Article

Automatic Sleep Stages Classification Using EEG Entropy Features and Unsupervised Pattern Analysis Techniques

Jose Luis Rodríguez-Sotelo [1]**, Alejandro Osorio-Forero** [2]**, Alejandro Jiménez-Rodríguez** [2]**, David Cuesta-Frau** [3,*]**, Eva Cirugeda-Roldán** [3] **and Diego Peluffo** [4]

[1] Grupo de Automática, Universidad Autónoma de Manizales, Antigua estación del ferrocarril, Manizales 170002, Colombia; E-Mail: jlrodriguez@autonoma.edu.co

[2] Grupo de Investigación de Neuroaprendizaje, Universidad Autónoma de Manizales, Antigua estación del ferrocarril, Manizales 170002, Colombia; E-Mails: alejosorio1@gmail.com (A.O.-F.); aljiro@gmail.com (A.J.-R.)

[3] Technological Institute of Informatics, Polytechnic University of Valencia, Alcoi Campus, Plaza Ferrándiz y Carbonell, 2, Alcoi 03801, Spain; E-Mail: ecirugeda@giica.com

[4] Universidad Cooperativa de Colombia, Faculty of Medicine, Pasto 520002, Colombia; E-Mail: diego.peluffo@campusucc.edu.co

* Author to whom correspondence should be addressed; E-Mail: dcuesta@disca.upv.es; Tel.: +34-966528505.

Abstract: Sleep is a growing area of research interest in medicine and neuroscience. Actually, one major concern is to find a correlation between several physiologic variables and sleep stages. There is a scientific agreement on the characteristics of the five stages of human sleep, based on EEG analysis. Nevertheless, manual stage classification is still the most widely used approach. This work proposes a new automatic sleep classification method based on unsupervised feature classification algorithms recently developed, and on EEG entropy measures. This scheme extracts entropy metrics from EEG records to obtain a feature vector. Then, these features are optimized in terms of relevance using the Q-α algorithm. Finally, the resulting set of features is entered into a clustering procedure to obtain a final segmentation of the sleep stages. The proposed method reached up to an average of 80% correctly classified stages for each patient separately while keeping the computational cost low.

1. Introduction

Humans devote approximately one third of their life to sleep [1]. The biological rationale of this need remains unknown. However, there have been some attempts to elucidate the role of sleeping in human beings. It has been related to plasticity [2], and also to the reorganization of memory [3,4].

Unfortunately, sleep disorders and/or deprivation affect a great number of subjects worldwide. This entails a huge impact on public and individual health, including economic costs, reduced quality of life, co-morbidities, and even early mortality [5]. These are the main reasons why sleep is a subject of intense systematic research interest, including sleep stages detection and characterization.

Polysomnography (PSG) is the most commonly used technique in medicine for the diagnosis and understanding of the sleep phenomena [6]. PSG is a multimodal recording of different bio-signals during the whole sleep period at night. The Electroencephalogram (EEG) can be used to indirectly study the dynamics of the neural activity [7], which is time-varying during night-time [8] (pp. 193–208, Chapter 10).

Sleep is a highly nonstationary and nonlinear process whose statistics are time-varying but remain constant inside a short period of time known as sleep stage. Sleep analysis considers 5 different stages: wake (W), drowsiness (N1), light sleep (N2), deep sleep (N3), and rapid eye movement sleep (REM) [9].

The scoring and classification of sleep stages is manually performed by experts who analyze the PSG records in small time windows or epochs (commonly 30 s duration). This task is difficult, subjective and is often an exasperating time consuming process, thus leading to a low reliability and differences among scorers [10,11]. In order to effectively address these issues, automatic classification schemes have been progressively introduced as an objective way to assist the experts during this process [12,13].

The most commonly used techniques for automatic sleep stage classification are still supervised however. Namely, they require user intervention or manual data annotation. Among these techniques, neural networks (NN) [14] yield very good results, with up to 80% classification accuracy [14,15]. Other approaches have achieved 95% accuracy in more specific applications, such as differentiating alert states from drowsy and sleep states. All these supervised approaches strongly depend on the predefined scorer's labels [16], and therefore they lack of flexibility, adaptability, and reusability.

In this work, a novel unsupervised classification scheme for sleep stage determination based on EEG records is proposed. This technique computes entropy metrics from the signals to build a feature vector. This vector is optimized in terms of relevance using the recently developed Q-α based approach [17] in order to reduce computational cost. The resulting weighted feature vector is processed by a J-means clustering algorithm to obtain the final stage classification. The segmentation accuracy rate achieved in the experiments was up to 80% for each patient separately, comparable to that obtained with supervised methods.

2. Method

A representative experimental set was processed using the method proposed. A vector of EEG entropy features was computed for different epochs of 30 s of duration. For comparative purposes, a number of entropy metrics were utilized. Spare features were removed using the Q-α algorithm, and the final vectors were clustered according to a J-means scheme. Results were assessed both in terms of classification accuracy and computational burden. Each step of the method proposed is described in Sections 2.2–2.4 (see flow chart in Figure 1), including the experimental dataset employed for validation (Section 2.1).

Figure 1. Flow chart of the method proposed. The EEG records are first processed to extract features that are then selected to optimize the information density. Finally, a clustering algorithm creates the partition of records into sleep stages.
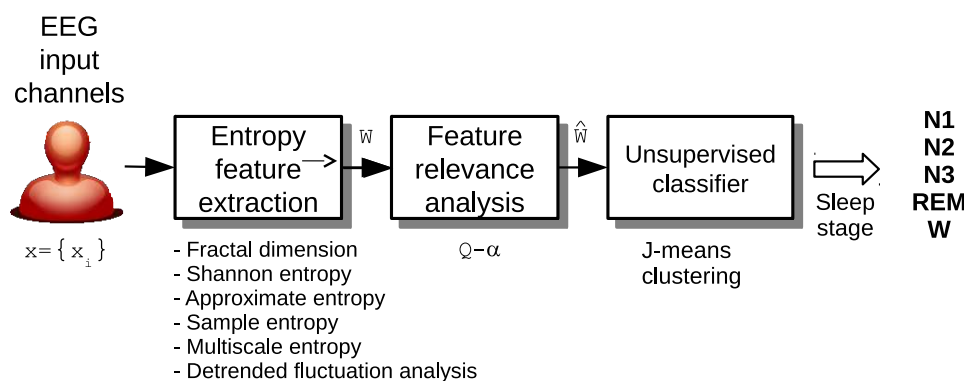
2.1. Experimental Dataset

The EEG signals used in the experiments were drawn from the SC Sleep-EDF Database [Expanded] [18]. This database is freely available through Physionet [19].

Specifically, only the Fpz-Cz and Pz-Oz EEG channels were analyzed [20]. These signals correspond to Caucasian normal males and females aged 25 to 34 years old, taking no medication. In total, 39 records sampled at 100 Hz and filtered from 0.5 to 100 Hz were employed in the experiments. Further details of the database can be found at [18,21].

Sleep stages were scored manually using Rechtschaffen and Kales (R & K) criteria [22]. Stage labels were provided together with the database by Physionet [19], using additionally recorded signals, and according to standard scoring rules [22]. Epochs marked as movement and unscored were rejected for further analysis (61 epochs). Information about the inter-rater reliability for sleep scoring can be found in [10,11]. For the specific purposes of the present study, sleep stages 3 and 4 were merged into a single deep sleep stage (N3).

2.2. Feature Extraction

Given an input discrete time series $x = \{x_1, \cdots, x_N | x_i \in \mathbb{R}, i \in \mathbb{Z} > 0\}$ of length N, a vector $\mathbf{V} \in \mathbb{R}^q$ is formed by q points, being q the number of features $\psi = \{\psi_1, \cdots, \psi_q | \psi_j \in \mathbb{R}, j \in \mathbb{Z} > 0\}$ used (number of 30 s epochs).

The aim of the feature extraction stage was to compare the performance of different entropy estimators when applied to the experimental set, and find out if some metrics are more sensitive to EEG patterns. The changes in sleep stages cause changes in the EEG, and these changes are also expected to be reflected in the entropy results, as observed in many other works, see [8] (pp. 193–208, Chapter 10) and [18,20,23]. The specific features computed in this study using the standard algorithms were: Fractal Dimension (FD), Detrended Fluctuation Analysis (DFA), Shannon entropy (H), Approximate Entropy (ApEn), Sample Entropy (SampEn), and Multiscale Entropy (MSE). A total of $q = 34$ complexity measures were included in the analysis.

2.2.1. Fractal Dimension

The estimation of the FD was applied in order to account for signal complexity and scale invariance. FD statistically quantifies how well a fractal matches the input data at different scales.

The method to compute FD was based on the box counting algorithm [24,25]. FD is estimated as the slope of the straight line fitted to the curve formed by the sequence $(\ln(L), (S(L)/L))$, where L is the size of the box, and $S(L)$ is the number of boxes. If the sampling interval of the input time series is Δt and the box size is $L = n\Delta t$, the number of boxes $S(L)$ can be obtained from:

$$S(n\Delta t) = \sum_{i=1}^{\text{mod}(N/n)} |\max(\Delta x_i) - \min(\Delta x_i)| \tag{1}$$

where $\Delta x_i = x_{n(i-1)+1}, x_{n(i-1)+2}, \cdots, x_{n(i-1)+n+1}$. The FD of x is estimated by counting the number of boxes needed to cover the curve [24]. A fractal dimension index was calculated for both EEG channels referred as FD.Fpz-Cz and FD.Pz-Oz (features ψ_1, ψ_2).

2.2.2. Detrended Fluctuation Analysis

It is not a true entropy estimator, but DFA is a measure that allows the detection of long-range power-law correlations in a time series [26]. The first step to obtain DFA is to compute the integrated time series as $y = \{y_k = \sum_{i=1}^{k} x_i\}$. Then, y is divided into N/L boxes of length L (the parameter L defines the time-scale). In each box, a line is fitted and its ordinate, denoted by y_k^L, is taken as the trend of the time series in that box. The integrated time series is detrended by subtracting y_k^L and the root mean square fluctuation is computed according to:

$$f(L) = \sqrt{\frac{1}{N} \sum_{i=k}^{N} (y_i - y_i^L)^2} \tag{2}$$

This process is repeated for different scales L. Finally, the scaling exponent that represents the DFA value is obtained from the slope of a linear fitting between $(\log_{10}(L), \log_{10}(f(L)))$. Three features were calculated from the DFA for each channel, the scaling exponents of the whole epoch (DFA-α), and scaling exponents (DFA-a1 and DFA-a2) before and after the relative estimated error correction (features $\psi_3 - \psi_8$). A detailed description of this algorithm can be found in [27].

2.2.3. Shannon Entropy

This is a measure of the data spread. It was calculated for each channel according to:

$$H(x) = -\sum_i p(x_i) \log(p(x_i)) \tag{3}$$

where $p(x_i)$ is the probability $p(x = x_i)$. In this work, we considered $p(x_i) = x_i^2$. This entropy measure has already been used in sleep EEG signal processing, yielding high values in wakefulness and REM sleep stages, and low values in N3 stages [28]. Shannon entropy was calculated for each channel independently (features ψ_9 and ψ_{10}).

2.2.4. Approximate Entropy

ApEn is related to the predictability or regularity of a time series. It was devised as an approximation of the Kolmogorov entropy of an underlying process [29].

The ApEn algorithm is basically a search for the repetitive patterns of length m commencing at sample i in which the distance induced by the maximum norm differs up to an error threshold r [30]. The ApEn statistic, given an input value for parameters m and r, is defined as:

$$\text{ApEn}(m, r, N) = \Phi_m(r) - \Phi_{m+1}(r) \tag{4}$$

where

$$\Phi_m(r) = \mathcal{E}\left\{\ln\left(\frac{c_i^m(r)}{N - m + 1}\right)\right\} \tag{5}$$

being $c_i^m(r)$ the number of vectors $\mathbf{x}_i \in \mathbb{R}^m$ such that $d(\mathbf{x}_i, \mathbf{x}_j) < r, \mathbf{x}_i = \{x_i, x_{i+1}, \cdots, x_{i+m-1}\}$, $1 \leq i, j \leq N - m + 1$. ApEn was computed for both Fpz-Cz and Pz-Oz channels (features ψ_{11} and ψ_{12}).

2.2.5. Sample Entropy

SampEn was an evolution of ApEn devised to solve the bias of ApEn due to counting self-matches. SampEn exhibits better statistical properties [30] than ApEn. It is computed in a similar fashion, but the final step becomes:

$$\text{SampEn}(m, r, N) = -\ln\left(\frac{A^m(r)}{B^m(r)}\right) \tag{6}$$

where $B^m(r)$ is defined as the mean of the number of vectors $\mathbf{x}_i \in \mathbb{R}^m$, such that $d(\mathbf{x}_i, \mathbf{x}_j) < r$ with $i \neq j$, divided by $N - m + 1$. The value of $A^m(r)$ is defined similarly for $\mathbf{x}_i \in \mathbb{R}^{m+1}$. SampEn was calculated for two template lengths ($m = 0$ and $m = 1$) for each channel (features $\psi_{13} - \psi_{16}$).

2.2.6. Multiscale Entropy

MSE is an estimator of the complexity of a time series in which entropy measures are applied at different time-scales. The calculation is carried out in two steps [31]. First, a coarse grained time series is constructed for the different scales τ considered ($1 < \tau \leq 9$ in this work), from the original signal x:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i \tag{7}$$

Second, an auxiliary single-scale entropy measure (we chose SampEn) is computed over the resulting coarse-grained series. MSE was computed for scales $n = 1, 2, \cdots, 9$ for the two channels (features $\psi_{17} - \psi_{34}$).

2.3. Feature Relevance Analysis

The previous feature extraction stage yields a set of p vectors $\mathbf{V} \in \mathbb{R}^q$, where the q components correspond to the features computed as described in Section 2.2. Next, in order to reduce the computational cost of the classification stage and improve the classification speed, a dimensionality reduction method, termed feature relevance analysis, takes place. This method provides a reduced subset $q' < q$ of the input features in p, but maximizing the preserved relevant discriminatory information. This optimization is carried out following a matrix projection scheme, similar to that used in Principal Component Analysis (PCA) methods. The output of this stage is a feature significance quantitative score that, using a threshold, enables or disables the exclusion of that feature. The value selected for the threshold was an accumulated variance criterion of 98% [32].

Specifically, this work applies a relevance analysis method based on the Q-α algorithm recently developed by us, and applied successfully to Electrocardiographic records [17]. This relevance analysis was applied to each patient separately. Features were retained in the final set if they were found relevant for at least 20 subjects.

The Q-α algorithm is an optimization procedure. Given a matrix $\mathbf{W}$ that contains the p feature vectors $\mathbf{V}$ of length q, $\mathbf{W} \in \mathbb{R}^{p \times q}$ the objective is to find a new matrix $\hat{\mathbf{W}}$ with $q' < q$ features, $\hat{\mathbf{W}} \in \mathbb{R}^{p \times q'}$. To obtain $\hat{\mathbf{W}}$, it is necessary to solve the optimization problem given by the general expression:

$$\max_{\mathbf{Q},\mathbf{A}} tr\left(\mathbf{Q}^T \mathbf{A}\mathbf{A}\mathbf{Q}\right) \tag{8}$$

where $\mathbf{A}$ is a symmetric, positive, affinity matrix given by $\mathbf{A} = \mathbf{W}\text{diag}\left(\alpha\right)\mathbf{W}^T$, tr is the matrix trace, α a weight vector $\alpha \in \mathbb{R}^q$ with $\|\alpha\| = 1$, and $\mathbf{Q}$ an orthonormal $p \times p$ matrix. The difference between $\mathbf{W}$ and $\hat{\mathbf{W}}$ can be determined by the Euclidean norm with regard to $\mathbf{A}$, *i.e.*, $\left\|\mathbf{W} - \hat{\mathbf{W}}\right\|_{\mathbf{A}}^2$, where the squared is used to simplify. It can be shown that $\left\|\mathbf{W} - \hat{\mathbf{W}}\right\|_{\mathbf{A}}^2 = tr(A) \sum_{i=(q-q')+1}^{q} \lambda_i$, where λ_i denotes the i-th eigenvector from $\mathbf{W}^T\mathbf{W}$. Thereby, the purpose of minimizing this difference is that it could be represented in terms of the maximization of its complement, such as $tr\left(\mathbf{W}^T\mathbf{A}\mathbf{W}\right)$. Because $\mathbf{A} = \mathbf{W}diag\left(\alpha\right)\mathbf{W}^T$, with an initial value of $\alpha = 1_q$, Equation (8) could be introduced as:

$$\begin{aligned} tr\left(\mathbf{Q}^T \mathbf{W}\mathbf{W}^T \mathbf{W}\mathbf{W}^T \mathbf{Q}\right) = \\ tr\left(\mathbf{Q}^T \mathbf{A}\mathbf{A}\mathbf{Q}\right) = \sum_{i=1}^{r} \lambda_i^2 \end{aligned} \tag{9}$$

The Equations (8) and (9) are analogous to the quadratic optimization problem described by [33] and is solved by the algorithm Q-α. This stage is also unsupervised since the 98% variance criterion fixes

when the Q-α finishes the search without user intervention or training datasets. This method sorts the features according to their relevance measured in terms of contribution to whole variance, and when the threshold is reached, no more features are included in the final feature set. A full description of the method could be found in [17].

2.4. Sleep Stages Unsupervised Classifier

The automatic classification is based on a J-means approach [34]. The objective of this stage is to obtain an optimal partition of the feature vectors in $\hat{\mathbf{W}}$ in such a way that the resulting disjoint clusters correspond to different sleep stages. Additional clusters can be added in order to collect outliers, namely, artifacts that can not be classified as any sleep stage because their dissimilarity to any centroid is too large (noise, movement artifacts, interferences, sensor contact issues).

The algorithm attains a local optimum by heuristic reassignment of neighborhoods using a jumping paradigm [34], as follows:

(1) *Initialization:* A standard k-means clustering is used to set an initial partition of the feature vectors and the centroids. This reduces the temporal cost of the partition calculation.
(2) *Search:* Given a tolerance threshold (4 standard deviations of the intra-cluster distance), find the unoccupied points (feature vectors that do not belong to any cluster).
(3) *Update:* Add a new cluster centroid at some unoccupied location and find the index of the best centroid to delete. Update the partition according to the new centroids.
(4) *Finalize:* If a local minimum is found in the previous iteration, stop. For each resulting cluster, a sleep stage can be assigned as the most frequent class (using a k-Neighbors method), which in clinical practice could be done by a whole cluster manual scoring. Otherwise return to step 2.

3. Results and Discussion

The experiments were carried out using a standard Windows 8 personal computer and MATLAB$^{®}$ environment tools. The most relevant features obtained were (using feature names and EEG channels): FD.Fpz-Cz, FD.Pz-Oz, H.Fpz-Cz, DFA-α.Fpz-Cz, DFA-α.Pz-Oz, DFA-a1.Pz-Oz, ApEn.Fpz-Cz, ApEn.Pz-Oz, MSE1.Pz-Oz, SampEn2.Pz-Oz.

Table 1. Recall and precision for sleep stages identification using the set of relevant features. Results expressed as mean(variance).

| Stage	J-means		NN	
	Recall	**Precision**	**Recall**	**Precision**
N1	0.15 (0.28)	0.14 (0.26)	0.35 (0.23)	0.42 (0.24)
N2	0.91 (0.07)	0.84 (0.07)	0.84 (0.09)	0.89 (0.10)
N3	0.59 (0.43)	0.39 (0.29)	0.43 (0.22)	0.46 (0.24)
REM	0.38 (0.44)	0.34 (0.40)	0.75 (0.26)	0.52 (0.31)
W	0.84 (0.16)	0.87 (0.10)	0.93 (0.08)	0.73 (0.18)

Classification methods were tested for each subject independently (1048 ± 223 epochs per subject). Mean results are presented in Tables 1 and 2, for each patient in Figure 2, and globally in Figure 3, including results from standard PCA. Table 1 shows the performance in terms of Recall ($T_p/(T_p + F_n)$) and Precision ($T_p/(T_p + F_p)$) related to each sleep stage. The parameters T_p, F_n, and F_p are the normal values used to quantify the quality of classification results: true positives, false negatives, and false positives, respectively. For comparative purposes, this table also includes the results obtained using a supervised classification approach based on Neural Networks (NN).

Table 2. Performance achieved depending on the specific entropy metric employed and other common features for Sleep stage detection using the method proposed. Time is normalized (1.00 corresponds to the slowest case).

| Feature	Accuracy	Kappa	Time
FD	0.78 (0.06)	0.61 (0.13)	0.62 (0.04)
DFA	0.75 (0.06)	0.56 (0.14)	0.62 (0.05)
H	0.65 (0.09)	0.37 (0.12)	0.81 (0.10)
ApEn	0.74 (0.05)	0.54 (0.12)	0.56 (0.02)
SampEn	0.73 (0.06)	0.51 (0.16)	0.63 (0.04)
MSE	0.69 (0.06)	0.42 (0.13)	0.88 (0.08)
Absolute Power	0.74 (0.06)	0.53 (0.11)	1.00 (0.29)
Asymmetry	0.70 (0.07)	0.46 (0.08)	0.63 (0.03)
Central Power	0.70 (0.07)	0.47 (0.11)	0.69 (0.04)
Coherence	0.70 (0.07)	0.46 (0.12)	0.69 (0.04)
Phase Coherence	0.67 (0.08)	0.38 (0.12)	0.75 (0.04)
Power Ratios	0.80 (0.05)	0.67 (0.08)	0.75 (0.05)
Relative Power	0.77 (0.06)	0.61 (0.11)	0.69 (0.03)

**Figure 2.** Results obtained for each patient separately.
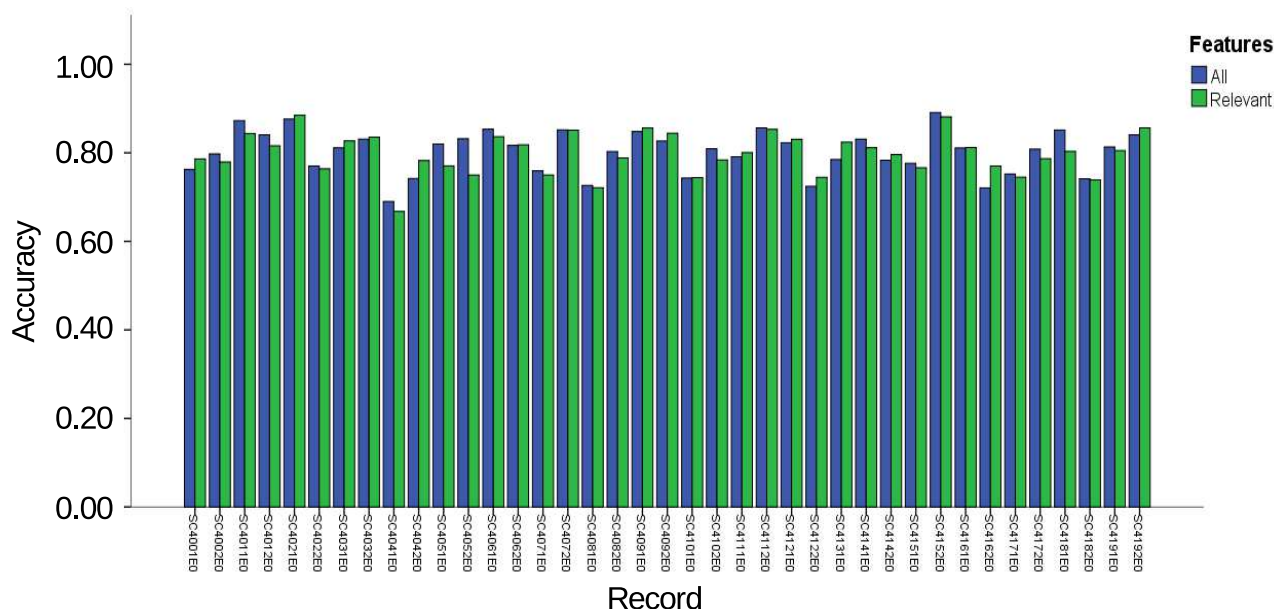
The method proposed exhibited comparable results to that of NN, despite being NN a supervised method (it includes labeled training patterns). The NN method outperforms the method proposed when detecting N1 and REM stages, but the opposite is found when detecting N2, N3 and REM sleep. These results support the studies that propose the use of unsupervised clustering techniques in order to address the sleep stage classification problem [35]. The performance archived by this sleep stage detection algorithm is similar (or superior) to previous results, for both supervised [23,36] and unsupervised applications [35]. Clustering methods do not depend on a previous training process, and therefore tend to be less influenced by signal differences among subjects or specific conditions [37]. The final partition groups are then scored. As an example, Figure 4 depicts a comparison between manual and automatic scoring results.

Figure 3. Performance of the NN and the proposed clustering-based classification method in terms of the set of features used (standard PCA or Q-α) for (**a**) accuracy and (**b**) Kappa coefficient. The results of the method proposed are highlighted (Average accuracy: 0.81).
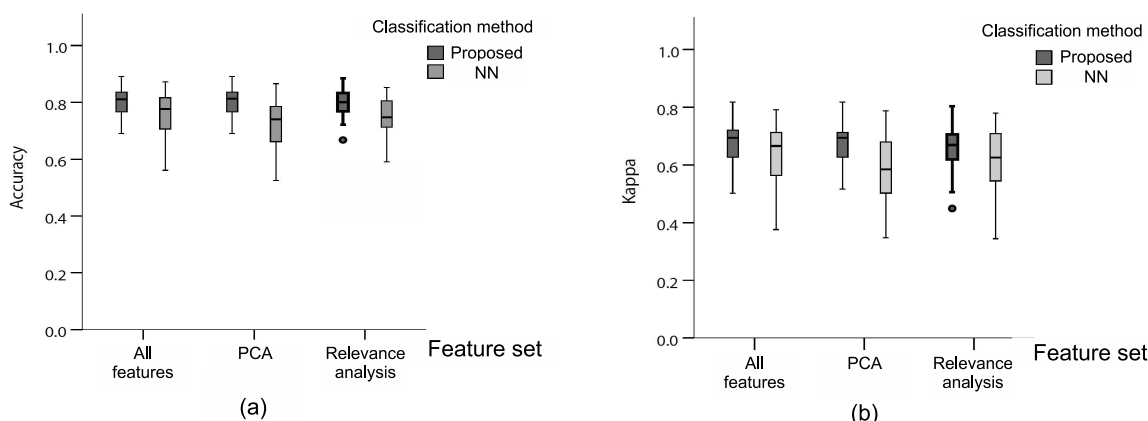
Figure 4. Example of result comparison using manual and automatic scoring. The hypnograms represent the class obtained using the method proposed in contrast to manual labels for a set of epochs from a single subject.
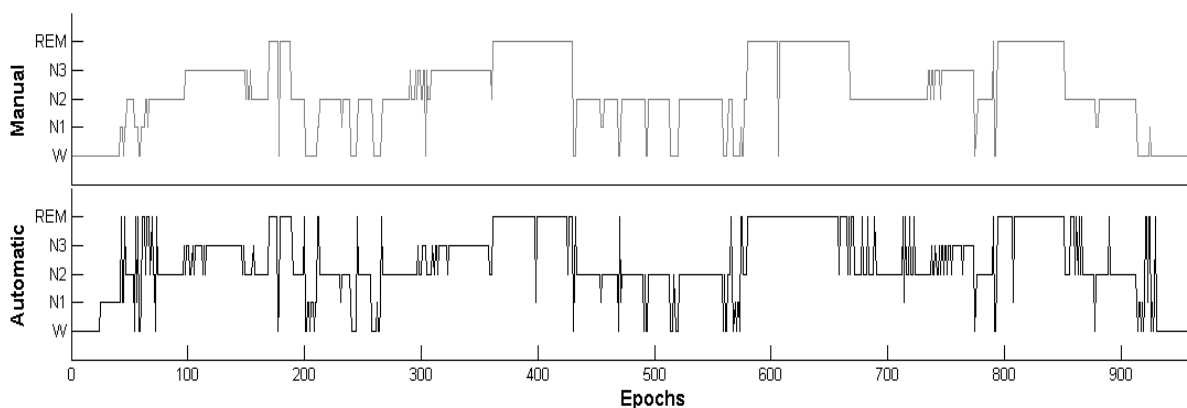
Recall and precision for sleep stages, with both classification methods, indicated better recognition of wakefulness and N2 stage (Table 1). These results are analogous to previous studies for automatic sleep stage scoring [38]. Also, results about the poor classification of N1 stage have been observed in previous inter-scorers agreement studies, specially when using the R&K rules [10,11], maybe due to the similarity

of the EEG activity described between REM and drowsy states [8], stages to which electromyographic (EMG) and electrooculographic (EOG) activity are fundamental for manual classification. In further studies, the inclusion of such signals could improve the performance. In fact, some works highlight the importance of these signals in automatic classification [39]. Some approaches even used a single electrode site to discriminate among sleep stages with Kappa values up to 0.74 using a supervised neural network approach [40], although no information about the signal features was included or specifications about the classifier.

Inter-scorer agreement for sleep stages has been reported to be around 0.68 Kappa and 76.8 accuracy. Similar results have been achieved using the proposed method (Figure 3). Also, our results are comparable to those found by novel automatic classification techniques that group epochs into wakefulness, REM and NREM sleep stages [41]. In this later study the use of six polygraphic channels (including EEG, EMG and EOG) led to a Kappa value around 0.51 ± 0.14, compared to manual scoring.

In another study [42], using linear discriminant analysis and entropy measures alone, they reached sensitivity and Kappa levels up to 76% and 0.65% respectively, which were higher than the ones obtained with the same measure in this work. Nevertheless, although we use the same SampEn and embedding dimension for Multiscale Entropy, the r value was set to a fixed value of 0.2 instead of including the standard deviation, which could account for the result difference.

In Table 2, the results are shown as the accuracy level $((T_p + T_n)/(T_p + F_n + T_n + F_p))$, Kappa statistic [43], and normalized computing time. These results are listed as a function of the specific entropy metric used for feature extraction: FD, DFA, H, ApEn, SampEn, and MSE. Although the results are quite similar both in terms of accuracy and computational cost, ApEn seems to reach an optimal tradeoff between these two issues, and FD yields the most accurate results. On the contrary, H and MSE offer a poorer performance, and should be discarded in favor of other entropy metrics.

Table 2 also shows the results achieved using the same method proposed in this paper, but computing classical features employed in most of the similar sleep stages studies instead of entropy metrics: the absolute power, power asymmetry, central power, coherence, phase coherence, power ratios, and relative power. Examples of application of these features and more details about them can be found in [36,44,45]. The classification results are similar to those obtained with the entropy metrics. Only the features based on the Power Ratios seem to outperform the previous results, but at a significantly higher computational cost. In fact, the main strength of the entropy features seems to be a general lower computational cost, with equal or even better accuracy.

In addition, the relevance analysis does not significantly appear to harm the classification performance of the method, as shown in Figure 3. The three cases tested: using all the raw features extracted, reducing the number of features using a standard feature selection method based on PCA, or the proposed relevance analysis method, yield a similar performance for Kappa and Accuracy. Furthermore, the NN method seems to be more sensitive to the feature selection stage, and in most of these experiment variants, the new method still outperforms NN.

This method has a very important novelty, which is using a non-supervised approach. Usually, this kind of methods perform worse than supervised ones since they do not have as much carefully selected input information and user feedback as these last have. The method proposed, not only does not underperform in comparison to supervised methods, but even exhibits a slightly better performance.

It improves the classification accuracy, but also exhibits faster calculations when the features selected were based on entropy estimators. Maybe the improvement is not very big in quantitative terms, but from the implementation point of view, this method can be more easily introduced in daily clinical practice since required user feedback is minimized. This method reaches a trade-off between being accurate enough as supervised methods, but requiring far less user intervention, the main advantage of unsupervised methods.

The sleeping brain has been described as a complex system with continuous, rather than discrete, transitions [46]. To study similarity of stages among subjects, the proposed method was applied to the whole set of subjects merged into a single matrix of 40,826 epochs by 56 relevant features.

With this method, a low similarity between subjects was found, specially in stages N1 and REM sleep. The low recall and precision values for these stages could be due to intra-subjects variability in the EEG signals, as described in [47,48]. Furthermore, these results are consistent with previous reports that suggest a significant EEG similarity between these stages [8], reflected in low discrimination values for both automatic classification [42,49] or inter-scorer agreement [11].

The global clustering time using PCA (0.30 ± 0.16 s) or the proposed Q-α optimization (0.20 ± 0.11 s) are lower than that needed for the complete set of features (0.69 ± 0.44 s). Therefore, the method proposed reduces considerably the computation time needed for an unsupervised sleep stage classification, while preserving the relevant features, and therefore, the accuracy [17]. Hence, the Q-α relevance analysis can be recommended as a competitive clustering technique for sleep stage classification, along with an FD, ApEn, or Power Ratio feature extraction method.

Figure 5. Results using an additional experimental database from subjects aged 18 to 79 and taking temazepam. The performance decreases but it also applies to other methods such as that based on NN.
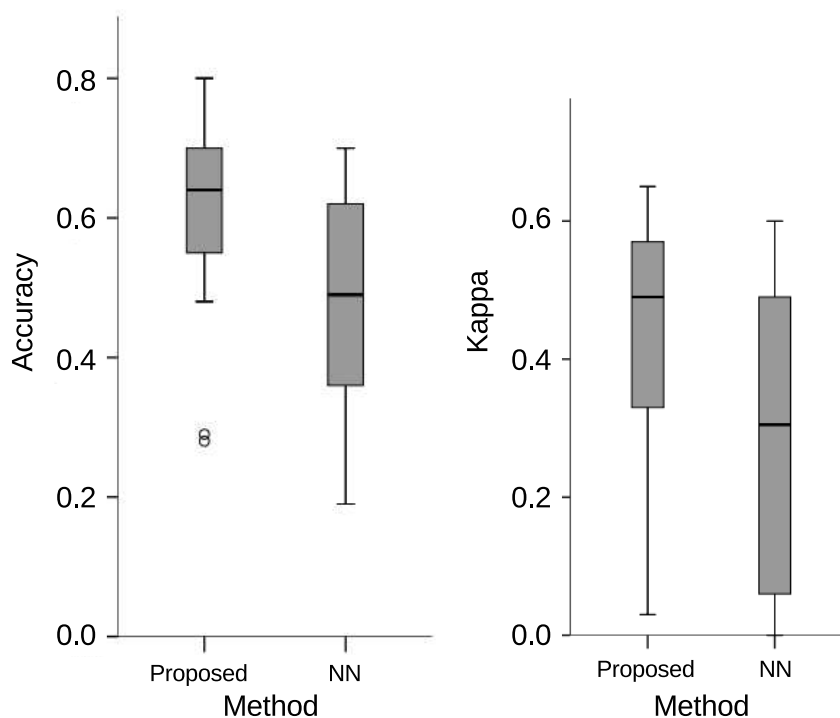
The performance was further validated using a different experimental database. It is also available at Physionet from a study of temazepam effects on sleep [19,50]. It contains 22 sleep files of male and female subjects (aged 18 to 79) sampled at 100 Hz. Half of the records were from subjects that took temazepam before going to sleep.

These additional results show a reduction of accuracy and Kappa values (Figure 5). This could be due to the greater subject variability in this dataset. For instance, age range and temazepam have different effects on sleep EEG dynamics [51]. However, this reduction is higher with the NN approach compared to the proposed one. The resulting confusion matrix is presented in Table 3.

Table 3. Confusion matrix for the automatic procedure computing the whole set of 40,826 epochs for all the patients together to test inter-subject variability.

		Prediction outcome				
		W	**N1**	**N2**	**N3**	**REM**
Actual value	**W**	3333	2046	1074	21	329
	N1	177	1082	624	39	882
	N2	884	915	8155	1198	6647
	N3	42	69	1539	4255	400
	REM	484	1017	1738	25	3851

4. Conclusions

This paper describes a new method to automatically detect sleep stages on EEG records, which is of great medical and social interest. This method proposes to use entropy metrics as the features of the EEG epochs to be analyzed, and also introduces a feature selection stage to reduce the algorithm cost.

The core of the method is the so-called Q-α relevance analysis. This method has been successfully used in similar applications to ECG signals [17], and it is exported in this work to the new field of EEG relevant feature selection. The reduction of the number of features is confirmed not to damage the performance of the method, keeping the segmentation capability almost intact, as should be expected.

The study explores the capabilities of a plurality of entropy metrics for comparative purposes. The results confirm the validity of the approach, from which it can be concluded that almost any metric could perform satisfactorily, specially ApEn and FD.

The experiments were repeated using more classical methods in this field such as NN. Furthermore, the feature extraction stage was also replaced by a standard stage based on other signal parameters. In all cases, the performance of the new method was at least as accurate as the references assessed, or even more on most cases, but at a low computational cost, and being an unsupervised method (no user intervention needed).

Most elements of the proposed method are already implemented in a MATLAB$^{\circledR}$ library called EEGLAB, an open source tool for assisted sleep staging. This software application was previously presented in [52].

Current efforts in most entropy-based EEG studies are still mainly focused on the identification of certain patterns (such as seizure detection [53]). Other works employ entropy as a metric for diagnostic

purposes, to discriminate among subjects with neural alterations (citeGarn2014 Bachiller2014. Our results, in addition to proposing a suitable method for automatic sleep stages detection, also suggest a differential effect of sleep stages on EEG entropy features. These results have been also supported by studies of anesthesia depth [54]. EEG features are also very important to understand sleep dynamics. A comparison of the entropy features among sleep stages, along with an analysis of the significance or meaning of this kind of features, is still object of research work.

In summary, we proposed in this paper a method to classify sleep stages based on computing entropy metrics of EEG records, optimizing the resulting set using a relevance analysis, and including a clustering algorithm to obtain the final partition. Such a method exhibits better flexibility and adaptability in comparison to other supervised methods. It also reduces the computational cost, which is very important, taking into account the impending electronic health record explosion.

Acknowledgments

Author Contributions

Jose Luis Rodríguez-Sotelo, Alejandro Osorio-Forero, Alejandro Jiménez-Rodríguez and Diego Peluffo contributed equally to the development of the research of the sleep stages classification, including the selection of the experimental dataset, the implementation of the feature extraction and relevance stages, and the clustering algorithm. They wrote the initial version of the paper.

David Cuesta-Frau led the research project TEC 2009-14222. He defined the experiments related to entropy metrics, the specific functions to use, and the scope of the results. He also defined the initial structure of the paper, the final version of the paper, the revisions, and the responses to the reviewers.

Eva Cirugeda-Roldán studied the entropy metrics, implemented their code, and carried out the specific tests using these metrics.

All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Saper, C.B.; Fuller, P.M.; Pedersen, N.P.; Lu, J.; Scammell, T.E. Sleep state switching. *Neuron* **2010**, *68*, 1023–1042.
2. Halasz, P.; Bodizs, R. *Dynamic structure of NREM sleep*; Springer: London, UK 2013.
3. Rauchs, G.; Desgranges, B.; Foret, J.; Eustache, F. The relationships between memory systems and sleep stages. *J. Sleep Res.* **2005**, *14*, 123–140.

4. Landmann, N.; Kuhn, M.; Piosczyk, H.; Feige, B.; Baglioni, C.; Spiegelhalder, K.; Frase, L.; Riemann, D.; Sterr, A.; Nissen, C. The reorganisation of memory during sleep. *Sleep Med. Rev.* **2014**, *18*, 531–541.

5. Hublin, C.; Partinen, M.; Koskenvuo, M.; Kaprio, J. Sleep and mortality: A population-based 22-year follow-up study. *Sleep* **2007**, *30*, 1245–1253.

6. Lovin, P.A.; Ehrenpreis, A.B. The role of polysomnography in the differential diagnosis of chronic insomnia. *Am. J. Psychiatry* **1988**, *145*, 346–349.

7. Steriade, M.; McCormick, D.A.; Sejnowski, T.J. Thalamocortical oscillations in the sleeping and aroused brain. *Science* **1993**, *262*, 679–685.

8. Niedermeyer, E., da Silva, F., Eds. *Electroencephalography*; Lippincott Williams and Wilkins: Philadelphia, PA, USA, 2005.

9. Iber, C. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*; American Academy of Sleep Medicine: Westchester, NY, USA, 2007.

10. Hopfe, H.D.; Anderer, P.; Zeitlhofer, J.; Boeck, M.; Dorn, H.; Gruber, G.; Heller, E.; Loretz, E.; Moser, D.; Parapatics, S. Interrater reliability for sleep scoring according to the Rechtschaffen Kales and the new AASM standard. *J. Sleep Res.* **2009**, *18*, 74–84.

11. Hopfe, H.D.; Kunz, D.; Gruber, G.; Klosch, G.; Lorenzo, J.L.; Himanen, S.L.; Kemp, B.; Penzel, T.; Roschke, J.; Dorn, H. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J. Sleep Res.* **2004**, *13*, 63–69.

12. Fraiwan, L.; Khaswaneh, N.; Ylweesy, K. Automatic sleep stage scoring with Wavelet Packets based on single EEG recording. *Proc. World Acad. Sci. Eng. Technol. Paris* **2009**, *54*, 385–488.

13. Vuckovic, A.; Radivojevic, V.; Chen, A.; Popovic, D. Automatic recognition of alertness and drowsiness from EEG by an artificial neural network. *Med. Eng. Phys.* **2002**, *24*, 349–360.

14. Robert, C.; Guilpin, C.; Limoge, A. Review of neural network applications in sleep research. *J. Neurosci. Methods* **1998**, *79*, 187–193.

15. Ronzhina, M.; Janousek, O.; Kolarova, J.; Novakova, M.; Honzik, P.; Provaznik, I. Sleep scoring using artificial neural networks. *Sleep Med. Rev.* **2012**, *16*, 251–263.

16. Subasi, A.; Ercelebi, E. Classification of EEG signals using neural network and logistic regression. *Comput. Methods Programs Biomed.* **2005**, *78*, 87–99.

17. Rodriguez-Sotelo, J.L.; Peluffo-Ordonez, D.; Cuesta-Frau, D.; Castellanos-Dominguez, G. Unsupervised feature relevance analysis applied to improve ECG heartbeat clustering. *Comput. Methods Programs Biomed.* **2012**, *108*, 250–261.

18. Kemp, B.; Zwinderman, A.; Tuk, B.; Kamphuisen, H.; Oberye, J. Analysis of a sleep-dependent neural feedback loop: The slow-wave microcontinuity of the EEG. *IEEE–BME* **2000**, *9*, 1185–1194.

19. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220.

20. Van Sweden, B.; Kemp, B.; Kamphuisen, H.; van der Velde, E. Alternative electrode placement in (automatic) sleep scoring (Fpz-Cz / Pz-Oz versus C4-A1 / C3-A2). *Sleep* **1990**, *3*, 279–283.

21. Mourtazaev, M.; Kemp, B.; Zwinderman, A.; Kamphuisen, H. Age and gender affect different characteristics of slow waves in the sleep EEG. *Sleep* **1995**, *7*, 557–564.

22. Rechtschaffen, A.; Kales, A. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*; US Department of Health, Education, and Welfare: Bethesda, MD, USA, 1968.

23. Fraiwan, L.; Lweesy, K.; Khasawneh, N.; Wenz, H.; Dickhaus, H. Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. *Comput. Methods Programs Biomed.* **2012**, *108*, 10–19.

24. Raghavendra, B.S.; Dutt, N.D. Computing fractal dimension of signals using multiresolution box-counting method. *Int. J. Inf. Math. Sci.* **2010**, *6*, 50–65.

25. Shoupeng, S.; Peiwen, Q. A fractal-dimension-based signal processing technique and its use for nondestructive testing. *Russ. J. Nondestruct. Test.* **2007**, *43*, 270–280.

26. Peng, C.K.; Buldyrev, S.V.; Havlin, S.; Simons, M.S.; Eugene, H.; Goldberger, A.L. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **1994**, *49*, 1685–1689.

27. Peng, C.; Havlin, S.; Stanley, H.; Goldberger, A. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos* **1995**, *5*, 82–87.

28. Fell, J.; Roschke, J.; Mann, K.; Schaffner, C. Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures. *Electroencephalogr. Clin. Neurophysiol.* **1996**, *98*, 401–410.

29. Pincus, S. Approximate entropy (ApEn) as a complexity measure. *Chaos Interdiscip. J. Nonlinear Sci.* **1995**, *5*, 110–117.

30. Richman, J.S.; Moorman, J.R. Physiological time–series analysis using Approximate Entropy and Sample Entropy. *Am. J. Physiol.–Heart Circulatory Physiol.* **2000**, *278*, H2039–H2049.

31. Costa, M.; Goldberger, A.L.; Peng, C.K. Multiscale entropy analysis of biological signals. *Phys. Rev. E* **2005**, *71*, 021906.

32. Toennies, K.D.; Celler, A.; Blinder, S.; Moeller, T.; Harrop, R. R.: Scatter segmentation in dynamic SPECT images using principal component analysis. In Proceedings of the SPIE (Medical Imaging 2003), San Diego, CA, USA, Part I, 15 February 2003; pp. 507–516.

33. Wolf, L.; Shashua, A. Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weighted-based approach. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 378–384.

34. Hansen, P.; Mladenovic, N. J-Means: a new local search heuristic for minimum sum of squares clustering. *Pattern Recognit.* **2001**, *34*, 405–413.

35. 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*, 7922–7928.

36. Koley, B.; Dey, D. An ensemble system for automatic sleep stage classification using single channel EEG signal. *Comput. Biol. Med.* **2012**, *42*, 1186–1195.

37. Hese, P.V.; Philips, W.; Koninck, J.D.; de Walle, R.V.; Lemahieu, I. Automatic detection of sleep stages using the EEG. In Proceedings of the 23rd Annual International Conference of the Engineering in Medicine and Biology Society, 25–28 October 2001; Volume 2, pp. 1944–1947.

38. Oropesa, E.; Cycon, H.L.; Jobert, M. Sleep stage classification using Wavelet Transform and neural network. *Int. Comput. Sci. Inst.* **1999**. Available online: http://www.researchgate.net/publication/216570220_Sleep_Stage_Classification_Using_Wavelet_Transform__Neural_Network (accessed on 10 December 2014)

39. Krakovska, A.; Mezeiova, K. Automatic sleep scoring: A search for an optimal combination of measures. *Artif. Intell. Med.* **2011**, *53*, 25–33.

40. Shambroom, J.R.; Fabregas, S.E.; Johnstone, J. Validation of an automated wireless system to monitor sleep in healthy adults. *J. Sleep Res.* **2012**, *21*, 221–230.

41. Swarnkar, V.; Udantha, R.A. Bispectral analysis of single channel EEG to estimate macro-sleep-architecture. *Int. J. Med. Eng. Inform.* **2014**, *6*, 43–64.

42. Liang, S.F. Automatic stage scoring of single-channel sleep EEG by using Multiscale Entropy and autoregressive models. *IEEE Trans. Instrum. Meas.* **2012**, *61*, 1649–1657.

43. Viera, A.; Garrett, J.M. Understanding interobserver agreement: The Kappa statistic. *Fam. Med.* **2005**, *5*, 360–363.

44. Weiss, B.; Clemens, Z.; Bodizs, R.; Halasz, P. Comparison of fractal and power spectral EEG features: Effects of topography and sleep stages. *Brain Res. Bull.* **2011**, *84*, 359–375.

45. Susmakova, K.; Krakovska, A. Discrimination ability of individual measures used in sleep stages classification. *Artif. Intell. Med.* **2008**, *44*, 261–277.

46. Eckehard, O.; Achermann, P.; Wennekers, T. The sleeping brain as a complex system. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **2011**, *369*, 3697–3707.

47. Buckelmuller, J. Trait-like individual differences in the human sleep electroencephalogram. *Neuroscience* **2006**, *138*, 351–356.

48. Dongen, H.P.V.; Vitellaro, K.M.; Dinges, D.F. Individual differences in adult human sleep and wakefulness. Leitmotif for a research agenda. *Sleep* **2005**, *28*, 479–496.

49. Fraiwan, L. Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. *Comput. Methods Programs Biomed.* **2012**, *108*, 10–19.

50. Kemp, B.; Janssen, A.; Roessen, M. A digital telemetry system for ambulatory sleep recording. *Sleep-Wake Research in The Netherlands* **1993**, *4*, 129–132. Available online: http://physionet.mit.edu/pn4/sleep-edfx/Papers/1993-Kemp—telemetry.pdf (accessed on 10 December 2014).

51. Dijk, D.J. Effects of seganserin, a 5-HT2 antagonist, and temazepam on human sleep stages and EEG power spectra. *Eur. J. Pharmacol.* **1989**, *171*, 207–218.

52. Rodriguez-Sotelo, J.; Osorio-Forero, A.; Jimenez-Rodriguez, A.; Restrepo, F. A new tool for assisted sleep staging and transitory sleep patterns analysis in EEG signals. In Proceedings of the IX Congreso Nacional X Seminario Internacional de NEUROCIENCIAS. COLNE., 15–17 May 2014; pp. 163–164.

53. Zhang, Z.; Chen, Z.; Zhou, Y.; Du, S.; Zhang, Y.; Mei, T.; Tian, X. Construction of rules for seizure prediction based on approximate entropy. *Clin. Neurophysiol.* **2014**, *125*, 1959–1966.

54. Khan, J.; Venkatraghavan, L.; Mariappan, R. Entropy as an indicator of cerebral perfusion in patients with increased intracranial pressure. *J. Anaesthesiol. Clin. Pharmacol.* **2014**, *30*, 409–411.