Journal of Neuroscience Methods xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Computational Neuroscience

Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines

Tarek Lajnef[a], Sahbi Chaibi[a], Perrine Ruby[b], Pierre-Emmanuel Aguera[b], Jean-Baptiste Eichenlaub[c], Mounir Samet[a], Abdennaceur Kachouri[a,d], Karim Jerbi[b,e,*]

[a] Sfax National Engineering School (ENIS), LETI Lab, University of Sfax, Sfax, Tunisia
[b] DYCOG Lab, Lyon Neuroscience Research Center, INSERM U1028, UMR 5292, University Lyon I, Lyon, France
[c] Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA
[d] Higher Institute of Industrial Systems of Gabes (ISSIG), University of Gabes, Gabes, Tunisia
[e] Psychology Department, University of Montreal, QC, Canada

ARTICLE INFO

Article history:
Received 20 June 2014
Received in revised form 15 January 2015
Accepted 16 January 2015
Available online xxx

Keywords:
Electroencephalography (EEG)
Sleep scoring
Oscillations
Polysomnography
Decision-tree
Support vector machine (SVM)
Linear Discriminant Analysis (LDA)
Hierarchical clustering
Machine learning
Dendrogram

ABSTRACT

Background: Sleep staging is a critical step in a range of electrophysiological signal processing pipelines used in clinical routine as well as in sleep research. Although the results currently achievable with automatic sleep staging methods are promising, there is need for improvement, especially given the time-consuming and tedious nature of visual sleep scoring.
New method: Here we propose a sleep staging framework that consists of a multi-class support vector machine (SVM) classification based on a decision tree approach. The performance of the method was evaluated using polysomnographic data from 15 subjects (electroencephalogram (EEG), electrooculogram (EOG) and electromyogram (EMG) recordings). The decision tree, or dendrogram, was obtained using a hierarchical clustering technique and a wide range of time and frequency-domain features were extracted. Feature selection was carried out using forward sequential selection and classification was evaluated using *k*-fold cross-validation.
Results: The dendrogram-based SVM (DSVM) achieved mean specificity, sensitivity and overall accuracy of 0.92, 0.74 and 0.88 respectively, compared to expert visual scoring. Restricting DSVM classification to data where both experts' scoring was consistent (76.73% of the data) led to a mean specificity, sensitivity and overall accuracy of 0.94, 0.82 and 0.92 respectively.
Comparison with existing methods: The DSVM framework outperforms classification with more standard multi-class "one-against-all" SVM and linear-discriminant analysis.
Conclusion: The promising results of the proposed methodology suggest that it may be a valuable alternative to existing automatic methods and that it could accelerate visual scoring by providing a robust starting hypnogram that can be further fine-tuned by expert inspection.

1. Introduction

Sleep is characterized by continuous changes in brain, eye, muscle, respiratory and heart beat activity. These changes are monitored with polysomnographic recordings, which measure, during a full night of sleep, different types of physiological data typically including the electroencephalogram (EEG), electro-oculogram (EOG), electromyogram (EMG) and electrocardiogram (ECG). Physiologically speaking, sleep states are split into two broad types: rapid eye movement (REM sleep) and non-rapid eye movement (non-REM sleep). The latter consists of 4 stages (S1, S2, S3 and S4). These distinct sleep stages are associated with distinct physiological and neuronal features which are generally used to identify the sleep stage a person is in. This process called sleep scoring, or sleep staging, is a critical step in a range of electrophysiological signal processing pipelines used in clinical routine as well as in sleep research.

In clinical routine, sleep studies are usually performed for the diagnosis of pathologies, such as insomnia, hypersomnia, circadian rhythm disorders, epilepsy and sleep apnea. Sleep scoring often relies on visual analysis of the recordings to establish a hypnogram that depicts in time the different sleep stages. The analysis generally follows established guidelines for sleep stage classification, such as

* Corresponding author at: Psychology Department, University of Montreal, QC, Canada. Tel.: +1 438 969 2103.
E-mail address: karim.jerbi@umontreal.ca (K. Jerbi).

http://dx.doi.org/10.1016/j.jneumeth.2015.01.022
0165-0270/

the ones introduced by Rechtschaffen and Kales (1968), where each segment of 30 s is labelled as Awake, S1–S4 or REM. A more recent classification manual proposed by the American Academy of Sleep Medicine (AASM) in 2007 (Iber et al., 2007), combines the non-REM stages S3 and S4 into a single stage of deep sleep (called N3), also known as slow-wave sleep (SWS). Both manuals propose to use EEG derivations (2 in the R&K manual, and 3 in the AASM one), 2 EOG electrodes and one EMG electrode.

While visual scoring remains the gold-standard, recent years have witnessed a surge in method developments for automatic or semi-automatic sleep staging (e.g. Agarwal and Gotman, 2001, 2002; Becq et al., 2005; Berthomier et al., 2007; Ma et al., 2011; Itil et al., 1969; Koley and Dey, 2012; Krakovska and Kristina, 2011; Larsen and Walter, 1970; Schaltenbrand et al., 1996; Sheng-Fu et al., 2012; Shing-Tai et al., 2012; Stanus et al., 1987; Steinn et al., 2005; Huang et al., 2014). Although these results obtained so far are promising, there is still room and a need for improvement, especially given the time-consuming and tedious nature of visual sleep scoring. Across existing methods, a wide range of physiological signatures, or features, have been extracted from polysomnographic (PSG) signals, including time-domain, frequency-domain, and time–frequency-domain features, and both linear and non-linear features have been explored. While some studies rely only on one or two features to perform sleep stage classification (e.g. Fraiwan et al., 2010; Šušmáková and Krakovská, 2008), several studies provide evidence for the utility of searching for an optimum combination of features (e.g. Grozinger et al., 2001; Sheng-Fu et al., 2012; Khalighi et al., 2013).

Beyond the specific electrophysiological features used, existing methods also differ in the type of classification framework used. Some machine learning techniques such as artificial neural networks have been widely used for sleep staging (Kerkeni et al., 2012; Marina et al., 2012; Ronzhina et al., 2012; Ma et al., 2011). A disadvantage of this method is the fact that the exact decision procedure remains hidden or implicit. Classification methods based on Bayesian probability (linear and quadratic discrimination, *k*-nearest neighbour), have also been used in sleep scoring (Fraiwan et al., 2010; Krakovska and Kristina, 2011) The requirement of a Gaussian distribution of data in these methods can sometimes be a limitation. Other approaches for automatic sleep scoring based on mathematical modeling and hidden Markov Models have also been proposed (Doroshenkov et al., 2007; Shing-Tai et al., 2012). Support vector machines (SVM) classification has also been used for sleep scoring (Steinn et al., 2005; Koley and Dey, 2012). Support vector machines, introduced in the early 90s (Boser et al., 1992; Cortes and Vapnik, 1995) are used in a wide range of learning problems such as pattern recognition, text categorization and medical diagnosis and they continue to draw a lot of attention in many fields including basic and clinical neuroscience.

In this paper, we propose a sleep staging procedure that achieves multi-class classification by embedding multiple SVMs in a decision tree (dendrogram) framework. The dendrogram-SVM method is applied to polysomnography data from 15 individuals and its performance is compared to expert visual scoring. In addition, to assess the added value of the proposed methodology, its results were benchmarked against two standard classification methods, which are "one-against-all" SVM and linear-discriminant analysis (LDA). In brief, the proposed classification pipeline, which is described in more detail below, consists of three main steps: (i) feature extraction from all EEG, EOG and EMG data of 15 individuals (covering both time and frequency domain features, combining linear and non-linear measures), (ii) dimension reduction and feature subset selection using forward selection and cross-validation within the training step and (iii) classification using a multi-class SVM based on a decision tree obtained via ascendant hierarchical clustering (AHC).

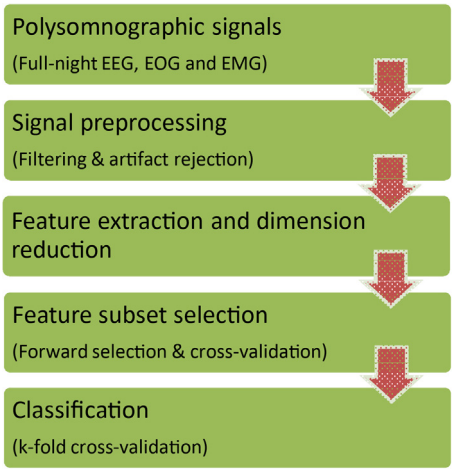


Fig. 1. Overview of the sleep stage classification pipeline.

In the following, we describe the full pipeline and the individual steps in detail. We then present the performance of the proposed DSVM method as measured by sensitivity, specificity and accuracy (using expert scoring as ground truth). Next, we report the comparison between the DSVM framework and standard "one-against-all" SVM and LDA applied to the same polysomnographic data set. Finally, we discuss the components of our method that could explain its higher performance compared to standard methods and also address potential limitations and ideas for further improvements.

2. Materials and methods

2.1. Polysomnographic data base

The data used in this study consists of polysomnographic (PSG) records in 15 healthy subjects aged 29.2 ± 8 years, which were collected at the DyCog Lab of the Lyon Neuroscience Research Center (Lyon, France) as part of a larger study exploring cognition during sleep (Eichenlaub et al., 2012, 2014; Ruby et al., 2013a,b). Each record contains EOG, EMG and 21 scalp-EEG channels. The EEG electrodes were positioned according to the International 10–20 system, the EOG electrodes were placed diagonally on the outer edges of the eyes, the EMG electrodes were positioned on the chin. All signals were recorded with a sampling frequency of 1000 Hz. The 15 PSG sleep recordings were visually scored by an expert in successive windows of 30-s using the R&K guidelines.

2.2. Methods

The sleep stage classification process can be divided in 5 distinct steps (Fig. 1). Once the polysomnographic signals are acquired, the EEG, EMG and EOG signals were filtered and all segments contaminated by artefacts were excluded (through a combination of automatic thresholds and visual inspection). This pre-processing step is comparable to standard EEG pre-processing and was fast (only taking a few hours to complete for all data sets). Different time and frequency domain features were calculated (see detailed below) and the most relevant ones were identified by straightforward statistical analysis (*t*-tests with *p* values revealing significant modulation by sleep stage). This is referred to here as the feature-space dimension reduction step. Next, the most discriminating features subsets were selected using a standard sequential forward selection procedure: Starting from the feature that provides the highest accuracy, we continue sequentially searching for the next feature that will then collectively provide the highest increase in

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

performance, the procedure is stopped as soon as the multi-feature decoding accuracy starts to drop. Note that, because the forward selection technique relies on cross-validation, it is performed by splitting the training data set itself into a training subset and a test subset (10-fold cross-validation). This is done to rule out any risk of over-fitting or over-learning in the feature selection step. Finally, sleep scoring is performed by feeding the selected features into a multi-class dendrogram-based SVM (i.e. based on a decision tree approach). This procedure is carried out at each repetition (also known as a *fold*) of the training–testing, where the overall data set is divided into a training set (10 individuals) and a test set (5 individuals).

All signal pre-processing, time–frequency analyses and data visualizations described throughout this study were performed using custom code written in MATLAB (Mathworks Inc., MA, USA). Various data format conversions, visual inspections and expert sleep scoring (hypnogram) were performed with in-house software package for electrophysiological signal analysis (ELAN) developed at INSERM U1028, Lyon, France (Aguera et al., 2011).

2.2.1. Signal pre-processing and visual sleep stage scoring

EEG signals are typically contaminated by a number of artefacts (muscle artefact, ECG, eye movements and blinks). In order to minimize the effect of these phenomena in our study EEG signals were filtered with a band pass filter with cutoff frequencies at 0.2 and at 40 Hz. All polysomnographic signals were then segmented into 30 s epochs in line with the segmentation used in sleep scoring standards. Data segments with any remaining prominent artefacts were excluded from subsequent analysis.

Since expert visual scoring will be used to benchmark the results of the automatic classification method, sleep stages were first visually scored offline according to standard criteria (Silber et al., 2007) by two experts to derive hypnograms based on 30-s epochs to determine the vigilance state (wake, S1, S2, SWS, or REM). The percentage of consistency between the two experts scoring was 76.73% with a kappa coefficient of 0.71 (epoch-by-epoch comparison).

2.2.2. Features extraction

It has been shown in previous reports that sleep stages classification can be achieved using a wide-range of polysomnographic parameters (e.g. Šušmáková and Krakovská, 2008; Grozinger et al., 2001). The features used can broadly be divided into linear (including time and frequency domain features) and non-linear measures. While the time-domain and non-linear features are computed directly from the time series of the signals, the frequency domain features used here were extracted from the signal's power spectral density (PSD) estimation. Note that the linear and non-linear time-domain features as well as the spectral entropy were computed for all channels (C3, Cz, EOG1, EOG2 and EMG), however, the frequency-domain features we used were computed only for EEG signals. In total, 102 features were extracted for each class (i.e., sleep stage) yielding a $N \times 102$ matrix, where N is the number of epochs in a given class.

2.2.3. Linear features

2.2.3.1. Linear prediction error energy. Linear prediction (LP) is a signal processing technique which consists of estimating a future sample of a discrete time signal from the previous sample. It has been used in a variety of applications including the analysis of speech and biomedical signals (Lajnef et al., 2010). A linear predictive filter estimates the spectral characteristics of the signal window by calculating the coefficients of a FIR filter. This estimation is an optimization process that involves calculation of filter coefficients to achieve minimum modelling error (Altunay et al., 2010). Mathematically, a linear predictive filter is defined by: $y[n] = \sum_{k=1} a_k s(n-k)$, where $y[n]$ represent samples predicted by the LP filter, a_k are filter coefficients and $s[n]$ are the time series samples. Note that a_k values are determined by minimizing the variance σ_e^2 of the error $e[n]$, where the latter is defined as: $[n] = s[n] - y[n] = s[n] - \sum_{k=1} a_k s(n-k)$, and its variance σ_e^2 is defined by $\sigma_e^2(a_k) = 1/N \sum_{n=0}^{N-1} e^2[n] = 1/N \sum_{n=0}^{N-1} \left(s[n] - \sum_{k=1} a_k s(n-k)\right)^2$. The error energy of each epoch is then calculated as: $E = \sum e^2$.

2.2.3.2. Time domain features. Many time-domain features have been used in the literature (Krakovska and Kristina, 2011). These features mainly consist of statistical measures applied directly to the time series. The list of time-domain features we used in this study is listed in Table 1.

2.2.3.3. Frequency domain features. The frequency-domain features that we used consist of total and relative spectral power, power ratios and spectral entropy. All these were extracted from power spectral density (PSD) estimation. Several methods for estimating the power spectrum exist and are generally broadly divided into parametric and non-parametric methods (Krakovska and Kristina, 2011). We applied the widely used Welch's averaged periodogram method (Oppenheim and Schafer, 1999). Each 30-s epoch (i.e. 30,000 samples at 1000 Hz sampling) were divided into six non-overlapping segments (5000 samples each) to which we applied a Hamming window. The final spectral density was achieved as the average of the spectral densities of all six segments. Moreover, we computed the power in five distinct frequency bands by averaging the power at each frequency bin of the given intervals. The five bands were: delta (δ, 0.5–4.5 Hz), theta (θ, 4.5–8.5 Hz), alpha (α, 8.5–11.5 Hz), sigma (σ, 11.5–15.5 Hz), beta (β, 15.5–32.5 Hz).

We then calculated the following frequency-domain features for each epoch in all individuals: (a) Total power, (b) Relative power (P_{rel}) in each of five frequency bands (dividing absolute power in each band by the sum of powers across all frequencies over the whole segment), (c) Power ratios in all 16 combinations across the five frequency bands (e.g. delta/alpha, delta/beta, delta/sigma, alpha/beta, alpha/theta, etc.) and finally (c) Spectral entropy (*Sen*), a measure of the regularity of the signal, which was introduced in 1996 by Pardey et al. (1996) and which can be computed from the relative power P_{rel} as follows: $Sen = -(1/\log N) \sum_{f=1}^{N} P_{rel}(f) * \log P_{rel}(f)$, where N is the total number of frequency bins and f is the value of each frequency bin. The spectral entropy of a pure sine wave is zero and that of uncorrelated white noise is one. This measure has been used in a wide range of applications, such as in the assessment of the depth of sedation from EEG recordings (Ferenets et al., 2006).

2.2.4. Non-linear features

2.2.4.1. Permutation entropy. Permutation entropy is a non-linear measure that characterizes the complexity of time series (Bandt and Pompe, 2002). It has been applied in order to monitor the depth of anaesthesia from EEG signals (Olofsen and Sleigh, 2008; Shalbaf et al., 2013) and has been shown to be a promising tool to reveal abnormalities of cerebral activity in patients with absence epilepsy (Ferlazzo et al., 2014). Like any other measures of entropy, permutation entropy is a convenient measure of regularity, complexity or flattening in the frequency distribution. Permutation entropy accounts for the temporal information contained in a time series and comes with very low computational costs (Zanin et al., 2012). When the EEG is dominated by high frequencies, the entropy is maximal (close to 1) and conversely, when the signal is dominated by low frequencies the entropy takes its minimum values (Bandt and Pompe, 2002).

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

Table 1
Full list of the time domain features extracted from the data.

Name	Definition	Formula	
Var	Variance (characterizes the dispersion of a distribution or a sample)	$var = \frac{1}{n-1}\sum_{i=1}^{N}(x(i)-\bar{x})^2$	(5)
Std	Standard deviation (square root of the variance)	$std = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x(i)-\bar{x})^2\right]^{1/2}$	(6)
RMS	Root mean square (square root of the arithmetic mean of the squares of the original values)	$RMS = \left[\frac{1}{n-1}\sum_{i=1}^{n}x(i)^2\right]^{1/2}$	(7)
Kurt	Kurtosis coefficient (a measure of the compression degree of a distribution)	$K = \frac{E(x-\bar{x})^4}{[E(x-\bar{x})^2]^4}$	(8)
Skew	Skewness (measures the degree of asymmetry of the distribution)	$S = \frac{E(x-\mu)^3}{\left[\sqrt{E(x-\mu)^2}\right]^3}$	(9)
Per75	Percentile75 (value below which 75% of the observations fall)	$card\left\{y(i)/y(i) < prctile75eeg\right\} = \frac{75n}{100}$	(10)

The principle of permutation entropy is based on the idea that by evaluating the frequency of occurrence of permutation patterns of the elements of a time series, one can infer information about the dynamics of the system at hand. Critically, the probability with which each possible pattern is present can reveal information about the dynamics of the signal. More generally, to each time series it is possible to associate a probability distribution Π, whose elements π_i are the frequencies associated with the i possible permutation patterns ($i = 1,\ldots,N!$ where N is known as the embedding dimension). The permutation entropy (PE) is defined as the Shannon entropy associated to such distribution: $PE = -\sum_{i=1}^{N!}\pi_i \log \pi_i$. A detailed explanation of permutation entropy can be found in a comprehensive review by Zanin et al. (2012) and the seminal paper by Bandt and Pompe (2002).

2.2.4.2. Teager energy operator. The second non-linear feature we used is the Teager energy operator (TEO), also known as the Teager–Kaiser energy operator. It is a non-linear quadratic operator initially introduced by Kaiser (1990) to measure the real physical energy of a system. One of its advantages is that it allows for the detection of instantaneous changes in the signal such as discontinuities, increases or decreases of amplitude and frequency. TEO has been used in numerous signal processing applications (e.g. Erdamar et al., 2012). The discrete-time TEO for real-valued signal x is given by $E(n) = x^2(n) - x(n-1)x(n+1)$.

2.3. Feature pre-processing and dimension reduction

Once all features are extracted, we apply a two-step pre-processing stage where we first search for outliers (features with values twice higher than the standard deviation of all values of the same feature in the same class). Second, in order to reduce the dimension of the feature space, we excluded features which, upon statistical examination with standard t-test, appeared to be the least discriminant between classes: we ran t-test statistical testing to compare the mean of each feature across all pairs among the 5 stages (Awake, S1, S2, SWS and REM). In other words, 15 t-tests were performed for each feature. Each time a feature achieved significance in any of the comparisons its score was increased by 1. The top 32 features (listed in Table 2) were kept for further analysis. This analysis step was common to all methods tested here. Its aim is to reduce the feature space to 32 features (from a total of 102 features extracted). We chose this univariate statistical approach as a generic and straight-forward dimensionality reduction step. Dimensionality reduction is required for the subsequent multiple feature selection procedure. Of course, it is theoretically conceivable that some features that are relevant in the context of multi-feature classification get dropped at this dimension reduction step. Alternative tools, including multivariate statistics could be considered in the future.

2.4. Multi-class SVM classification

Numerous reports in the literature provide evidence for the high performance of SVM in particular for high dimensional classification problems (Huang et al., 2002; Melgani and Bruzzone, 2004). In principle, SVMs are designed for binary classification problems (discrimination between two classes), however, as in many classification tasks, automatic sleep scoring requires discrimination between multiple classes (Awake, S1, S2, SWS and REM). Therefore, if one wants to benefit from the putative advantages of SVM classification, a multi class SVM framework needs to be implemented. Two of the most widely used approaches for multi-class SVM classification are the *One-Against-All* (OAA) and the *One-Against-One* (OAO) approaches. The OAA framework consists of using a binary SVM to distinguish each class from all other classes and the decision is obtained by applying a winner-takes-all strategy. By contrast, in the OAO multi-SVM approach a dedicated classifier is trained for each of all possible pairs of classes. In other words, for a total of Q classes, one would need to train $n = C_Q^2$ classifiers and the decision is then obtained by performing a majority vote (max-wins voting).

Table 2
List of selected features after the t-test procedure.

1. Perm_entropy_EMG	12. delta/sigma_C3	23. beta/alpha_C3
2. Perm_entropy EOG1	13. theta/delta_Cz	24. beta/sigma_C3
3. Rel Power delta_C3	14. theta/alpha_Cz	25. sigma/delta_C3
4. Rel Power theta_Cz	15. theta/beta_Cz	26. sigma/theta_C3
5. RelPower alpha_C3	16. theta/sigma_Cz	27. sigma/alpha_C3
6. Rel Power beta_C3	17. alpha/delta_C3	28. sigma/beta_C3
7. Rel Power sigma_C3	18. alpha/theta_C3	29. Kurtosis_EOG1
8. Spectral entropy_Cz	19. alpha/beta_C3	30. Per75_EOG1
9. delta/theta_C3	20. alpha/sigma_C3	31. Kurtosis_EOG2
10. delta/alpha_C3	21. beta/delta_C3	32. Per75_EOG2
11. delta/beta_C3	22. beta/theta_C3	

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

In this study, however, we propose to implement and evaluate a different multi-class strategy: a decision-tree-based support vector machine approach for sleep stage classification. Hierarchical clustering is used to design an optimal hierarchical structure of the decision tree. The rationale here is that associating decision tree architecture with binary SVMs combines the advantages of the efficient computation of decision trees and the high classification accuracy of SVMs.

2.4.1. Dendrogram multi-class SVM

Although they are not as well-known and established as OAO and OAA multi-class SVM methods, decision-tree-based multi-SVM classification has been explored in the machine learning and computer science literature (Takahashi and Abe, 2002; Benabdeslem and Bennani, 2006; Gjorgji et al., 2009; Bala and Agrawal, 2011). Here we use a decision-tree SVM classification method which will refer to in the following as Dendrogram-SVM (DSVM). A dendrogram (from the greek word *dendro*, which means *tree*) generally refers to a tree diagram that illustrates the arrangement of nodes produced by hierarchical clustering. Dendrograms are for example often used in computational biology to illustrate the clustering of genes or samples. Simply put, this method is a decision-tree classification framework where each binary classification node is fulfilled by a binary SVM. The first step of the procedure applied consists therefore in defining the tree with its binary branchings, i.e. the structure of the dendrogram. This was done here by computing the mean values of the features for each class (center of gravity), to which we then applied ascendant hierarchical clustering (AHC). Hierarchical clustering is a cluster analysis method which seeks to build a hierarchy of clusters; The ascendant (or agglomerative) hierarchical clustering technique is a bottom-up approach (as opposed to the top-down descendant or divisive hierarchical clustering method). In the ascendant approach used here, observations that stem from each final cluster (class) are sequentially merged as we move up the hierarchy (Hastie, 2009; Tuffery, 2011). The dendrogram that resulted from the AHC analysis was then used as the backbone for the multi-class classification with a distinct binary SVM assigned to each node.

2.4.2. SVM parameter optimization

The performance of SVM classification can be significantly enhanced by tuning its hyper-parameters (Gaspar et al., 2012). One relevant SVM parameter is the penalty factor C, which determines the trade-off between complexity and proportion of non-separable samples. Another critical SVM optimization is the careful selection of the kernel and its parameters. We selected, for each SVM of the dendrogram, the best performing kernel (i.e. either linear kernel: $k(x,y) = x^T y + c$, or a radial-basis function RBF-kernel: $k(x,y) = \exp(-x - y/2\sigma_2)$. To achieve this optimization without risk of overfitting the data, the identification of the best kernel and its parameters was performed at each fold on the training set only. We first compared linear to default RBF-SVM performance, and then if RBF-SVM showed better accuracy, the optimal *sigma* parameter for the RBF-kernel was selected using a 10-fold cross-validation on the training data (repeated N times for each of the N values of *sigma* to be tested). Note that, to achieve this, the *total training data* was itself split 10 times into a training (90%) and a testing (10%) subset. This SVM parameter optimization procedure was carried out individually for all SVMs of the dendrogram.

2.4.3. Feature subset selection

Another key factor in optimizing SVM performance is obviously an appropriate selection of the feature set. As explained above, feature space dimension was first reduced by keeping the most relevant ones (as determined by their two-class discrimination ability probed with statistical tests, see Section 2.3 for details). Next, to further optimize feature selection, we used a standard sequential forward procedure selection (SFPS) (e.g. Krakovska and Kristina, 2011) which, here also, was only performed within the training data because it involves cross-validation and we want to avoid any risk of over-fitting. The SFSP method can be summarized as follows: First, for a given SVM, the best feature is selected based on its binary classification ability. At each iteration, a new feature is selected and added so that the extended set maximizes the correct classification rate. As soon as the classification accuracy starts to drop, the algorithm terminates and the optimal feature set for this SVM is stored. This forward selection procedure was applied for each node of the dendrogram to select the best feature subset for each individual SVM.

2.4.4. Classification

Now that the optimal parameters and data features for each SVM of the tree have been set using training data, the classification can be tested by training the decision-tree multi-class SVM classifier and testing it on the test set. This overall cross-validation was repeated three times. The whole database was divided into three randomly determined folds. Each fold was divided into training (10 subjects) and testing sets (the remaining 5 subjects). The classification process was performed 10 times in each fold for robustness. A confusion matrix was computed each time and an overall confusion matrix was derived by averaging all 30 confusion matrices.

2.4.5. Effect of inter-scorer variability

Two additional analyses were includes in order to assess the potential effect of inter-scorer variability on the reported performances. First of all, in addition to the classification analysis that used expert 1 as ground truth, we also ran the same classification pipeline but this time using the second scorer (expert 2) as reference. Second, the classification performance was also re-computed once again by using only the portion of the data (ca. 75%) for which experts 1 and 2 show 100% agreement.

2.4.6. Performances measures

We computed several performance metrics in order to evaluate the proposed DSVM framework. Three within-class classification metrics were used and two global classification measures were used.

(a) Within-class classification metrics: First of all, for the evaluation of performance on a class by class basis (e.g. ability of the method to correctly detect REM sleep epochs), we computed three standard measures (Fraiwan et al., 2010) which are sensitivity (SE), specificity (SP) and accuracy (AC). These require first the calculation of the true positives (TP), false negatives (FN), true negatives (TN) and false positives (FP) by comparing the classification results to the expert classification as ground truth.

The three *within-class* classification performance measures are defined as:

(i) Sensitivity (SE) = TP/(TP + FN), i.e. true positive rate.
(ii) Specificity (SP) = TN/(TN + FP), i.e. true negative rate.
(iii) Accuracy (AC) = (TN + TP)/(TP + TN + FP + FN).

(b) Global classifier performance metrics: Overall accuracy (i.e. the accuracy calculated across all classes) is often used as a single metric of global performance in sleep staging. Overall accuracy across classes is computed as the sum of all the correctly classified observation of each class divided by the sum of all observations. However, overall accuracy is not a very reliable metric for global classifier performance, because it yields biased results if the number of samples varies significantly across the different classes (Alberg et al., 2004). Therefore, in order to assess the performance of the classifier as a whole, we also compute percent (%) correct classification across all

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

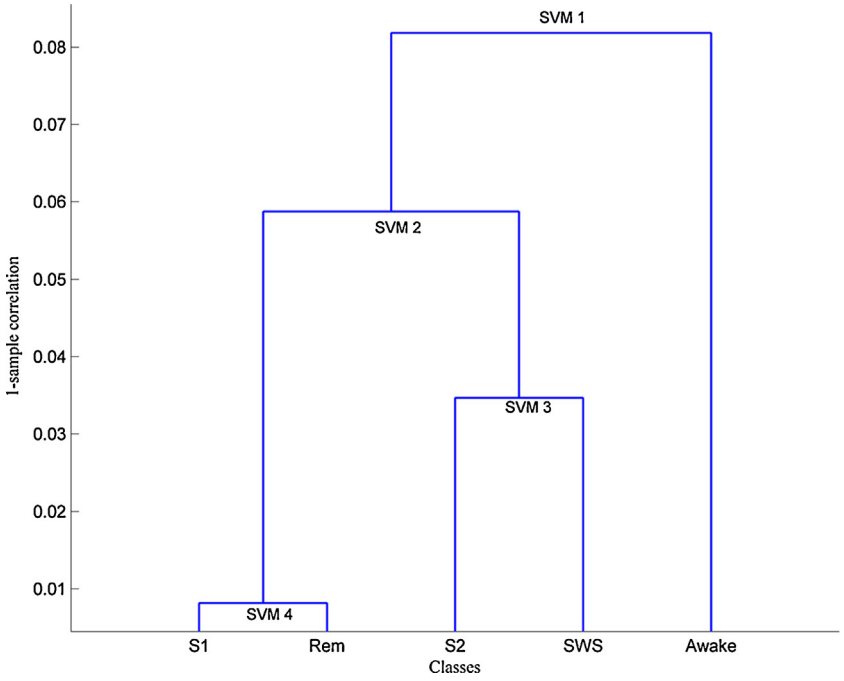


Fig. 2. Dendrogram computed via ascending hierarchical clustering (ACH) depicts the multiple SVM taxonomy generated for the five classes (Awake, S1, S2, SWS and REM) using the 32 feature space across all 15 subjects (see Table 2). The obtained dendrogram consist of the following 4 binary SVMs. SVM1: [Awake] vs. [SWS/S2/REM/S1], SVM2: [S1/REM] vs. [S2/SWS], SVM3: [S2] vs. [SWS] and SVM4: [S1] vs. [REM].

classes. This measure is sometimes called "decoding accuracy" and it is simply the mean of the sensitivity measures across all classes, in other words, the mean of the diagonal elements of the confusion matrix. Decoding accuracy (DA) allows for comparison of the classification performance with the theoretical chance level (i.e. 20% for a 5-class classification task) but also for comparison among classifiers.

3. Results

3.1. Dendrogram generation

The hierarchical cluster analysis step yielded the dendrogram shown in Fig. 2. At the top of the tree (i.e. the root node), the first binary decision occurs for Awake versus the rest of the stages {S1, S2, SWS, REM}. The Awake class is thus a terminal node (also known as leaf) and, when training SVM1, will be considered to be a negative class, while the remaining merged four classes are positive. Similarly, the second binary classifier in the tree (*SVM2*) will be trained considering elements of {S2, SWS} as negative and elements of {S1, REM} as positive. *SVM3* considers elements of {S2} as negative and {SWS} as positive. Finally, *SVM4* discriminates elements of {S1} from those of {REM} ones. Note that the structure of this decision tree was obtained after several clustering parameters and methods were tested. Using the distance maximization between classes as selection criterion, we found that the clustering based on correlation computing provided the best discrimination.

3.2. Feature subsets selection

The forward selection procedure described above (Section 2.4) helped identify for each of the four binary SVMs of the dendrogram a specific set of features to be used in order to train the individual SVM classifiers when running the overall multi-class classification process. This SVM-specific feature selection procedure was carried out for each of the three folds of the cross-validation procedure. The features that were selected via the forward selection approach systematically across all three folds are listed in Table 3.

Fig. 3 shows an illustrative example of the increase in correct classification rate (% decoding accuracy) with an increasing number of features, sequentially growing via the forward selection procedure. Note that these were obtained using cross-validation within the training data sets. The stopping criterion is the drop in classification rate (grey dashed line).

3.3. Classification results

Overall, our findings suggest that our proposed DSVM multi-class framework (trained on 10 individuals and tested on 5 other individuals) provided strong global sleep-stage classification performances, whether evaluated (a) in a stage by stage mode (Fig. 4), or (b) in terms of multi-class classification performance (Fig. 5) and importantly, (c) when compared to other standard classification methods (Figs. 6 and 7). Each of these three main findings will now be described in more detail.

3.3.1. Class prediction precision

Fig. 4 shows the sensitivity (SE), specificity (SP) and accuracy (AC) of DSVM class detection. All five classes are detected with a specificity and accuracy higher than 0.8. The best performances were obtained for epochs of REM sleep (with SE, SP and AC of

Table 3
List of features selected for each binary SVM via forward selection.

SVM1	'alpha/sigma' 'sigma/beta' 'theta/sigma' 'alpha/sigma' 'sigma/theta' 'beta/theta'
SVM2	'beta/sigma' 'rel power delta' 'delta/alpha' 'perm entropy_EMG' 'alpha/beta' 'alpha/delta' 'rel power theta' 'beta/alpha' 'alpha/theta' 'rel power alpha' 'sigma/delta'
SVM3	'sigma/alpha' 'permu entropy_EOG1' 'rel power sigma' 'spectral entropy'
SVM4	'alpha/delta' 'alpha/sigma' 'rel power delta' 'spectral entropy' 'permut entropy_EOG1' 'sigma/beta' 'sigma/theta 'beta/theta' 'alpha/beta' 'sigma/alpha'

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

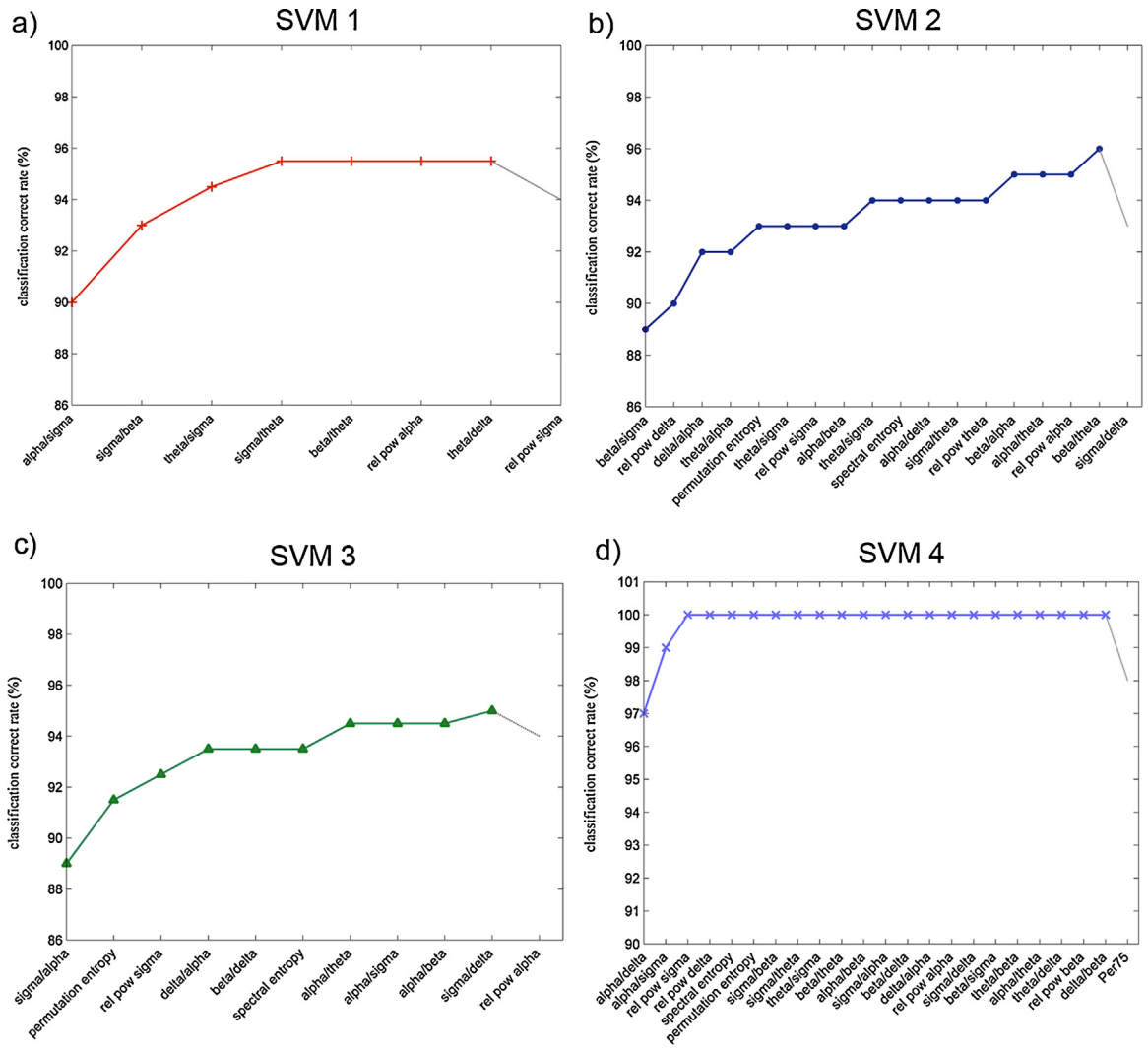


Fig. 3. Illustrative example of the increase in correct classification rate by including an increasing number of features selected via the forward selection procedure. This was applied only on the training data in order to identify the best features to use in each of the binary 4 SVMs that build up the dendrogram.

0.97, 1 and 0.98 respectively). In addition, sensitivity was above 0.7 for all stages, except for stage S1 (0.41), and it even exceeded 0.9 for Awake and REM. In other words, the sensitivity, specificity and accuracy results were high for all classes, with one exception being

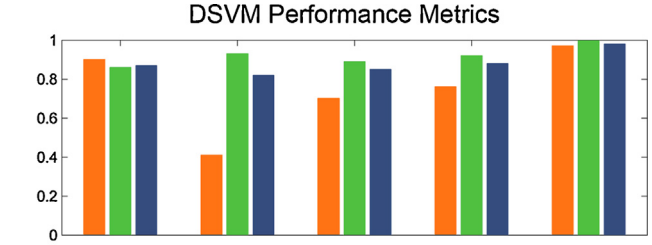


Fig. 4. Class-specific metrics for DSVM performance: Sensitivity (SE), Specificity (SP) and Accuracy (AC), all of which evaluate the precision of classifier prediction as compared to expert scoring.

the drop in sensitivity for S1. However, the specificity and accuracy for S1 were very high (0.93 and 0.82 respectively). The explanation for the drop in sensitivity just for the case of S1 can be inferred from the confusion matrix of the DSVM classifier (Fig. 7a). Beyond depicting the rate of correct classification (or sensitvity) for each stage in its diagonal elements, the values in the off-diagonal elements quantify the rate at which an epoch of class *i* (row) was mistakenly classified as belong to class *j* (column). Hence, the DSVM confusion matrix (Fig. 7a) shows that 37% of S1 epochs were incorrectly classified as Awake epochs. This strong confusion between Awake and S1 accounts for the low sensitivity found for S1. Note that the confusion matric is not symmetrical (only 8% of Awake epochs were falsely classified as S1). Inspection of the other diagonal (0.9, 0.7, 0.76 and 0.97 for Awake, S2, SWS and REM) and off-diagonal values for the corresponding rows confirms the robustness of DSVM classification.

3.3.2. Individual performances

Fig. 5 shows for each subject the percent decoding accuracy (mean sensitivity across all classes) that was obtained for each subject, when the given subject was part of the test set (i.e. when the data of the subject was not used for classifier training). Results show that decoding accuracies for all subjects were well above the chance level of 20% (a chance of 1 out of 5 of randomly detecting the correct

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

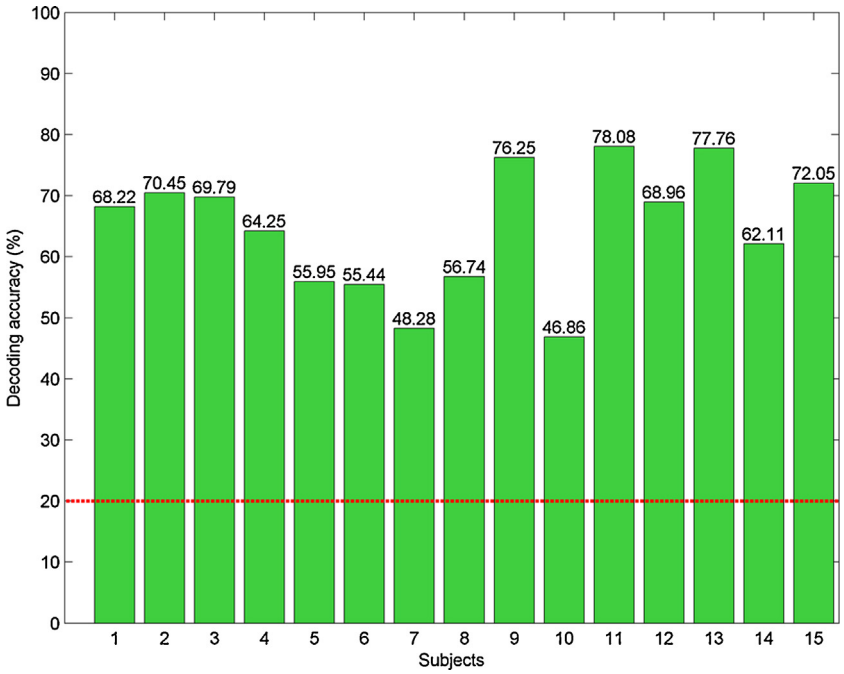


Fig. 5. Subject by subject decoding accuracy (mean sensitivity across all classes) for all 15 subjects. The dashed line indicates the classification theoretical chance level for 5 classes (i.e. 20%). Each of these values represents the mean of the diagonal elements of the confusion matrix for a given subject.

class). Note here that although such theoretical chance levels are not reliable for small sample sizes (Combrisson and Jerbi, 2015), the very high number of segments used here for each sleep stage (see Table 4) and high classification percentages obtained, allows to safely assess our results by comparison to such theoretical levels. Furthermore, DSVM classification performance varies across individuals with a standard deviation of 10.37%. The highest decoding accuracy (%) was found for S11 (79%) and the lowest for S10 (46%). Nevertheless, correct prediction of sleep stages was above chance level for all 15 subjects.

3.3.3. Comparison with standard classifiers

The performance of our method was also compared with two other widely used multi-class classifiers: The One-Against-All SVM (which was implemented using "Libsvm" library) and the Linear Discriminate Analysis (LDA). These two classifiers were trained and tested using the identical data partitions (3 folds of 10 training sets versus 5 testing sets) of our data base. Fig. 6 shows the direct comparison between decoding accuracy achieved using DSVM compared to the results of standard one-against-all and LDA. These findings suggest that the proposed Dendrogram-SVM yields the best performance with an overall decoding accuracy of 74.74%. It also has a better accuracy for all stages (expect S1 where the LDA classifier showed the highest performance with 73% of epochs correctly classified).

Finally, the three upper panels of Fig. 7 show the confusion matrices associated with each one of the three classifiers: DSVM, one-against-ALL SVM and LDA. Broadly speaking, sleep stages are mostly confused with adjacent elements in the matrix. This could be explained by the fact that sleep is a continuous process with stronger similarities among certain pairs of consecutively occurring stages (Awake and S1, or S1 and S2). Furthermore, because of this similarity, the transition between some of these adjacent stages may be harder to distinguish. Note this applies both to the trained classifier and to the human expert. All three classifiers show their highest decoding performance for "REM" and "Awake" stages. The S1 stage leads to the lowest correct decoding for the SVM based classifiers with 41% and 32% with the DSVM method and OAA-SVM respectively. However, the LDA classifier showed a good detection of this stage with 73% of epochs correctly classified.

Interestingly, Fig. 7 (panels d–f) also shows that if we recompute the performance of the proposed DSVM classification pipeline, but this time using expert 2 rather than expert 1 for ground-truth epoch labelling, we obtain very similar results with a mean DA across classes equal to 74.87% (std 2.12). Note that the original DA obtained when using expert 1 as ground truth was 74.74%. Beyond being very similar, these decoding accuracies are also close to the percent consistency among the two experts (76.73%). In addition, computing the DSVM classification performance of our method using only the data segments for which both experts showed 100% agreement yielded a mean DA of 81.81% (std 3.24). Fig. 7f summarizes DSVM performances across the three ground-truth scenarios (Exp1 as reference, Exp2 as reference, or agreement between Exp1 and Exp2 as reference), measured respectively as decoding accuracy (mean of confusion matrix diagonals) and overall accuracy (see Section 2.4.6). This last scenario yielded a decoding accuracy (mean sensitivity) of 81.81%, mean specificity of 94.4% and overall accuracy of 92.41%. Note that using data segments where multiple experts agree comes with the advantage of reduced human scoring errors (increased labelling reliability) but also implies the presence of fewer ambiguous (possibly transition phase) data segments.

Table 4
Number of epochs (30-s data segments) of each stage ($n = 15$).

Sleep stages	Awake	S1	S2	SWS	REM	Total
Number of epochs	880	921	4575	3456	2370	12,202

4. Discussion

This paper proposes and evaluates the performance of a multi-class decision-tree-based SVM framework (Dendrogram-SVM) for the automatic classification of 5 sleep stages from full-night polysomnographic recordings in 15 individuals (Scalp-EEG, EOG

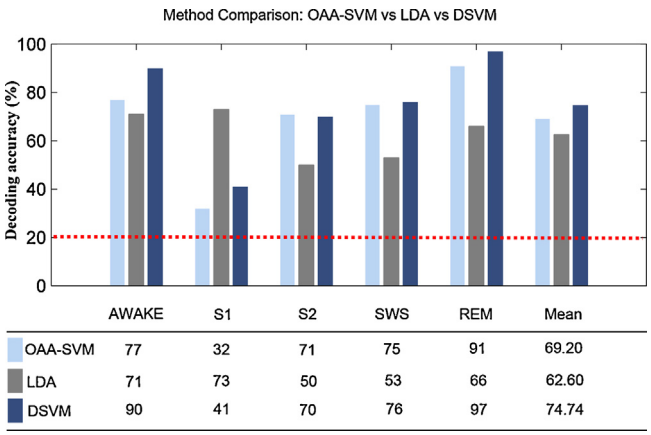
Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

and EMG). Our findings demonstrate that DSVM provides a high class by class correct detection rate and that the decoding accuracy across all tested individuals was well above the 20% chance level. By using a 3-fold cross validation and the 10 times averaging procedure, the correct detection rate was 74.74% with a standard deviation of 2.52%. The mean sensitivity, specificity and accuracy were 0.74, 0.92 and 0.88 respectively (compared to expert visual sleep scoring).

Moreover, the human-to-human agreement (consistency between the experts scoring) which was 76.73% is reasonably close to the machine-to-human agreement obtained with the proposed DSVM classifier, which yielded a decoding accuracy of 74.74% compared to expert 1 (and 74.87% compared to expert 2).

Most importantly, when applying more standard multi-class techniques (One-against-all SVM, or linear discriminant analysis) to the same data set (15 subjects), the decoding accuracy comparisons showed the overall superiority of the DSVM approach: Mean sensitivity for DSVM across all 5 classes was 75%, compared to 69% and 63% for OAA-SVM and LDA respectively. Sensitivity to REM epochs, for instance, was 97% in DSVM, compared to 91% and 66% for OAA-SVM and LDA respectively (Fig. 6).

The rationale for this study was to investigate whether we can improve sleep stage classification performance by combining the strengths of SVM classification (speed, applicability to relatively large sample sets, linear and non-linear kernel parameters, etc.) with the efficiency of decision-tree procedures. Moreover, standard SVM analysis is by definition a binary classifier, so using it here at each node (branching) of a decision tree structure, provides an efficient multi-class SVM framework. Our findings are in line with this hypothesis since the results obtained with DSVM did indeed outperform those achieved with more standard multi-class classification methods. While the DSVM framework proved to be more efficient for sleep stage classification for our 15 subject data set, it of course remains to be seen whether this results can be generalized to other EEG signal classification problems.



	AWAKE	S1	S2	SWS	REM	Mean
OAA-SVM	77	32	71	75	91	69.20
LDA	71	73	50	53	66	62.60
DSVM	90	41	70	76	97	74.74

Fig. 6. Comparison between the classification performance (sensitivity) of three methods: OAA-SVM, LDA and DSVM. The last column represents decoding accuracy (i.e. mean of the diagonal elements of the confusion matrix).

All three methods tested in this paper start with the identical feature extraction step, where of a wide range of features were computed from the EEG, EOG and EMG signals (linear time and frequency domain and non-linear time-domain features). Next, the feature space was reduced by excluding the least discriminant features following statistical quantification using *t*-tests. The same subset of selected features was then fed into the three classification methods. It is from here on that the three classification pipelines differ: Dendrogram-SVM, OAA-SVM and LDA.

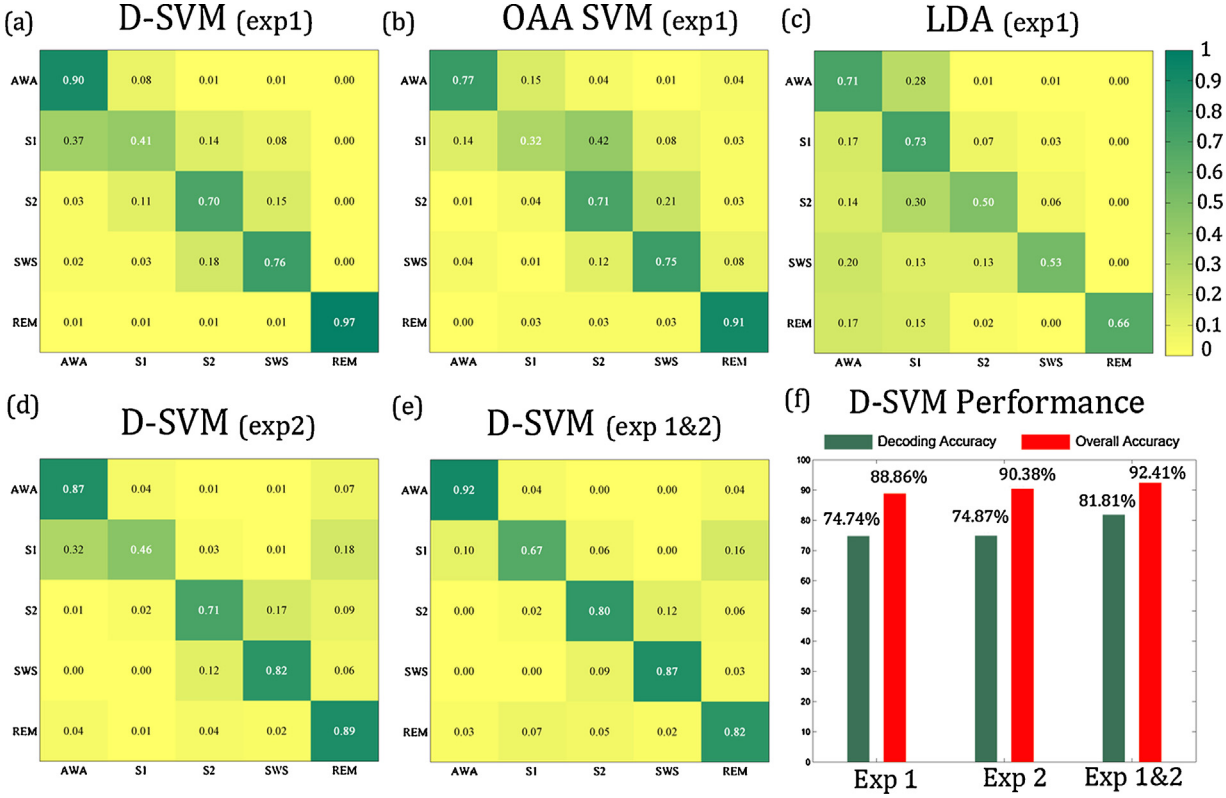


Fig. 7. Mean confusion matrices obtained with (a) DSVM, (b) OAA-SVM and (c) LDA, (d) DSVM (but with expert 2 staging as ground truth reference), (e) DSVM (but only using data segments where experts 1 and 2 were in agreement, i.e. ca.75% of segments), (f) Mean classification performance metrics for DSVM for three ground-truth scenarios (Exp1, Exp2, Exp1&2): Green bars depict decoding accuracy (i.e. mean of confusion matrix diagonals, or mean sensitivity), red bars depict the overall accuracy metric (see Section 2.4.6) which is often used in the sleep staging literature, and can thus be useful for comparison purposes.

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

Our findings suggest that the generation of an optimized dendrogram structure using ascending hierarchical clustering is an important source of performance enhancement. It reduces the number of classifiers and execution time. In particular, compared to the one-against-one SVM approach that would require 15 SVMs to run a 5-class classification the dendrogram approach boils down the classification to four binary SVMs. The dendrogram also optimizes the choice of clusters of classes to compare at each node, in a data driven manner. Another advantage of our method is that there is no need for a classification decision stage; any given sample is assigned at the end of the tree to one class. This said, it is important to keep in mind that the added value and novelty of the DSVM does not only stem from embedding multiple SVMs into the nodes of a decision-tree (which, as mentioned above, has also been explored in a few other classification studies). The strength and hence higher performance found with DSVM most likely arise from the combination of procedures that make up the pipeline. These include the optimization of the individual SVM parameters (linear or RBF kernels and sigma value) and the enhanced feature subset selection to be used with each binary SVM (based on sequential forward selection). In other words, the higher classification performance achieved with the DSVM framework proposed here stems on one hand from the combination of SVM classification and decision-tree architecture, and on the other hand from the ability within this framework to optimize certain parameters of the classifiers and the feature space.

Note that the feature selection procedure consists of two steps: (i) an initial dimensionality reduction is achieved via statistical testing (*t* test) and (ii) a feature selection based on a sequential forward selection applied recursively on the training set only. In order to quantify the added value of these feature selection steps, we re-analysed the data in two ways: First, we re-ran the same analysis pipeline proposed but this time we excluded the forward selection method. This resulted in the overall mean decoding accuracy across the 3 folds dropping from 74.74% to 65%. Second, we also re-ran the classification pipeline excluding both the forward selection procedure and the initial dimensionality reduction step. The mean decoding accuracy across the 3 folds dropped in this case from 74.74% down to 62%. These results suggest that the dimension reduction step and the forward selection procedure both positively impact decoding accuracy.

Although it is not possible to establish a systematic one-to-one relationship between electrophysiological features and each single sleep stage, the sleep literature does provide some idea about the predominance of specific features in different sleep stages. For instance, S1 is often associated with theta oscillations, S2 with the presence of K-complexes and spindles, while SWS contains delta oscillations and REM sleep is characterized among other things with bursts in the EOG signal superimposed on low frequency EEG (such as theta waves). The Awake state displays, among other things, string increases in alpha waves when the eyes are closed, and higher beta activity when the eyes are open. These characteristics are neither exhaustive nor necessarily verified by all sleep studies, but they do provide general guidelines for visual identification of sleep stages (Rechtschaffen and Kales, 1968; Iber et al., 2007; Šušmáková and Krakovská, 2008). Interestingly, the list of features that were automatically selected for each SVM of the dendrogram in our study show a reasonable overlap with these standard observations (see Fig. 2 for the SVMs of the dendrogram, and Table 3 for the list of associated features). The consistency is obviously only partial and the extensive list of features explored here go beyond the standard set reported in sleep staging manuals.

Our results show that the DSVM model classifies the majority of stages with high sensitivity, except for S1 (41%) which is often misclassified as Awake (as can be seen in the confusion matrix). In fact, S1 detection has also been shown to be associated with the weak correct detection rates in several other studies, with sensitivity values of 7.7% (Gunes et al., 2010), 35.12% (Sheng-Fu et al., 2012) and 33.6% (Shing-Tai et al., 2012). In addition to its similarity with Awake from an electrophysiological point of view, the relatively small number of available S1 epochs in a full night recording compared with other stages can also partly explain the low sensitivity; A low number of S1 epochs means that there are less data to train the model on, and thus the results on the test epochs can be weaker than for epochs where much more data was available to properly train the classifier. Interestingly, our findings also show that correct S1 classification was significantly increased when using LDA, compared with the two SVM-based classification methods. This observation suggests that a combined LDA-SVM model could be worth exploring in the quest for an efficient model with high performances for all stages.

Further improvements can be expected via the use of techniques that make a stronger use of the temporal dimension of the data, such as hidden Markov models (HMMS). Several studies have used Markov chains to classify sleep stages, in particular via Gaussian Observation HMMs (GOHMMs) or discrete HMMs (DHMMs). To date the results, expressed in terms of mean decoding accuracy across stages, are either comparable (Shing-Tai et al., 2012 [DA = 77.81%]) or lower (Flexer et al., 2002 [DA = 41.5%], Flexer et al., 2005 [DA = 60.1%]; Doroshenkov et al., 2007 [DA = 61.94%]) than the performance we report here (DA = 74.74%). Generally, the performances reported so far with HMMs seem to be lower than the best performances reported using SVM. Combining HMM and SVM might therefore be a promising alternative worth exploring (e.g. Lopes and Perdigão, 2007). Other promising methods include the use of unsupervised feature learning architecture called deep belief nets (DBNs), a probabilistic generative model with deep architecture that searches the parameter space by unsupervised greedy layerwise training (Längkvist et al., 2012).

Importantly, we have shown (Fig. 7) that the decoding accuracies and overall accuracies obtained with the DSVM method were robust to changes in expert reference (ground-truth labelling). When only using the subset of data segments for which both experts fully agreed during visual scoring (ca. 75% of all segments), the decoding accuracy increased to 81.81% and the overall accuracy metric reached 92.41%.

Note that the performances of automatic sleep staging methods reported in the literature show a high degree of variability. While the results reported here are promising, there are reports of higher classification accuracy and many studies with lower sensitivities than the values found here. There are several reasons for these discrepancies which need to be acknowledged. The first reason is the variability in the quality and size of the recorded data set (the application of our methodology on data sets reported in the literature will undoubtedly lead to results that are either slightly higher or slightly lower than the ones we obtained with the 15-subject polysomnographic data set we used). Secondly, inter-expert variability in visual sleep scoring is an acknowledged limitation that can also affect the results reported using automatic methods, and hampers direct comparison between studies (The benchmark against which the methods are compared in each study also shows inter-study variability). In addition, one needs to keep in mind that, beyond the difference in machine learning technique applied, various studies also differ in the feature selection procedure. Therefore, for all the reasons above, future studies will benefit from data sharing among researchers and rely on identical expert scoring as benchmark. Moreover, what is critical to bear in mind when evaluating the rate of correct classification is that, irrespective of the absolute values DSVM achieved, it outperformed the two more standard methods when applied on the same data and with the same initial feature space. We might therefore expect that the performance enhancements observed here might also hold

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

when comparing these three methods on other data sets. Of course, this remains to be tested. Nevertheless, from a pragmatic point of view, the most useful method might turn out to be the method that provides the most reliable starting point for a semi-automatic procedure where experts can easily fine-tune the automatic hypnogram output.

The expanding literature on the modification of various neurophysiological indices during various sleep stages constitutes a promising resource for ideas of features to test in the context of automatic sleep scoring. It is possible that higher performances could be achieved by exploring the discriminative power of further sleep specific neuronal phenomena: Quantifying the presence of K-complex waves (Colrain, 2005; Loomis et al., 1938), sleep spindles (Andrillon et al., 2011; Contreras and Steriade, 1996), bursts of high-frequency gamma oscillations (Ayoub et al., 2012; Dalal et al., 2010; Le Van Quyen et al., 2010; Valderrama et al., 2012; Worrell et al., 2012), monofractal and multifractal properties of the human sleep EEG (Weiss et al., 2009, 2011; Zorick and Mandelkern, 2013) and including them in the proposed DSVM method could potentially lead to an even better classification. The detection of some of these phenomena might be enhanced by recent methodological developments (Ahmed et al., 2009; Babadi et al., 2012; Chaibi et al., 2012, 2013, 2014; Jaleel et al., 2014; Nonclercq et al., 2013; O'Reilly and Nielsen, 2014a,b; Warby et al., 2014; Worrell et al., 2012). Furthermore, features such as cross-frequency interactions, long-range coupling among distant electrodes and long-range temporal correlations may also provide efficient novel markers for distinct sleep stages.

5. Conclusion

The aim of this study was to implement and evaluate an automatic sleep staging framework that could eventually help neuroscience researchers and clinicians by reducing the analysis time of polysomnographic signals while enhancing the quantitative nature and robustness of the scoring procedure. The results obtained here with data from 15 polysomnographic recordings demonstrate the utility for sleep scoring of a dendrogram-based multi-class SVM combined with a number of optimization steps related to feature selection and to SVM parameter selection. In addition to achieving high sensitivity, specificity and accuracy, the proposed DSVM pipeline outperforms two standard multi-class procedures (OAA-SVM and LDA). Finally, we believe that the methods described in this paper may also prove to be useful for the investigation of sleep disorders such as sleep apnea. More generally, the utility of dendrogram-SVM illustrated here for sleep stage classification may prove to also be of high interest for a wide range of EEG-based decoding problems such as the monitoring of cognitive brain states or decoding of motor intentions in the context of brain–computer interface (BCI) research (Astrand et al., 2014; Besserve et al., 2007; Jerbi et al., 2009, 2011; Krusienski and Wolpaw, 2009; Lotte et al., 2007).

Acknowledgements

Tarek Lajnef was in part supported by travel funds from EDST doctoral program and the LETI Laboratory, Sfax, Tunisia. Jean-Baptiste Eichenlaub is supported by the Fyssen Foundation. This work was partly performed within the framework of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program ANR-11-IDEX-0007. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program.

Appendix A. Software availability.

The functions that were developed and used in this study to perform multiclass Dendrogram-SVM classification have been made available online. The provided code can be used to generate a dendrogram based on hierarchical cluster analysis which is then used to train and classify multi-class data using the decision-tree SVM. We hope that this set of tools will help students and researchers replicate and extend our analyses. The code can be downloaded from Mathwork's File Exchange platform at the following URL: http://www.mathworks.com/matlabcentral/fileexchange/48632-multiclass-svm-classifier

References

Agarwal R, Gotman J. Computer-assisted sleep staging. IEEE Trans Biomed Eng 2001;48(December (12)):1412–23.
Agarwal R, Gotman J. Digital tools in polysomnography. J Clin Neurophysiol 2002;19(April (2)):136–43.
Aguera PE, Jerbi K, Caclin A, Bertrand O. ELAN: a software package for analysis and visualization of MEG, EEG, and LFP signals. Comp Intell Neurosci 2011;2011:158970.
Ahmed B, Redissi A, Tafreshi R. An automatic sleep spindle detector based on wavelets and the Teager energy operator. Conf Proc IEEE Eng Med Biol Soc 2009;2009:2596–9.
Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J Gen Intern Med 2004;19(May (5)):460–5.
Altunay S, Telatar Z, Erogul O. Epileptic EEG detection using the linear prediction error energy. Expert Syst Appl 2010;37:5661–5.
Andrillon T, Nir Y, Staba RJ, Ferrarelli F, Cirelli C, Tononi G, et al. Sleep spindles in humans: insights from intracranial EEG and unit recordings. J Neurosci 2011;31:17821–34.
Astrand E, Enel P, Ibos G, Dominey PF, Baraduc P, Ben Hamed S. Comparison of classifiers for decoding sensory and cognitive information from prefrontal neuronal populations. PLOS ONE 2014;9(January (1)):e86314.
Ayoub A, Mölle M, Preissl H, Born J. Grouping of MEG gamma oscillations by EEG sleep spindles. Neuroimage 2012;59:1491–500.
Babadi B, McKinney SM, Tarokh V, Ellenbogen JM. DiBa: a data-driven Bayesian algorithm for sleep spindle detection. IEEE Trans Biomed Eng 2012;59:483–93.
Bala M, Agrawal RK. Optimal decision tree based multi-class support vector machine. Inform Slovenia 2011;35:197–209.
Bandt C, Pompe B. Permutation entropy: a natural complexity measure for time series. Phys Rev Lett 2002;88, 174–102.
Becq G, Charbonnier S, Chapotot F, Buguet A, Bourdon L, Baconnier P. Comparison between five classifiers for automatic scoring of human sleep recordings. Stud Comp Intell 2005;4:113–27.
Benabdeslem K, Bennani Y. Dendogram-based SVM for multi-class classification. J Comp Inform Technol 2006;4:283.
Berthomier C, Drouot X, Herman-Stoïca M, Berthomier P, Prado J, Djibril, et al. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. Sleep 2007;30:1587–95.
Besserve M, Jerbi K, Laurent F, Baillet S, Martinerie J, Garnero L. Classification methods for ongoing EEG and MEG signals. Biol Res 2007;40:415–37.
Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory ACM; 1992. p. 144–52.
Chaibi S, Bouet R, Jung J, Lajnef T, Samet M, Bertrand O, et al. Developement of Matlab-based Graphical User Interface (GUI) for detection of high frequency oscillations (HFOs) in epileptic patients. IEEE; 2012. p. 56–62, Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6152445
Chaibi S, Sakka Z, Lajnef T, Samet M, Kachouri A. Automated detection and classification of high frequency oscillations (HFOs) in human intracereberal EEG. Biomed Signal Process Control 2013;8(November (6)):927–34.
Chaibi S, Lajnef T, Sakka Z, Samet M, Kachouri A. A reliable approach to distinguish between transient with and without HFOs using TQWT and MCA. J Neurosci Methods 2014;232:36–46.
Colrain IM. The K-complex: a 7-decade history. Sleep 2005;28:255–73.
Combrisson E, Jerbi K. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods 2015., http://dx.doi.org/10.1016/j.jneumeth.2015.01.010.
Contreras D, Steriade M. Spindle oscillation in cats: the role of corticothalamic feedback in a thalamically generated rhythm. J Physiol 1996;490:159–79.
Cortes C, Vapnik VN. Support vector networks. Mach Learn 1995;20:1–25.
Dalal SS, Hamamé CM, Eichenlaub JB, Jerbi K. Intrinsic coupling between gamma oscillations, neuronal discharges, and slow cortical oscillations during human slow-wave sleep. J Neurosci 2010;30:14285–7.
Doroshenkov LG, Konyshev VA, Selishchev SV. Classification of human sleep stages based on EEG processing using hidden Markov models. Biomed Eng 2007;41:25–8.

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022

Eichenlaub JB, Morlet D, Ruby P. What is the specificity of the response to the own first-name when presented as a novel in a passive oddball paradigm? An ERP study. Brain Res 2012;1447:65–78.
Eichenlaub JB, Bertrand O, Morlet D, Ruby P. Brain reactivity differentiates subjects with high and low dream recall frequencies during both sleep and wakefulness. Cereb Cortex 2014;24:1206–15.
Erdamar A, Duman F, Yetkin S. A wavelet and Teager energy operator based method for automatic detection of K-Complex in sleep EEG. Expert Syst Appl 2012;39:1284–90.
Ferenets R, Lipping T, Anier A, Jäntti V, Melto S, Hovilehto S. Comparison of entropy and complexity measures for the assessment of depth of sedation. IEEE Trans Biomed Eng 2006;53:1067–77.
Ferlazzo E, Mammone N, Cianci V, Gasparini S, Gambardella A, Labate A, et al. Permutation entropy of scalp EEG: a tool to investigate epilepsies: suggestions from absence epilepsies. Clin Neurophysiol 2014;125:13–20.
Flexer A, Dorffner G, Sykacek P. Rezek I: an automatic, continuous and probabilistic sleep stager based on a hidden Markov model. Appl Artif Intell 2002;16:199–207.
Flexer A, Gruber G, Dorffner G. A reliable probabilistic sleep stager based on a single EEG signal. Artif Intell Med 2005;33:199–207.
Fraiwan L, Lweesy K, Khasawneh N, Fraiwan M, Wenz H, Dickhaus H. Classification of sleep stages using multi-wavelet time frequency entropy and LDA methods. Inf Med 2010;49.
Gaspar P, Carbonell J, Oliveira JL. On the parameter optimization of Support Vector Machines for binary classification. J Integr Bioinform 2012;9:201.
Gjorgji M, Dejan G, Ivan CA. Multi-class SVM classifier utilizing binary decision tree. Informatica 2009;33:233–41.
Grozinger M, Fell J, Roschke J. Neural net classification of REM sleep based on spectral measures as compared to nonlinear measures. Biol Cybern 2001;85:335–41.
Gunes S, Polat K, Yosunkaya S. Efficient sleep stage recognition system based on EEG signal using *k*-means clustering based feature weighting. Expert Syst Appl 2010;37:7922–8.
Hastie T. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, NY: Springer; 2009.
Huang C, Davis LS, Townshend JRG. An assessment of support vector machines for land cover classification. Int J Remote Sens 2002;23:725–49.
Huang C-S, Lin C-L, Ko L-W, Liu S-Y, Su T-P, Lin C-T. Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels. Front Neurosci 2014(September):8.
Iber C, Ancoli-Israel I, Chesson AL, Quan SF. The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specification. American Academy of Sleep Medicine; 2007.
Itil TM, Shapiro DM, Fink M, Kassebaum D. Digital computer classifications of EEG sleep stages. Electroencephalogr Clin Neurophysiol 1969;27(July (1)):76–83.
Jaleel A, Ahmed B, Tafreshi R, Boivin DB, Streletz L, Haddad N. Improved spindle detection through intuitive pre-processing of electroencephalogram. J Neurosci Methods 2014;233:1–12.
Jerbi K, Freyermuth S, Minotti L, Kahane P, Berthoz A, Lachaux JP. Watching brain TV and playing brain ball exploring novel BCI strategies using real-time analysis of human intracranial data. Int Rev Neurobiol 2009;86:159–68.
Jerbi K, Vidal JR, Mattout J, Maby E, Lecaignard F, Ossandon T, et al. Inferring hand movement kinematics with MEG, EEG and intracranial EEG: from brain-machine interfaces to motor rehabilitation. IRBM BioMed Eng Res 2011;32(1):8–18.
Kaiser JF. On a simple algorithm to calculate the 'energy' of a signal. In: Proc IEEE ICASSP'90; 1990. p. 381–4.
Kerkeni N, Ben Cheikh R, Bedoui MH, Alexandre Dogui M. Sleep stages classification by hierarchical artificial neural networks. IRBM 2012;33:35–40.
Khalighi S, Sousa T, Pires G, Nunes U. Automatic sleep staging: a computer assisted approach for optimal combination of features and polysomnographic channels. Expert Syst Appl 2013;40(December (17)):7046–59.
Koley B, Dey D. An ensemble system for automatic sleep stage classification using single channel EEG signal. Comp Biol Med 2012;42:1186–95.
Krakovska A, Kristina M. Automatic sleep scoring. A search for an optimal combination of measures. Artif Intell Med 2011;53:25–33.
Krusienski DJ, Wolpaw JR. Brain–computer interface research at the wadsworth center developments in noninvasive communication and control. Int Rev Neurobiol 2009;86:147–57.
Lajnef T, Chaibi S, Kachouri A, Samet M. Epileptic seizure detection using linear prediction filter. In: 12th international conference on Sciences and Techniques of Automatic control and computer engineering; 2010.
Längkvist M, Karlsson L, Loutfi A. Sleep stage classification using unsupervised feature learning. Adv Artif Neural Syst 2012;2012:1–9.
Larsen LE, Walter DO. On automatic methods of sleep staging by EEG spectra. Electroencephalogr Clin Neurophysiol 1970;28(May (5)):459–67.
Le Van Quyen M, Staba R, Bragin A, Dickson C, Valderrama M, Fried I, et al. Large-scale microelectrode recordings of high-frequency gamma oscillations in human cortex during sleep. J Neurosci 2010;30:7770–82.
Loomis AL, Harvey N, Hobart GA. Distribution of disturbance patterns in the human electroencephalogram, with special reference to sleep. J Neurophysiol 1938;1:413–30.
Lopes C, Perdigão F. Hybrid HMM/SVM speech event detector. In: 6th conference on telecommunications, Conftele 2007, v. 1; 2007. p. 601–4.
Lotte F, Congedo M, Lécuyer A, Lamarche F, Arnaldi B. A review of classification algorithms for EEG-based brain-computer interfaces. J Neural Eng 2007;4:R1–13.
Ma HY, Hu B, Jackson M, Yan JZ, Zhao W. A hybrid classification method using artificial neural network based decision tree for automatic sleep scoring. World Acad Sci Eng Technol 2011;79:279–84.
Marina R, et al. Sleep scoring using artificial neural networks. Sleep Med Rev 2012;16:251–63.
Melgani F, Bruzzone L. Classification of hyper spectral remote sensing images with support vector machine. IEEE Trans Geosci Remote Sens 2004;42:1778–90.
Nonclercq A, Urbain C, Verheulpen D, Decaestecker C, Van Bogaert P, Peigneux P. Sleep spindle detection through amplitude-frequency normal modelling. J Neurosci Methods 2013;214:192–203.
Olofsen E, Sleigh JW, Dahan A. Permutation entropy of the electroencephalogram: a measure of anaesthetic drug effect. Br J Anaesth 2008;101:810–21.
Oppenheim AV, Schafer RW. Discrete-time signal processing. Prentice Hall; 1999.
O'Reilly C, Nielsen T. Assessing EEG sleep spindle propagation. Part 1: theory and proposed methodology. J Neurosci Methods 2014a;221:202–14.
O'Reilly C, Nielsen T. Assessing EEG sleep spindle propagation. Part 2: experimental characterization. J Neurosci Methods 2014b;221:215–27.
Pardey J, Roberts S, Tarassenko L. A review of parametric modeling techniques for EEG analysis. Med Eng Phys 1996;18:2–11.
Rechtschaffen A, Kales A, editors. A manual of standardized terminology, techniques and scoring system for sleep stages of human subject. Washington, DC: US Government Printing Office, National Institute of Health Publication; 1968.
Ronzhina M, Janoušek O, Kolářová J, Nováková M, Honzík P, Provazník I. Sleep scoring using artificial neural networks. Sleep Med Rev 2012;16(June (3)):251–63.
Ruby P, Blochet C, Eichenlaub JB, Bertrand O, Morlet D, Bidet-Caulet A. Alpha reactivity to complex sounds differs during REM sleep and wakefulness. PLOS ONE 2013a;8(11):e79989.
Ruby P, Blochet C, Eichenlaub JB, Bertrand O, Morlet D, Bidet-Caulet A. Alpha reactivity to first names differs in subjects with high and low dream recall frequency. Front Psychol 2013b;13(4):419.
Schaltenbrand N, et al. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep 1996;19:26–35.
Shalbaf R, Behnam H, Sleigh JW, Steyn-Ross A, Voss LJ. Monitoring the depth of anesthesia using entropy features and an artificial neural network. J Neurosci Methods 2013;218:17–24.
Sheng-Fu L, Chin-En K, Yu-Han H, Yu-Shian C. A rule-based automatic sleep staging method. J Neurosci Methods 2012;205:169–76.
Shing-Tai P, Chih-En K, Jian-Hong Z, Sheng-Fu L. A transition-constrained discrete hidden Markov model for automatic sleep staging. BioMed Eng Online 2012:11–52.
Silber MH, Ancoli-Israel S, Bonnet MH, Chokroverty S, Grigg-Damberger MM, Hirshkowitz M, et al. The visual scoring of sleep in adults. J Clin Sleep Med 2007;3(March (2)):121–31.
Stanus E, Lacroix B, Kerkhofs M, Mendlewicz J. Automated sleep scoring: a comparative reliability study of two algorithms. Electroencephalogr Clin Neurophysiol 1987;66(April (4)):448–56.
Steinn G, Runarsson TP, Sven S. Automatic sleep staging using support vector machines with posterior probability estimates. In: International conference on computational intelligence for modelling, control and automation 2005 and international conference on intelligent agents, web technologies and internet commerce; 2005. p. 366–72.
Šušmáková K, Krakovská A. Discrimination ability of individual measures used in sleep stages classification. Artif Intell Med 2008;44(November (3)):261–77.
Takahashi F, Abe S. Decision-tree-based multiclass support vector machines. In: Proceedings of the ninth international conference on neural information processing; 2002. p. 1418–22.
Tuffery S. Data mining and statistics for decision making. Chichester, West Sussex; Hoboken, NJ: Wiley; 2011.
Valderrama M, Crépon B, Botella-Soler V, Martinerie J, Hasboun D, Alvarado-Rojas C, et al. Human gamma oscillations during slow wave sleep. PLoS ONE 2012;7(4):e33477.
Warby SC, Wendt SL, Welinder P, Munk EG, Carrillo O, Sorensen HB, et al. Sleep-spindle detection: crowd sourcing and evaluating performance of experts, non-experts and automated methods. Nat Methods 2014;11:385–92.
Weiss B, Clemens Z, Bódizs R, Vágó Z, Halász P. Spatio-temporal analysis of monofractal and multifractal properties of the human sleep EEG. J Neurosci Methods 2009;185(December (1)):116–24.
Weiss B, Clemens Z, Bódizs R, Halász P. Comparison of fractal and power spectral EEG features: Effects of topography and sleep stages. Brain Res Bull 2011;84(April (6)):359–75.
Worrell GA, Jerbi K, Kobayashi K, Lina JM, Zelmann R, Le Van Quyen M. Recording and analysis techniques for high-frequency oscillations. Prog Neurobiol 2012;98(September (3)):265–78.
Zanin M, Zunino L, Rosso OA, Papo D. Permutation entropy and its main biomedical and econophysics applications: a review. Entropy 2012;14:1553–77.
Zorick T, Mandelkern MA. Multifractal detrended fluctuation analysis of human EEG: preliminary investigation and comparison with the wavelet transform modulus maxima technique. PLoS ONE 2013;8(July (7)):e68360.

Please cite this article in press as: Lajnef T, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods (2015), http://dx.doi.org/10.1016/j.jneumeth.2015.01.022