the other PSG signals provided: ECG – electrocardiogram, EMG – electromyogram, respiration based signals – respiratory efforts and airflow. We observe that a similar or improved accuracy is obtained in most cases when using all dimensionality reduction techniques, which is a promising result as it allows to reduce the computational load while maintaining performance and in some cases also improves the accuracy of automated sleep stage detection. In our study, using autoencoders for dimensionality reduction maintains the performance of the model, while using PCA and FA the accuracy of the models is in most cases improved.

keywords: autoencoders; EEG; factor analysis; LSTM; MLP; principal component analysis; random forests; sleep staging.

Introduction

When investigating sleep related problems, health care professionals make use of polysomnographic (PSG) recordings to monitor and analyze the patients during sleep. PSG studies can span several hours and may include signals such as (EEG) electroencephalogram, (ECG) electrocardiogram, (EMG) electromiogram, respiration related signals, etc.

For identifying sleep patterns and placing a diagnostic, clinicians usually analyze all the recorded signals manually to classify the different stages of sleep. The annotation process is cumbersome and can take a significant amount of time since signals should be annotated in short time windows. According to the American Academy of Sleep Medicine (AASM) Guidelines, sleep stages should be labeled on 30 s epochs [1]. Automatic sleep stage classification algorithms can ease the burden of manually annotating each epoch.

The applications of automatic sleep stage detection algorithms mainly help physicians in providing a faster and more accurate diagnosis by facilitating the annotation process. A typical sleep study does not only imply a significant amount of work for the annotator, but it also brings a significant discomfort to the patient. The patient normally conducts a sleep study in a sleep clinic, while connected to the PSG recording equipment. This brings a certain degree of discomfort and impacts the sleep quality. With the advent of wearable technologies and remote health monitoring, the process of recording sleep studies can be simplified and improved. The patient could potentially record a sleep study at home, with a more user friendly equipment. Automatic sleep scoring algorithms can also be of potential help in this case. Based on the output of the algorithms, sleep disorders can more easily be diagnosed and alarms can be triggered in case of severe respiratory or cardiac problems.

By using a smaller set of signals, the equipment that needs to be applied to the patient is less bothersome. If by using only EEG signals, a sufficient accuracy is obtained

***Corresponding author: Alexandra-Maria Tăuţan**, University Politehnica of Bucharest, Splaiul Indepenţei 313, 060042, Bucharest, Romania, E-mail: alexandra.tautan@upb.ro
Alessandro C. Rossi and Ruben de Francisco, Onera Health, Eindhoven, The Netherlands
Bogdan Ionescu, University Politehnica of Bucharest, Bucharest, Romania

from the automatic sleep scoring algorithm, all other signal types would not be required in a remote recording.

Automatic sleep staging algorithms combined with wearable technologies can open the possibility of wide population screening. Some sleep disorders can be underlying symptoms of bigger health problem or a prodromal symptom of serious diseases. For instance, some motor types of neurodegenerative diseases present rapid eye movement (REM) sleep behavior disorder (lack of atonia during the REM sleep stage) years prior to the actual onset of the motor symptoms [2]. Its earlier detection can help in providing adequate treatment sooner.

Many methods have been proposed for automatic sleep stage classification [3, 4]. A more detailed view is provided in section 4. A typical algorithm setup extracts features from recorded PSG signals and uses them as input for a classification algorithm. The obtained model can then predict sleep stage labels on unseen data. One of the factors contributing to the performance of the model is the information used for training. The type of input data and its variety has a direct impact on the patterns that the model can recognize. The biomedical signals used to characterize sleep patterns show a large variation in characteristics from person to person and from healthy individuals to those suffering from diverse pathologies. Training on a larger data set might capture more of the signal variations and provide a better prediction power. However using large amounts of data can become computationally expensive.

Dimensionality reduction techniques are often used when dealing with large data sets, to ease the computational requirements while maintaining a good performance. They can also be useful in representing data into formats that enhance the properties of each class, which in this case is represented by sleep stages.

In this paper, our contributions are focusing on dimensionality reduction for preprocessed PSG data, namely: (i) use of a simple autoencoder network as a dimensionality reduction technique, and (ii) study the effects of different dimensionality reduction methods (FA, PCA, AE) on different types of classifier models. By using less computational load and less memory, more information can be added to the model, potentially broadening its applicability as more variability is included.

The paper is organized as follows. The next section provides a brief overview of previous work on automatic sleep stage detection and dimensionality reduction. The section Materials and Methods presents different aspects of the methods used: raw signals used and data set preprocessing, extracted features, dimensionality reduction techniques and classifiers. The section Results details the experimental results while the last section concludes the paper.

Previous work

The topic of automatic sleep stage detection is abundantly present in scientific literature. Many algorithm variations are available based on the type of input signals and method of classification used. Here we focus on studies that use EEG data as input. Methods for automatic sleep scoring can make use of single channel EEG recordings [5, 6] or of multiple EEG channels [7]. Classification models are in most cases built on extracted features. Features are extracted from the time domain [8–10], frequency domain [7, 11], time-frequency domain [12, 13] or a different representation of the data [14]. These types of features are generally used as input for classic algorithms such a support vector machines [13], k-nearest neighbor [14], random forests (RF) [6] etc.

In recent years, neural networks have also been extensively used in the problem of automatic sleep stage classification. Different architectures were created including for instance convolutional [15] or long-short term memory (LSTM) [8] layers or well known deep neural network architectures such as VGGNet [16]. Some of the neural network methods use extracted features as input, while some are able to use raw data as input represented either in time [5] or in the form of spectrograms [16, 17]. Some neural network architectures explore the properties of convolutional layers to extract frequency features from the signal prior to the output classification layers [15, 18].

Regardless of the type of input data used for the training of the sleep scoring model, the data amounts can be large as signals are annotated in 30 s epochs and several hours of recordings are collected during a sleep study. The high dimensionality of the data can become problematic due to memory constraints and the requirements for computational power.

Different methods have been proposed in literature to reduce the size of large data sets and represent information in a more convenient way for further processing [14, 19]. For instance, data can be transformed in different dimensions that would better capture data variability using FA or PCA [19–21]. Fan et al. [21] uses multi-scale entropy combined with PCA to extract features and automatically detect sleep stages from the MIT-BIH database. The final accuracy reached 87.9%.

AE can compress the input data to different degrees. When applied to automatic sleep stage detection, AE networks are mostly used for classification purposes [5, 22].

Tsinalis et al. [5] uses a single frontal EEG channel from the sleep EDF dataset to classify sleep stages. Features are extracted using the wavelet transform and a stacked sparse autoencoder with 20 layers is used for classification. The highest obtained mean accuracy is of 88%. Najdi et al. [22] uses the ISRUC data set to automatically classify sleep stages using a feature based model. Features are extracted from EEG, electrooculogram and EMG data and are selected using a discriminative feature selection algorithm. These are fed into a stacked sparse autoencoder resulting in a classification accuracy of 82.2%. Feature extraction or dimensionality reduction can also be obtained using autoencoders. Prabhudesai et al. [23] used a convolutional autoencoder to extract spectral features representative of the different EEG power bands. These features were combined with a linear discriminant analysis classifier to obtain the sleep stages. Statistically better results were obtained when using autoencoders for feature extraction.

Feature selection can also be considered a technique of reducing the amount of information while maintaining the level of performance. The author in [24] compares various feature ranking algorithms and explores the use of autoencoders for further feature transformation. Different classifiers are used for EEG based sleep scoring. By using feature ranking in the classification, the obtained mean accuracy was 75%. When adding autoencoders for feature transformation, the overall classification reached 82.2%.

Although dimensionality reduction techniques have been used previously in automatic sleep stage detection, in this study we aim to compare several techniques of data representation and compression while observing their effects on the results obtained with different types of classification algorithms such as RF, multilayer perceptron (MLP) and LSTM.

Materials and methods

Our approach for creating an automated sleep scoring model was to extract features from raw sensor data (EEG, ECG, EMG and respiration based signals recorded during sleep) and use these features as input to various classifiers. Dimensionality reduction techniques were applied after feature extraction as presented in Figure 1. The obtained representation of the data is used for training the classifiers. The model obtained is afterwards tested. Performance evaluation was conducted for each model in a cross-validation scenario.

Data set description

In this study, we used the PSG data set provided by Massachussets General Hospital on Physionet as part of the challenge *'You Snooze, You Win: The PhysioNet/Computing in Cardiology Challenge 2018'* [25, 26]. The data set was chosen due to the abundance of annotated data according to the AASM guideline. The available training set contains PSG data from 994 subjects. The AASM guideline defines five stages of sleep: Wakefulness (W), N1 (non-REM light sleep), N2 (non-REM light sleep), N3 (non-REM deep sleep) and REM [1]. A label is placed every 30 s of recording. This results in a total of approximately 2×10^6 annotated epochs that can be used as input for classification.

Each PSG recording contains six EEG channels (F3M2, F4M1, C3M2, C4M1, O1M2, O2M1), a submentalis EMG, ECG and signals monitoring respiratory effort from the chest, abdomen and an airflow signal.

Raw signals and preprocessing

A PSG study contains multiple physiological recordings which can also be used for the automatic classification of sleep stages. Using only EEG data can simplify the physical setup required for recording. In our study, we have used two signal combinations: six EEG channels, six EEG channels combined with EMC, ECG and respiratory based signals [25]. The latter combination was chosen as it includes information from all recorded signals, thus increasing further the size of the data set, while adding more aspects that can improve classification.

Sleep stages are distributed unequally throughout sleep. Non-REM and REM sleep have a cyclic structure. Non-REM sleep is prevalent throughout the night taking up to 75–80% of the total sleep time, whereas REM sleep represents the remaining 20–25% [27]. N1 sleep generally occurs right after wakefulness and represents a small percent out of the total sleep time. N2 represents the majority of sleeping time, while deep sleep from the N3 stage is generally around 10% of the total sleep time [27]. Therefore PSG data sets are unbalanced, as more instances of one class are available for training the classifier than others. To provide the classifiers with an equal amount of data from each sleep stage class, we have preprocessed the distribution of classes in the data set. We have selected the smallest available class and randomly sampled the other classes such that all sleep stages will be equally represented in the input seen by the

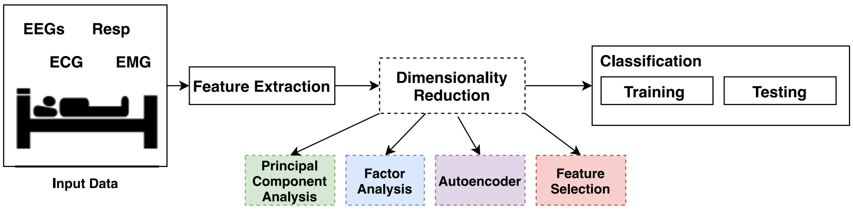

Figure 1: Overview of the method of obtaining the automatic sleep stage model from polysomnographic recordings.

Table 1: Extracted features in time and frequency domain from each electoencephalographic (EEG), electrocardiogram (ECG), electromyogram (EMG) and Respiratory-based signals [11, 33]. All features are extracted on 30 s windows. The last column represents the total number (#) of features extracted from one epoch of one signal.

		Name	Description	Name	Description	#
EEG	Time	meanA	Mean amplitude	maxA	Maximum amplitude	28
		skewS	Skewness	kurtosisS	Kurtosis	
		stdS	Standard deviation			
	Frequency	meanP	Mean spectrum value in EEG band	maxP	Maximum spectrum value in EEG band	
		minP	Minimum spectrum value in EEG band	stdF	Standard deviation of the EEG band spectrum	
		kurtosisF	Kurtosis of the EEG band spectrum	Delta/theta	Ratio between mean spectrum power in delta and theta band	
		Delta/theta	Ratio between mean spectrum power in delta and theta band	Theta/alpha	Ratio between mean spectrum power in theta and alpha band	
		Delta/alpha	Ratio between mean spectrum power in delta and alpha band			
ECG	Time	RR interval	Mean interval between detected R peaks of the ECG epoch	BPM	Beats per minute	14
		RMSSD	Root mean square of HRV signal	SDNN	Standard deviation of HRV signal	
		minHRV	Minimum of HRV signal	maxHRV	Maximum of HRV signal	
		skewHRV	Skewness of HRV signal	kurtosisHRV	Kurtosis of HRV signal	
		entropyHRV	Spectral entropy of HRV signal			
	Frequency	TF	Total frequency – mean power spectrum of HRV <0.4 Hz	VLF	Very low frequencies - mean power spectrum of HRV <0.04 Hz	
		LF	Low frequencies - mean power spectrum of HRV between 0.04 and 0.15 Hz	HF	High frequencies - mean power spectrum of HRV between 0.15 and 0.4 Hz	
		LFHF	Ratio between low and high frequencies			
EMG	Time	meanE	Mean amplitude	minE	Minimum amplitude	11
		maxE	Maximum amplitude	skewE	Skewness	
		kurtosisE	Kurtosis	varianceE	Variance	
		rmsE	Root mean square of the EMG epoch	entropyE	Spectral entropy of the EMG epoch	
	Freq	MaxFE	Frequency at which the power spectrum is maximum in the EMG epoch	maxPSDE	Maximum of the power spectrum in the EMG epoch	
		skewE	Skewness			
Resp	Time	meanR	Mean amplitude	skewR	Skewness	12
		kurtosisR	Kurtosis	varianceR	Variance	
		stdR	Standard deviation	nPeaks	Number of peaks detected in an epoch	
		meanNPeaks	Mean distance between two consecutive peaks in an epoch	stdNPeaks	Standard deviation of the distance between two consecutive peaks in an epoch	
		skewNPeaks	Skewness of the distance between two consecutive peaks in an epoch			
	Freq	MaxFR	Frequency at which the power spectrum is maximum	maxPSDR	Maximum of the power spectrum in the epoch	
		meanPSDR	Mean of the power spectrum in the epoch			

classifiers. This is also the case in the data set we have chosen for our experiments.

Extracted features

Biomedical signals are typically non-stationary and so the features used to characterize them are diverse, from both time and frequency domains. A window of 30 s was selected for feature extraction to match the annotated sleep stages. An overview of all the extracted features is presented in Table 1. Each stage of sleep is represented differently by the characteristics of the various PSG signals. All of the signals show changes corresponding to the physiological states from the different cycles of sleep.

EEG: Information in the EEG signal is encoded both in the amplitude fluctuations from the time domain as well as in the frequency changes of the signal. Most often the frequency content of the EEG is used as an indication of the state of the subject. In this work, we considered four EEG frequency bands: *Delta* – between 0.5 and 4 Hz, *Theta* – between 4 and 8 Hz, *Alpha* – between 8 and 12 Hz, *Sigma* – between 12 and 20 Hz. The five stages of sleep defined by the AASM guidelines [1], comprise different EEG rhythms. The wakefulness state mostly consists of alpha

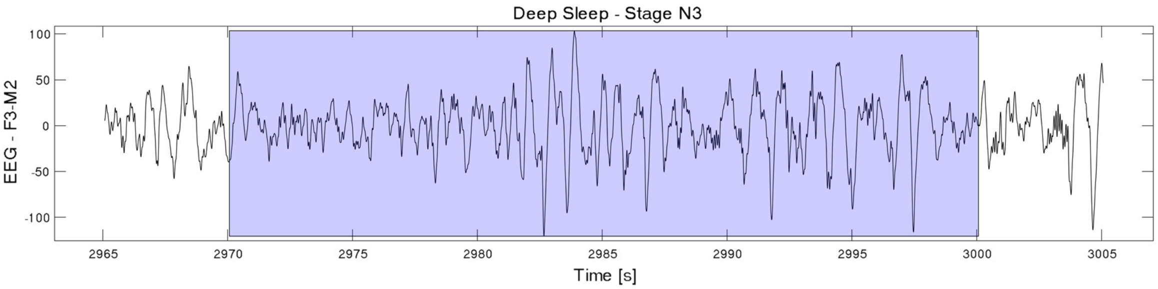

Figure 2: Selected epoch from the EEG F3-M2 channel from subject tr03-0005 of the database annotated as *Wake*. The highlighted portion represents the annotated section.

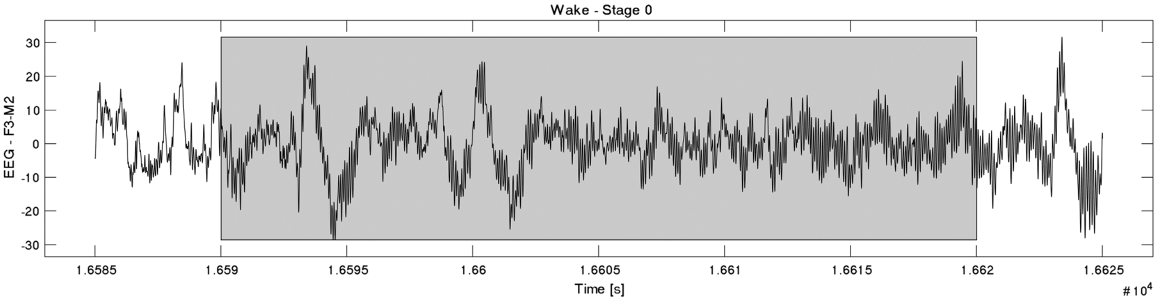

Figure 3: Selected epoch from the EEG F3-M2 channel from subject tr03-0005 of the database annotated as *N3*. The highlighted portion represents the annotated section.

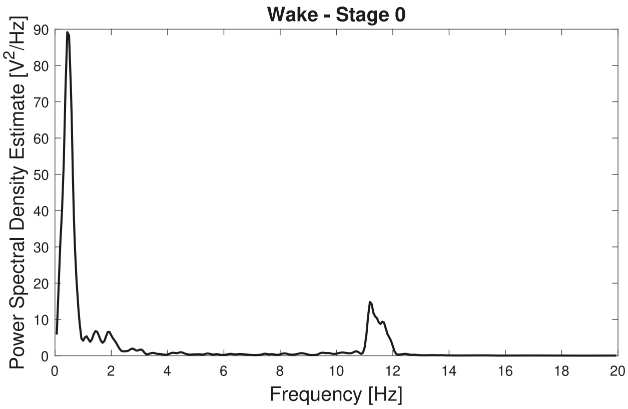

Figure 4: Power spectral density estimate of the EEG epoch represented in Figure 2.

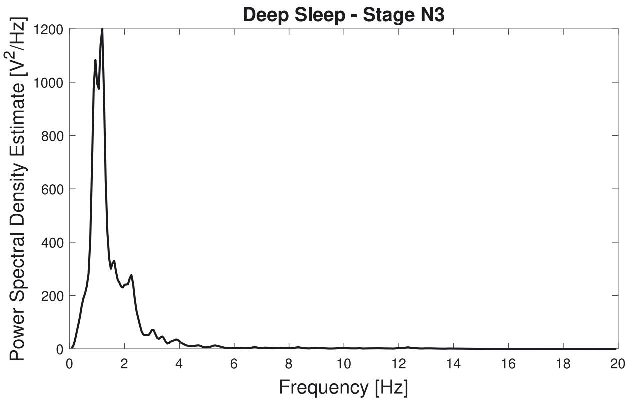

Figure 5: Power spectral density estimate of the EEG epoch represented in Figure 3.

and sigma EEG oscillations. The N1 state has predominant theta activity combined with slow sigma, while the N2 state has predominant theta activity with minimal alpha waves. The N3 sleep stage, also known as deep sleep, contains mostly delta waves. REM activity is characterized mostly by the presence of theta activity on the EEG signals [27].

Figures 2, 3 show two EEG epochs in time for the wake and N3 stage respectively. Figures 4, 5 present the frequency domain representations of the same EEG epochs. Differences between the two EEG epochs can be observed both in the time and frequency domain. As it can be seen from Figure 2 and Figure 4, strong alpha waves are present on the EEG trace with a characteristic frequency of around 11–12 Hz. When looking at deep sleep (Figure 3), the amplitude in time of the signal is significantly increased while the rhythm is decreased. This can also be seen in the frequency domain representation, where alpha waves are no longer visible and the spectrum is dominated by strong low frequency components (delta waves). Both time and frequency domains can be used to characterize sleep stages from EEG signals.

From each of the six EEG channels provided in the selected database, 5 time domain features and 23 frequency domain features were extracted [7, 11]. The frequency domain features characterizing in-band power (meanP, maxP, minP, stdP, kurtosisF — see Table 1) were extracted for each of the defined EEG frequency bands. Therefore a total of 28 features were extracted from each EEG channel.

ECG: The most relevant information regarding sleep from the ECG signal includes characteristics of heart rate (HR) and heart rate variability (HRV). HR is significantly lower during deep sleep and the other stages of non-REM sleep when compared to the wakefulness state. During REM sleep, the HR increases slightly but does not reach the level from wakefulness [28]. To capture these fluctuations, features were extracted from both the time and frequency domain from the HR and HRV signals constructed from the detected ECG R peaks [29, 30]. The extracted features are detailed in Table 1. A total of 14 features were used.

EMG: Muscle tone changes throughout the different stages of sleep. These changes are captured with the help of EMG measurements. During non-REM sleep, the muscle tone and therefore the EMG activity is similar to that found in wakefulness. During REM sleep, the muscles experience atonia in order to no enact dreams. This can be seen on the EMG signal, which presents significantly less activity than in other sleep phases. A total of 11 time and frequency domain features were extracted from the provided EMG signal as detailed in Table 1.

Respiration: During the different sleep cycles, respiration has a similar pattern as heart rate (HR). During non-REM sleep, the respiratory activity is lower than compared to wakefulness. In REM sleep, the respiratory activity increases but does not reach the level of the activity from wakefulness [27]. These fluctuations are monitored in the chosen data set using three signals – obtained from chest and abdomen belts measuring respiratory efforts and a sensor monitoring airflow. Both time and frequency domain features are extracted from each individual signal. The respiratory signals show peaks and valleys corresponding to the inspiration and expiration effort. Theses peaks are also detected and features are extracted from the differences between consecutive peaks [31]. All features are detailed in Table 1. A total of 12 features are obtained for each of the three respiratory signals, so a total of 36 features characterize the respiration effort during sleep.

In our experiments, when using all six EEG channels, a total of 168 features per epoch are used as input to the classifier. When combining the EEG channels with ECG, EMG and Respiration based features, the input increases to 205 features per epoch. Taking the size of the data set into consideration (see Section 4), the amount of input instances to the classifier is computationally expensive. In some cases, additional features do not improve performance but actually reduce it [32]. Selecting features or representing data sets in a more suitable manner for classification can improve performance, while reducing computational costs. Dimensionality reduction techniques are applied for the two feature sets.

Dimensionality reduction

The aim of our dimensionality reduction efforts was to reduce the computational power and memory requirements, while attempting to improve performance though different data representations. When experimenting with different techniques, we have kept the output data sizes constant for comparison. Experiments were performed with a total number of features being reduced to 3, 5, 30, and 50 components only. These components were then fed to the classifier.

Principal component analysis (PCA): PCA is a linear transformation that maps the input data into a lower dimension while maintaining the variance of the original data set [19]. PCA finds a linear mapping between the covariance matrix of the data which results in a number of principal eigenvectors. The number of eigenvectors represents a lower dimensionality of the input information and so allows for a dimensionality reduction of the original dataset while maintaining relevant information. In our experiments, we used an incremental PCA implementation [34] with a batch size of 64. The number of eigenvectors was changed to 3, 5, 30, and 50.

Factor analysis (FA): FA analyzes the correlations between different variables by creating a set of common factors or latent variables [20]. FA assumes that the initial information can be grouped by the correlation between variables. The variables within one group are highly correlated with each other but little correlated to variables form other groups. Thus one group can represent a factor. In our case, the number of factors was changed to 3, 5, 30, and 50.

Autoencoders (AE): Artificial neural networks are made up of networks of neurons organized in layers. Each one of the neurons has an activation function, which combined with the input provided, creates an output. This is fed to connected neurons from the next layer. The connections between neurons are weighted and these determine if the neuron is excitatory (passes information) or inhibatory (stops information). [35].

AE are a type of neural network containing a minimum of three layers: input, hidden and output layers. As the size of the output layer is forced to be the same size as the input layer, the hidden layer encodes the information provided as input [36]. If the hidden layer has fewer neurons than the input layer, the result of the encoding process can be used as a compression of the initial data set thus reducing its dimension while maintaining relevant information. In our implementation, we made use of a sparse autoencoder (a single hidden layer) and varied the number of neurons of the hidden layer between 3, 5, 30, and 50. These components were extracted from the model and used as a reduced representation of the original set in the classification process.

Our implementation is illustrated in Figure 6. All implemented layers are fully connected. In fully connected networks, all neurons from one layer are connected to all neurons from the next layer. For the compilation of the model, we used an Ada Delta optimizer with a mean square error (MSE) loss function. The optimizer is used to tweak the weights of the neurons, according to the value of the loss function. Model fitting was performed using a batch size of 32 to avoid overfitting and within 20 epochs. These parameters were experimentally optimized from the convergence of the model using the MSE loss function.

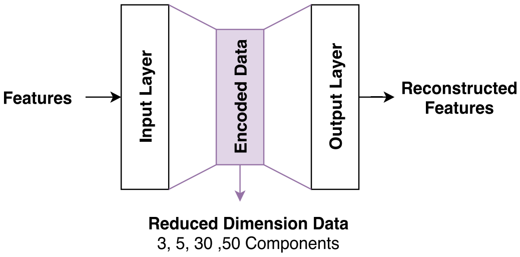

Figure 6: Dimensionality reduction with autoencoders.

Classification

Three representative classifiers have been selected for the evaluation. A decision tree based network – RF and two neural networks – MLP and LSTM Network.

Random forest (RF): Is an ensemble learning method that combines the output of several decision trees [37]. A classification decision tree is a type of predictive algorithm that uses descriptors of a problem to branch out in different directions creating new nodes. After several nodes have been created, a final decision point is reached. The final node of a tree is called a leaf. In our case, the descriptors are the features extracted from the data, while the leaves are the final sleep stage classes. Ensemble learning techniques combine the output of several predictive algorithms to provide a decision. In RF, decision trees are built by selecting a random sample of the input. The final decision is obtained by averaging the output class of all decision trees used. In our experiments, 10 decision trees were used. The minimum number of samples (descriptors) required for a decision to be made for each leaf was of minimum 10. RF has inherent feature selection implemented. Based on the decision trees created, the relevance of each feature can be determined. We used this information to analyze the most relevant features proposed which can also be used for feature selection.

Multilayer Perceptron (MLP): MLP is a type of feed forward neural network that has at least three layers: input layer, hidden layer(s) and output layer. In a feed forward network, the information flows only in one direction from input to output. The output layer represents the probability of predicting each type of class.

In our case, the input layer will contain the features while the output layers the predicted sleep stage. The number of hidden layers can be increased, creating a deep network. Each hidden layer is made up of neurons called perceptrons [38]. Variations on the number of hidden layers and layer sizes were performed. Best results were obtained when using three hidden layers of 500 perceptrons per layer. In our network architecture, all layers were fully connected and neurons had a $tanh()$ activation function. While compiling the model, an Adam optimizer was used with a binary cross-entropy loss function.

Long-short term memory (LSTM): Is a type of recurrent neural network that contains memory units and gates which makes it possible to selectively use prior temporal information for current state predictions [39, 40]. These types of networks are useful when dealing with time varying information, such as biomedical signals. For our implementation, we used a simple architecture containing a single LSTM layer with 128 units. The output layer contained five neurons representative of the five sleep stages. Compilation of the model was performed similarly to the MLP network: an Adam optimizer was used with a binary crossentropy loss function.

Evaluation

Following the best practice from literature [5, 11], each generated model was evaluated in a 10-fold cross-validation. The data set was split into 10 parts, nine were used for training and the remaining one was used for testing. When selecting the data for creating the training and testing folds, a stratified approach was used. Each fold preserved the percentages of the elements from each class. The performance was compared using the mean accuracy and $F1$-score over all sleep stage classes:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$F1_{score} = 2 * \frac{\mathrm{Recall} * \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \tag{4}$$

where TP – true positive, TN – true negative, FP – false positive, FN – false negative.

Cohen's Kappa coefficient (Kappa) was used to quantify the agreement between predicted classes and the provided annotations. This coefficient is generally used for evaluation inter-rater agreements and is defined as:

$$Kappa = \frac{Accuracy - P_e}{1 - P_e} \tag{5}$$

where P_e – probability of detection of each sleep stage determined in this case from the predicted values and the provided annotations.

Results

Effects of dimensionality reduction

Figure 7 presents an overview of the results of the variations performed for the different dimensionality reduction techniques when using six EEG channels. When applying no dimensionality reduction, the baseline difference between the three classifiers can be observed. For the automated sleep stage scoring using RF, MLP and LSTM an accuracy of 86, 67 and 71% was obtained respectively.

Results obtained from the six EEG channel input and the six EEG channels combined with the other signals were compared. The summary is available in Table 2. For the RF classifier, when using more signals and no dimensionality reduction, the performance increased to 95% accuracy. However, this is no longer valid when looking at the results obtained with MLP and LSTM. The additional features contribute negatively to the performance of these models.

Feature importance

For studying which features are more relevant to the problem of automatic sleep stage scoring, we have made use of the inherent properties of RF to assign a feature

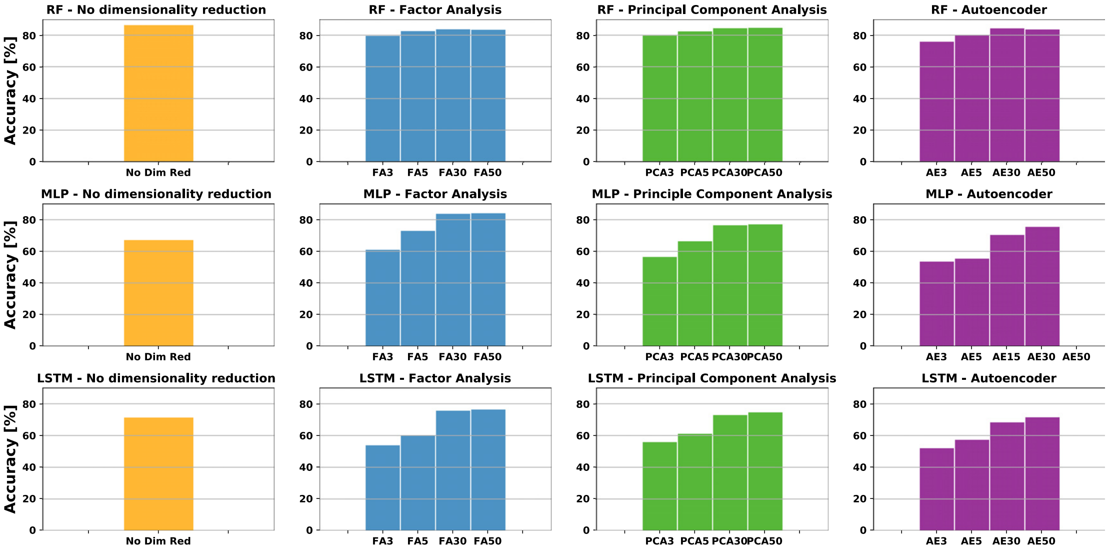

Figure 7: Effects of different dimensionality reduction techniques for six EEG channels (EEGs) on the accuracy of the models created using three classifiers. The first column is the results when using no dimensionality reduction. The second, third and fourth columns are the results when using FA, PCA and AE as dimensionality reduction. Results for RF, multilayer perceptron (MLP) and LSTM classifiers are presented in the first, second and third row respectively.

Table 2: Comparison of performance for Automated Sleep Stage Scoring using different dimensionality reduction techniques. RF – Random Forest, MLP – Multilayer Perceptron, LSTM – Long Short Term Memory, FA – Factor Analysis, PCA – Principal Component Analysis, AE – Autoencoders, Acc – Accuracy, F1 – F1-score, Kappa – Cohen's Kappa coefficient, EEGs – the six EEG channels used as input, EEGs + Sig – the six EEG channels used together with ECG, EMG and Respiration signals as input. All results are expressed in percentages.

	No dim reduction				FA			PCA			AE		
	[%]	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
RF	EEGs	86	71	63	84	84	80	85	85	81	84	84	80
	EEGs + Sig	95	86	83	90	91	88	91	92	89	85	85	81
MLP	EEGs	67	66	57	85	86	83	76	81	76	75	75	70
	EEGs + Sig	52	46	44	94	95	94	43	46	31	50	50	40
LSTM	EEGs	71	68	60	77	71	77	75	75	69	72	72	65
	EEGs + Sig	51	44	42	92	91	89	38	35	21	50	49	38

relevance (see Section 4). Since we used a variation of k-fold cross-validation (see Section 4), the most relevant feature for each fold was considered when proposing a set of selected features. When looking at the six EEG channels, the most relevant information was encoded in the frequency features of specific channels: the minimum power in the delta band of channel F3M2, C3M2 and O2M1 and in the ration between the theta and alpha band of channels C3M2, O1M2 and O2M1. When combining the EEG channels with the ECG, EMG and Respiration channels, the relevant EEG features are mixed with several respiration based features: the number of peaks detected from the signal obtained from the chest belt along with the mean, standard deviation and skew of the distance between consecutive peaks.

Discussion

For all dimensionality reduction techniques, FA, PCA, and AE, a larger number of output components, 30 or 50, resulted in a better model performance than when three or five outputs were used. No significant reduction in performance was observed for any technique. When using factor analysis or principal component analysis different projections of the data are obtained.

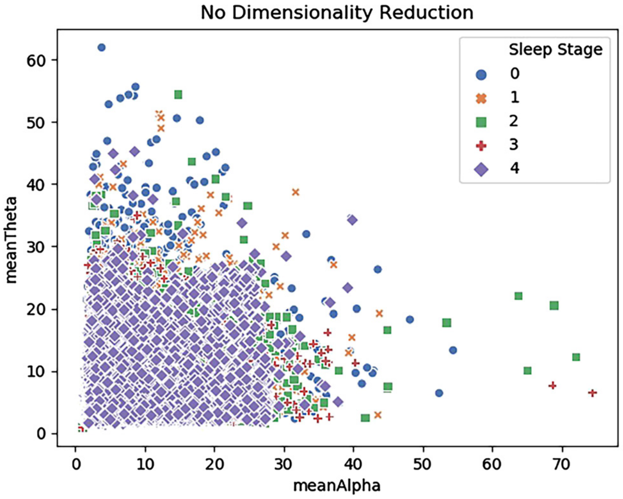

Figure 8: Representation of the meanTheta and meanAlpha values of one frontal EEG signal across all epochs in the pre-processed dataset. Sleep stages: 0 – Wake, 1 – N1, 2 – N2, 3 – N3, 4 – REM.

Figure 8 presents the values of meanAlpha plotted against the values of meanTheta obtained from the F3M2 EEG channel from all epochs in the pre-processed dataset. As alpha waves are present mostly during wakefulness, while theta waves appear during REM and light sleep, the feature values representing their mean spectral power should present a clear differentiation between at least wakefulness and stages N1, N2, and REM. Although there are some limitations in describing complex EEG frequency patterns only by the mean power spectrum of two defined bands, some differences between sleep stages are expected. From Figure 8, no clear differentiation can be observed.

Figures 9–11 present two of the components available after the transformation of the dataset through the use of PCA, FA, and the Autoencoder respectively. The pre-processed dataset was passed through the transformations which resulted in an output of 50 Components. Out of the new data representation, two components were chosen at random for the graphical representation. They were plotted for all epochs in the data set. As expected, the data is reshaped and has a lower dimension. For all data transformations, the visual separation between sleep stage classes seems clearer at visual inspection. Some components enhance the separation between specific stages of sleep. For instance, in the representation from Figure 9, stages 0 and 4 corresponding to wakefulness and REM sleep, are better separated. In Figure 10, classes 0, 2, and 4 corresponding to wakefulness, N2 and REM are better separated. When choosing other components for graphical visualization, different sleep stages can be better differentiated. In the case of dimensionality reduction through autoencoders, a less clear separation between the different stages of sleep can be observed in Figure 11. This is also reflected in the performance enhancements of the various classification models.

In some cases, a clearer separation between classes through features, results in an increase in performance. For instance, when applying factor analysis on the input to the MLP classifier, an increase of almost 20% in accuracy is obtained. This is also present when using the LSTM classifier and factor analysis, an increase of 5% is

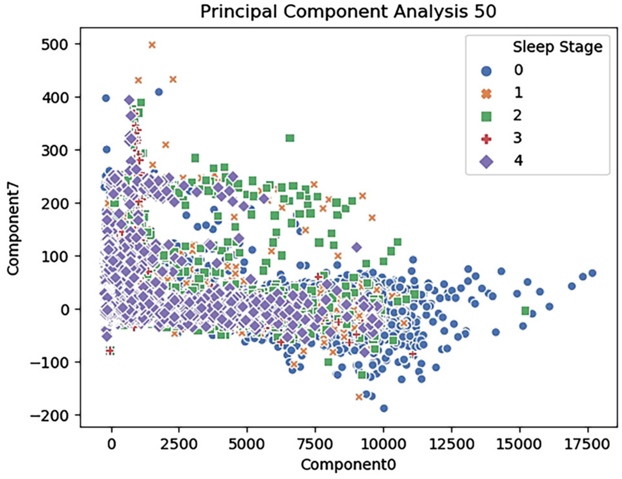

Figure 9: Representation of Component 7 vs. Component 0 of the principal component analysis (PCA) transformation with an output of 50 Components. Sleep stages: 0 – Wake, 1 – N1, 2 – N2, 3 – N3, 4 – REM.

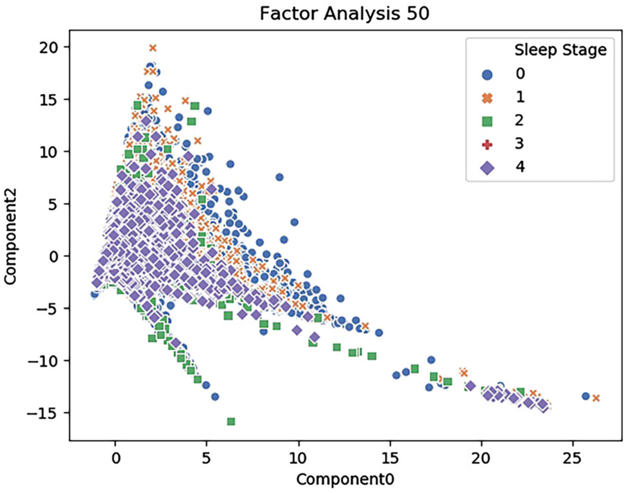

Figure 10: Representation of Component 2 vs. Component 0 of the factor analysis (FA) transformation with an output of 50 Components. Sleep stages: 0 – Wake, 1 – N1, 2 – N2, 3 – N3, 4 – REM.

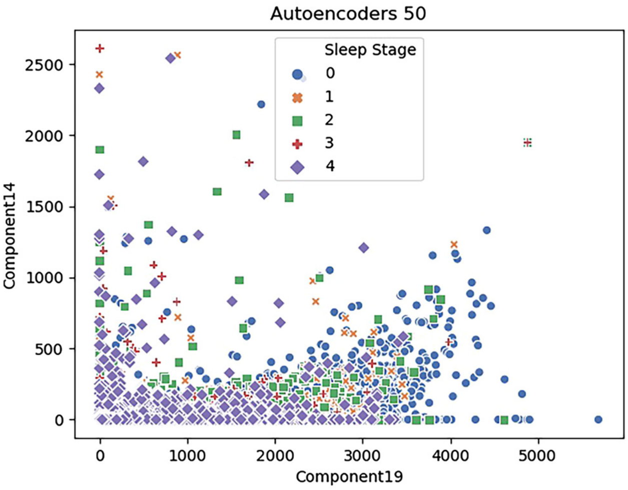

Figure 11: Representation of Component 19 vs. Component 14 of the Autoencoder dimensionality reduction with an output of 50 Components. Sleep stages: 0 – Wake, 1 – N1, 2 – N2, 3 – N3, 4 – REM.

observed. A similar increase is also obtained when using PCA prior to MLP and LSTM. When using autoencoders as a dimensionality reduction method, the performance remains in the same range as when no reduction is applied. While looking at the results of the RF classifiers, the obtained accuracies are in approximately the same range with a decrease of maximum 5% in accuracy.

In case of RF, that has inherent feature selection, the performance is increased as adequate input is selected. Although the performance is significantly lower when using MLP and LSTM with the combination of EEG and other signals, when applying factor analysis, the performance is significantly improved. Factor analysis seems the most suitable method for dimensionality reduction when using MLP and LSTM networks. When using autoencoders, the performance remains in the same range for all type of classifiers. Applying dimensionality reduction techniques for the selected PSG dataset, the accuracy of the obtained models using the three selected classifiers remains the same or in some cases it is even improved when compared to the case when no dimensionality reduction is applied.

The importance and selection of features is relevant when dealing with a large number of features used as input. The aim of feature selection is to reduce the redundancy of information and eliminate features that might have a negative impact on performance. This is highlighted by our results, where the RF classifier that inherently assigns a feature relevance (see Section 6) performs significantly and consistently better than neural networks without feature selection.

Overall, the performance obtained with the proposed methods is in range with the performance of other methods that can be found in literature. Alickovic et al. [13] obtained a higher accuracy of 97.25% using a subject specific sleep stage classification method applied to 20 subjects from the Sleep EDF dataset, on a single channel EEG. The approach used a discrete wavelet transform for feature extraction and PCA for dimensionality reduction prior to building a model with a support vector machine classifier. In comparison, our method has a slightly lower performance however we have created a general, non-patient specific model based on a large dataset. The advantage of using a larger dataset is the potential of capturing more variation in the patient population which can lead to a better discriminative power when presented with new data.

Biswal et al. [7] uses an extended version of the same Physionet Challenge sleep dataset. Their best result was of 85.76% accuracy when using expert-defined features extracted from six EEG channels as input to a recursive neural network classifier. Kuo et al. [41] uses the same sleep stage detection dataset for automated sleep scoring. All PSG signals are used as input to a bi-directional LSTM based classification algorithm. Within a hold out validation strategy, where 50% of the data is used for training and 50% of data is used for test, an accuracy of 82.9% is obtained. When comparing our results with other methods applied to the same dataset, we have obtained a performance increase of 10% when using features extracted from six EEG signals, ECG, EMG and respiratory signals with a random forest classifier.

A limitation of our approach is the use of a single dataset for training and testing. Although variations between patients are captured, different data sets might bring forth variations in the recorded physiological signals due to the different recording equipment used. Although the AASM guideline unifies the views on sleep architecture content, there is some degree of variability in the annotators' interpretation of the PSG signals and therefore in the provided annotations. The typical inter-scorer agreement is around 80% [42]. Expert sleep scoring is used as the ground truth in training machine learning algorithms. in general, the performance of automated sleep stage scoring methods is limited by the quality of the provided annotations and of the inter-scorer agreement.

Conclusions

In this paper, we have investigated several dimensionality reduction techniques for a PSG dataset applied to the problem of automated sleep stage detection. Autoencoders are efficient in maintaining a similar performance to when no dimensionality reduction is applied. However, statistical transformations such as factor analysis and principal component analysis can in some cases enhance the performance. The accuracy of the trained models is dependent not only on the input dataset and the dimensionality reduction applied, but also on the type of classifier used.

The highest performance for automated sleep scoring was of 95% accuracy when using all input signals (six EEG channels combined with ECG, EMG and Respiratory signals) with the RF classifier and applying no dimensionality reduction method. The MLP and LSTM classifiers significantly underperformed with respect to RF when no dimensionality reduction was used. When FA was applied their accuracy reached 94 and 92% respectively, significantly improving the performance. Dimensionality reduction techniques help in reshaping the input data such that computational power is reduced, and for some transformations the performance is increased.

The architecture used for the autoencoder was simple. Iterations on the architecture and parameters of the autoencoder can lead to a better representation of the data and thus increasing the performance. The networks used for the classification can also be further optimized. For instance, LSTM networks make use of temporal information. Sleep patterns as recorded through EEG signals show high variations over time. LSTM can be used to capture long and short-term dependencies within sleep patterns. Our future work will focus on exploiting time-based information as well for automatic sleep stage classification.

Research funding: None declared.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.

References

1. Berry Brooks RB, Gamaldo R, Harding CE, Lloyd SM, Quan RM, Troester SF, et al. The AASM manual for the scoring of sleep and associated events. Version 2.4. Darien: American Academy of Sleep Medicine; 2017.
2. Louis Erik K, Boeve FB. REM sleep behavior disorder: diagnosis, clinical implications and future directions. Mayo Clinic Proc;11: 1723–36.
3. Fiorillo L, Puiatti A, Papandrea M, Ratti PL, Favaro P, Roth C, et al. Automated sleep scoring: a review of the latest approaches. Sleep Med Rev 2019;48. https://doi.org/10.1016/j.smrv.2019.07.007.
4. Rahman MM, Bhuiyan MIH, Hassan AR. Sleep stage classification using single-channel EOG. Comput Med Biol 2018; 102:211–20.
5. Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv 2016, arXiv:1610.01683.
6. Hassan AR, Subasi A. A decision support system for automated identification of sleep stages from single-channel EEG signals. Knowl Base Syst 2017;128:115–24.
7. Biswal S, Kulas J, Sun H, Goparaju B, Westover MB, Bianchi MT, et al. SLEEPNET: automated sleep staging system via deep learning. arXiv 2017, arXiv:1707.08262v1.
8. Zhao D, Wang Y, Wang Q, Wang X. Comparative analysis of different characteristics of automatic sleep stages. Comput Methods Progr Biomed 2019;175:53–72.
9. Aboalayon K, Faezipour M, Almuhammadi W, Moslehpour S. Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation. Entropy 2016;18:272.
10. Memar P, Faradji F. A novel multi-class EEG-based sleep stage classification system. IEEE Trans Neural Syst Rehabil Eng 2017; 26:84–95..
11. Tautan A-M, Rossi AC, De Franciso R, Ionescu B. Automatic sleep stage detection using a single channel frontal EEG. In: The 7th IEEE International Conference on E-Health and Bioengineering – EHB Iasi, Romania: IEEE; 2019.
12. Hassan AR, Bhuiyan MIH. A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. J Neurosci Methods 2016;271:107–18.
13. Alickovic E, Subasi A. Ensemble SVM method for automatic sleep stage classification. IEEE Trans Instrum Meas 2018;667: 1258–65.
14. Hassan AR, Bhuiyan MIH. Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting. Comput Methods Progr Biomed 2017;140:201–10.
15. Yang Y, Zheng X, Yuan F. A study on automatic sleep stage classification based on CNN-LSTM. In: Proceedings of the 3rd International Conference on Crowd Science and Engineering Singapore, New York, NY, USA: Association for Computing Machinery; 2018. pp. 1–5.
16. Vilamala A, Madsen KH, Hansen LK. Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring. arXiv 2018, arXiv:1710.00633.
17. Patanaik A, Ong JL, Gooley JJ, Ancoli-Israel S, Chee MWL. An end-to-end framework for real-time automatic sleep stage classification. Sleep 2018:41. https://dx.doi.org/10.1093%2Fsleep%2Fzsy041.
18. Sun Y, Wang B, Jin J, WAng X. Deep convolutional network method for automatic sleep stage classification based on neurophysiological signals. In: 11th International congress on image and signal processing. Beijing, China: BioMedical Engineering and Informatics; 2018.
19. Van Der Maaten LJP, Postma EO, Van Den Herik HJ. Dimensionality reduction: a comparative review. J Mach Learn Res 2009;10:1–41. https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf.
20. Khosla N. Dimensionality reduction using factor analysis. Australia: Griffith University Queensland; 2006. Thesis (Masters).
21. Fan Y. Research on feature extraction of EEG signals using MSE-PCA and sleep staging. IEEE International Conference on

Signal Processing, Communications and Computing,Qingdao, China: ICSPCC, IEEE; 2018,2. p. 1–5.

22. Najdi S, Gharbali AA, Fonseca JM. Feature transformation based on stacked sparse autoencoders for sleep stage classification. IFIP Int Fed Inf Process 2017;499:144–53.
23. Prabhudesai KS, Collins LM, Mainsah BO. Automated feature learning using deep convolutional auto-encoder neural network for clustering electroencephalograms into sleep stages. In: 9th International IEEE EMBS Conference on Neural Engineering, San Francisco, CA, USA: IEEE; 2019.
24. Najdi S. Feature extraction and selection in automatic sleep stage classification. Lisbon: Universidade Nova de Lisboa; 2018. Master Thesis.
25. Ghassemi MM, Moody BE, Lehman LWH, Song C, Li Q, Sun H, et al. You snooze, you win: the PhysioNet/computing in cardiology challenge 2018. Comput Cardiol 2018:20–3. https://doi.org/10.22489/CinC.2018.049.
26. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circ J Am Heart Assoc 2000;23:101.
27. Colten HR, Altevogt BM. Sleep disorders and sleep deprivation: an unmet public health problem. Washington, D.C.: The National Academies Press; 2006. p. 34–9.
28. Chouchou F, Desseilles M. Heart rate variability: a tool to explore the sleeping brain?. Front Neurosci 2014;8:402.
29. Sedghamiz H. Matlab implemenation of Pan Tompkins ECG QRS detector. “msc3”; 2014.
30. Pan J, Tompkins WJ. Pan Tompkins 1985–QRS detection. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 1985;32:230–6.
31. Van Steenkiste T, Groenendaal W, Ruyssinck J, Dreesen P, Klerkx S, Smeets C, et al. Systematic comparison of respiratory signals for the automated detection of sleep apnea. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS; Honolulu, HI, USA; 2018. pp. 449–52.
32. Fonseca P, Long X, Radha M, Haakma R, Aarts RM, Rolink J. Sleep stage classification with ECG and respiratory effort. Physiol Meas 2015;36:2027–40.
33. Tautan A-M, Rossi AC, De Franciso R, Ionescu B. Automatic sleep stage detection: a study on the influence of various PSG input signals. In: 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society, Montreal, QC, Canada; EMBC; 2020.
34. Ross D, Lim J, Lin R-S, Yang M-H. Incremental learning for robust visual tracking. Int J Comput Vis 2008;77:125–41.
35. Hagan MT, Demuth HB, Beale MH, De Jesus O. Neural network design, 2nd ed, Sillwater, Oklahoma, USA: Martin Hagan; 1995.
36. Bengio Y. Learning deep architectures for AI trends in machine learning. Found Trends Mach Learn 2009;1:1–127.
37. Schroff F, Criminisi A, Zisserman A. Object class segmentation using random forests. In: Proceedings of the British machine vision conference; 2008.
38. Haykin SO. Neural networks: a comprehensive foundation, 2nd ed., Delhi, India: Pearson Education (Singapore) Pte. Ltd.; 1999.
39. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;8:1735–80.
40. Chetlur S, Woolley C, Vandermersch P, Cohen J, Tran J, Catanzaro B, et al. cuDNN: efficient primitives for deep learning. arXiv 2014, arXiv:1410.0759v3.
41. Kuo CE, Chen GT, Lin NY. Automatic sleep staging using deep long short-term memory: validation in large-scale datasets. In: Proceedings of the 2019 3rd International conference on computational biology and bioinformatics, Nagoya, Japan; ACM; 2019. pp. 58–64.
42. Rosenburg RS, Van Hour S. The American Academy of sleep medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9:81–7.