An Introduction and Comparison of Automatic Sleep Stage Scoring Methods

William King

Abstract

Introduction: Sleep stage scoring is important in determining medical information to diagnose patients with potential sleep disorders. While visual scoring of sleep stages is still in use, the use of an automatic sleep stage scoring process would be significantly faster and more efficient. This literature review serves to discuss the automatic sleep stage scoring process in terms of feature extraction, feature selection, and feature classification, while analyzing previous studies to determine suitable classifiers and features to use.

Methods: This review explores various types of features and methods used in the automatic sleep scoring processes of extraction, selection, and classification. These methods use single-channel electroencephalogram (EEG) data. Moreover, three selected studies that have implemented some of these methods are analyzed and evaluated to determine the pairing of features and classifiers that provides the greatest scoring accuracy.

Results: Based on the results, the use of time-frequency domain features with the use of wavelet transformation is the best feature to use. These features are best paired with the artificial neural network classifier, providing an 89% accuracy in worst case scenarios and a 97% accuracy in best case scenarios.

1. Introduction

Sleep is essential to the body and mind, rejuvenating them so that one can carry out their daily tasks. By cycling through the various sleep stages, one can obtain these physical and mental benefits. In stage 1 sleep, light sleep occurs, in which one can easily be awakened. Eye movement and muscle activity slow down, and random muscle contractions can occur[1]. In stage 2 sleep, eye movement stops completely, and brainwave activity slows down[1]. In stage 3 sleep, slow moving waves called delta waves begin to appear[1]. Upon reaching stage four, delta waves are exclusively produced, and deep sleep is entered, in which one cannot easily be awoken[1]. Finally, in the REM stage, eye movement becomes erratic and temporary paralysis occurs[1]. It is also in the REM stage that dreams occur[1]. REM stage sleep is often viewed as the most beneficial sleep stage, in which protein production and cognitive functionality is increased[1].

However, people are often deprived of this sleep stage, as well as other sleep stages due to sleep disorders of varying severity. These disorders are prevalent in our society, with around 60 million Americans experiencing sleep disorders such as insomnia, sleep apnea, and narcolepsy[1]. To diagnose and distinguish between those with mild sleep issues and those with more severe sleep disorders, sleep stage scoring is used, in which EEG analysis is used to extract data on an individual's sleep signals during the different sleep stages[2]. Although the manual process of visual sleep scoring is commonly used to an 83% accuracy, automatic processes that utilize

machine learning algorithms have been investigated in order to make the scoring process less expensive and more efficient[2].

Although automatic sleep stage scoring has abundant potential, the process has yet to be perfected. While it has achieved a greater accuracy of about 87% in healthy individuals, it has only achieved an accuracy of about 69% in patients [2]. This is due to the fluctuating occurrence of EEG signal bands in sleep patients, which lowers the accuracy of the EEG signals being recorded[2]. This lower accuracy rate poses a potential risk of misdiagnosis, which can have a negative impact on patient's health and well being. Due to this accuracy rate being lower than 75%, the process is considered substandard.

This paper reviews a series of automatic scoring methods used to analyze a patient for sleep disorders, and determines the best method used. Section 2 explains the various features used in the scoring process, as well as extraction, selection, and classification methods used in the automatic scoring process. Section 3 describes automatic scoring methods in detail. In section 4, the results of these methods will be shown. The final section is dedicated to a conclusion.

2. Materials and Methods

In this section, features used to identify sleep disorders in the scoring process will be described. Then, the processes in which these features are extracted, selected, and classified will be explained in detail.

2.1 Types of Features

Features are special aspects of signals that describe certain characteristic quantitatively. All features fall into four categories: time domain, frequency domain, time-frequency domain, and nonlinear features.

- Time Domain Features: Time domain features are real time features that are computed directly from the EEG signal. The most notable time domain features include Hjorth parameters and statistical parameters.
    - *Statistical Parameters:* mean, variance, standard deviation, skewness, kurtosis, threshold percentile, and median are statistical parameters that are commonly used. By applying these to EEG signals, relevant features can be extracted.[2]
    - *Hjorth Parameters:* Hjorth parameters were introduced by Bo Hjorth in 1970, and utilize three parameters to indicate statistical properties: activity parameters to measure the variance of a time series, mobility parameters to proportionate standard deviation of the power spectrum, and complexity parameters, to determine the change in frequency.[3]
- Frequency Domain Features
    - Frequency domain features are repeatedly used for describing changes for EEG signals with respect to their frequency of occurrence.[2]
- Time-frequency Domain Features

    - Time-frequency domain features utilize both the time and frequency domain, determining which frequencies occur at a specific time point.[2]
- Non-linear Features
    - Non-linear features consist of complex, non-linear characteristics, and are commonly found in EEG signals. As a result, they are widespread in the sleep stage scoring process.[2]

2.2 Feature Extraction

When automatically scoring the sleep stages, it is important to extract relevant, discriminative, and independent data, as this helps score in subsequent learning and generalization steps.[2] For the same reason, it is also important to reduce the amount of redundant information collected. [2] Feature extraction is the process of this extraction. Depending on the type of feature, various methods can be used to extract this information from the EEG signals. This information is extracted in sets of epochs, or 30 second intervals. [2]

Fourier Transformation is a commonly used means to extract specific features from a specified type of signal. [2] In this process, time signals are broken down into the frequency signals that it is made of, allowing for features to be transformed from the time domain into the frequency domain. [2] Conversely, *Inverse Fourier Transformation* allows for a feature to be transformed from the frequency domain back into the time domain.[4] Through the use of these two transformations, certain characteristics of signals that were previously unobtainable can be obtained.

Time Domain: Both statistical and Hjorth parameters utilize mathematical formulas to extract features, some of which are described below:[2]

Statistical Parameter Features	**Formula**	**Notes**
Mean (μ) – the average value of a dataset	$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$	X_n (n = 1, 2, 3, 4, …. N-2, N-1, N) is a time series.
Variance (var) – a measure of how far a set of numbers is spread out form the mean	$var = \frac{1}{N-1}\sum_{n=1}^{N} (x_n - \mu)^2$	X_n (n = 1, 2, 3, 4, …. N-2, N-1, N) is a time series. μ is the mean.
Standard Deviation (std) – a measure of dispersion of a dataset	$std = (var)^{\frac{1}{2}}$	Std is standard deviation. Var is the variance.
Median (M)	$M = \{\frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} for\ even\ N\ x_{\frac{N+1}{2}}\ f$	X_n (n = 1, 2, 3, 4, …. N-2, N-1, N) is a time series. N is the number of values in the series.
Hjorth Parameter Features	**Formula**	**Explanation**

Activity (H_a)	$H_a = \sqrt{var(x(t))}$	X is the measure of variance of a time series. Var is the variance.
Mobility (H_m)	$H_m = \sqrt{\frac{var(x(t)\frac{dx}{dt})}{var(x(t))}}$	X is the measure of variance of a time series. Var is the variance.
Complexity (H_c	$H_c = \frac{Hm(x(t)\frac{dx}{dt})}{Hm(x(t))}$	X is the measure of variance of a time series. Var is the variance.

Table 1: Time Domain Features

Frequency Domain:

- *Spectral Estimation* is a commonly used method of data analysis used to extract the spectral characteristics of EEG signals[5]. In this process, data from an extracted time series are compared to an artificially created time series consisting of sine and cosine functions[5]. Depending on the transformations applied to the sine and cosine functions, varying amount of cycles will occur over a period of time, changing the frequency of the signal[5]. When these sine and cosine frequencies are added together at different frequencies and amplitudes, the artificial time series is created[5]. In using spectral estimation in combination with Fourier Transformation, the time series can be re-expressed in a standard way, allowing for different time series to be more easily compared using the Power Spectral Density (PSD)[2].
    - *Parametric methods* are model-based approaches that use the signal model to create an estimate of the spectrum[2]. Parametric methods include autoregressive (AR), Moving average (MA) and Autoregressive moving average (ARMA)[2]. Parametric methods are best used when the signal being examined has a low signal-to-noise ratio and a long length[2].
    - *Non-parametric methods* calculate PSD values directly from signal samples of an epoch utilizing Fourier transformation.[2] Periodogram and Welch schemes are commonly used alongside the Fast Fourier Transform (FFT) algorithm for easy implementation.[2] However, non-parametric methods are unsuited for short signal intervals.[2]

Time-frequency Domain:

- *Signal Decomposition* is the act of breaking down signals to a series of basis functions.[2] This is most easily done using Short Time Fourier Transformation, in which Fourier Transformation is applied to each window of signals.[2] Another popular transformation utilized in signal decomposition is Wavelet Transformation (WT), in which various filters are applied to decompose a signal into coupled frequency scales.[2]
- *Energy Distribution* displays the distribution of energy through both the time and frequency domains.[2] The Choi-Williams distribution is often used to do this, and is an accurate method that conserves energy.[2] The smoothed pseudo Wigner-Ville distribution can also be used, and is capable of analyzing non-stationary signals.[2]

2.3 Feature Selection

Once the most relevant features have been extracted, redundancy can be further reduced through feature selection. Feature selection techniques are utilized to find discriminative subsets, or combinations, of features.[6] Upon finding these variables or features, they are selected into a new dataset, and remaining variables are no longer analyzed.[6] In doing this, irrelevant and redundant data can be further reduced.[6]

There are two different general approaches to feature selection: the exhaustive approach and the heuristic approach.[2] The exhaustive approach is most commonly used, evaluating every single possible subset of features.[2] This approach can be very tedious, as sometimes there are millions of possible feature subsets.[2] However, it ensures that the best possible subsets are chosen.[2] Conversely, the heuristic approach uses general rules of thumb to create sufficient methods of selecting features, often providing a good solution using less resources.[2]

Specific feature selection methods include Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), and Principle Component Analysis (PCA).[2]

- SFS takes relevant features and iteratively adds them to an empty set until a certain condition is met, creating a single feature subset as a solution.[6] This follows the heuristic approach, using the condition as a guideline to select features.[6] Once added to the subset, features cannot be removed.[6]
- SBS takes a subset full of features and removes them until a certain condition is met.[6] Similarly to SFS, the heuristic approach is used, removing these features until an acceptable feature subset is created.[6] SBS requires more computational power than SFS, and conversely cannot add features once they have been removed.[6]
- PCA identifies appropriate variables, or principal components, that are used to plot items on a graph such that data can be well differentiated.[6] By using a principal component, data can be plotted differently on the graph, creating unique clusters that provide different analysis.[6] Multiple components can also be used at once, creating a more specific, wider spread of information.[6]

2.4 Feature Classification

Feature classification uses classifiers to analyze data and identify characteristics. In sleep stage scoring, feature classification is used to assign a sleep stage to the epochs analyzed.[2] The classifiers applied can have linear or nonlinear boundaries, allowing for feature vectors of different classes to be separated.[2] Commonly used classifiers include K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM).

- KNNs are best used as a classifier when handling data that has multiple modes.[2] By assigning labels to an input based off of a majority vote of the k nearest samples, data that is significantly spread out can be classified (where k is a user defined constant).[2]
- SVMs are best used as a classifier when errors do not occur while handling data.[2] Rather than minimizing the impact of errors like other classifiers, the SVM minimizes the

probability of an error occurring.[2] This is done through a function that maximizes margin width around the separating plane of two classes, minimizing training error and thus increasing accuracy.[2]

- RFs utilize a series of tree structures to classify data, in which data is randomly inputted into the system multiple times.[2] After, roughly two thirds of the data is selected to be used.[2] These tree structures are trained separately of each other such that they are more robust, and can better handle interference.[2]
- ANNs utilize a system made up of connected neural nodes, similar to the human mind, to classify data.[2] There can be multiple layers to an ANN, which contain more data for large datasets.[2] ANNs can be particularly useful in helping computers emulate human thought and action.[2]
- GMMs are statistical models that utilize Gaussian functions and coefficients to estimate a continuous probability density of a Gaussian component.[2] This is done through the use of expectation maximization algorithms, which find the most likely outcomes and use them to determine the labels used for the data.[2]
- HMMs are a comprehensive version of the Markov Chain Model, in which a series of algorithms are used to classify information.[2] The Baum-Welch algorithm is used to determine parameters, and the Viterbi algorithm is then used to find the sequence of the data states.[2] This information is then used to find the output of the Markov Model.[2] The HMM can be beneficial to use as it is a dynamic classifier that can properly handle inputs that have been time warped.[2]

3. Studies of Automatic Sleep-Stage Scoring Methods

The studies featured in this review utilize a combination of the above feature extraction, selection, and classification methods in their sleep stage classification. The methods discussed utilize single-channel EEG data.

3.1 Frequency Domain Features and Hidden Markov Model

In this study, data was gathered from 20 subjects between the ages of 19 and 27 over a series of 30 second epochs.[7] Various frequency domain features were extracted, using the FFT to convert time features to frequency features where needed.[7] These features were then paired with the Discrete Hidden Markov Model (DHMM), a specific HMM that is more stable and accurate, serving as the classifiers.[7] 10 subjects were used to train the DHMM, while the other 10 subjects were used to test the DHMM.[7] The accuracy percentages of each test subject can be seen below:[7]

Subject No.	Wake (%)	S1 (%)	S2 (%)	SWS (%)	REM (%)	Overall (%)
1	100	38.43	86.89	93.97	85.19	85.22
2	95.31	67.35	83.73	87.5	78.76	83.71

3	100	24.66	89.42	100	90.44	90.41
4	93.33	8.53	69.53	100	83.91	78.95
5	100	31.33	67.77	100	87.82	77.09
6	75	28.77	93.78	96.84	87.26	92.14
7	100	22	90.84	100	97.33	92.62
8	100	9.11	76.97	78.91	94.96	80.26
9	84.48	48	71.59	97.51	100	84.11
10	40	58	85.25	94.44	95.76	88.35
Mean (std)	88.81 (19.1)	33.62 (19.1)	81.58 (9.4)	94.92 (6.9)	90.14 (6.8)	85.29 (5.5)

Table 2: Results from Study 1 (Frequency Domain and AHMM)

3.2 Time-frequency Domain Features with Support Vector Machine and k-Nearest Neighbour

In this study, data was gathered from 17 subjects (five men and 12 women) between the ages of 26 and 67 over a night's worth of 30 second epochs (mean duration 7 hours).[8] Time-frequency domain features were extracted after preprocessing the data and using WT, and were paired with the SVM and KNN classifiers.[8] For each classifier, the samples were divided into training and test sets. The average accuracy percentage of each classifier can be seen below:[8]

Classification	**Sleep Stages**						**Total**
Method	**Wake**	**NREM 1**	**NREM2**	**NREM3**	**NREM4**	**REM**	**Accuracy**
kNN	322/350	105/109	2159/3773	352/390	775/943	655/836	4414/6407
	= 92%	= 97%	= 58.2%	= 90.3%	= 82.2%	= 78.4%	= 68.9%
SVM	334/350	107/109	2328/3773	368/390	824/943	709/836	4684/6407

	= 95.6%	= 98.5%	= 61.8%	= 94.3%	= 87.4%	= 84.9%	= 73.1%

Table 2: Results from Study 2 (Time-frequency and kNN/SVM)

3.3 Time-frequency Domain Features and Artificial Neural Networks

In this study, data was gathered from 14 subjects (seven men and seven women) between the ages of 21 and 35 over a series of 30 second epochs.[9] Time-frequency domain features were extracted after transforming them using WT.[9] These features were then paired with the ANN classifier, consisting of one hidden layer and 12 inputs.[9] The average accuracy percentage of the ANN can be seen below:[9]

	Sleep Stages				**Total**
	Wake	**Stage1 + REM**	**Stage2**	**SWS**	**Accuracy**
Specificity	99.2±0.2	93.8±0.9	87.6±2.0	97.0±0.7	94.4±4.5
Sensitivity	79.7±3.6	85.7±2.5	87.5±1.9	84.0±2.7	84.2±3.9
Accuracy	98.5±0.2	91.4±0.5	87.5±0.6	94.6±0.4	93.0±4.0

Table 3: Results from Study 3 (Time-frequency and ANN)

Results and Conclusion

From the results of these three studies, the methods used to classify sleep stages can be evaluated primarily in terms of accuracy. In terms of feature extraction, the use of time-frequency domain features proved to be very useful. Although time domain, frequency domain, and nonlinear features can also be beneficial in certain scenarios, oftentimes it is beneficial for both the frequency components and the timing of EEG signals to be known.[8] Techniques such FT can help to extract time-frequency domain features, even if they are not present to begin with.[8] When deciding which classifier to best pair these features with, the ANN method boasts the highest accuracy of 89% in worst case scenarios and 97% in best case scenarios, while also producing satisfactory results in terms of specificity and sensitivity.[9] Depending on the situation and the

types of features extracted, other classifiers may be more beneficial or simplistic to use. However, of the studies featured in this review, the ANN classifier presents the most accurate classifier possible.

Abbreviations

Term	Acronym
Electroencephalogram	EEG
Rapid Eye Movement	REM
Sequential Forward Analysis	SFA
Sequential Backward Analysis	SBA
Principal Component Analysis	PCA
k-nearest Neighbour	KNN
Support Vector Machine	SVM
Artificial Neural Network	ANN
Random Forest	RF
Gaussian Mixture Model	GMM
Hidden Markov Model	HMM

Keywords

automatic sleep stage scoring; signal processing; features;

Acknowledgements

I would like to express my sincerest thanks to the Foundation of Student Science and Technology for the providing myself with the opportunity to participate in the Online Research Co-operative Program. I would like to specifically thank my mentor, Ahnaf Rashik Hassan, for his guidance and mentorship throughout the creation of this paper, as well as my co-op coordinator, Allen Flemington, for his constant support throughout the placement.

Bibliography

(1) American Sleep Association. What Is Sleep? Latest Research & Treatments | American Sleep Assoc https://www.sleepassociation.org/patients-general-public/what-is-sleep/ (accessed Nov 30, 2017).

(2) Boostani, R.; Karimzadeh, F.; Nami, M. A Comparative Review on Sleep Stage Classification Methods in Patients and Healthy Individuals. *Comput. Methods Programs Biomed.* **2017**, *140*, 77–91.

(3) Hjorth, B. EEG Analysis Based on Time Domain Properties. *Electroencephalogr. Clin. Neurophysiol.* **1970**, *29* (3), 306–310.

(4) Folland, G. B. *Fourier Analysis and Its Applications*; American Mathematical Society, 2009.

(5) Simões, H.; Pires, G.; Nunes, U.; Silva, V. FEATURE EXTRACTION AND SELECTION FOR AUTOMATIC SLEEP STAGING USING EEG.

(6) Jain, A. Feature Selection http://www.cse.msu.edu/~cse802/Feature_selection.pdf (accessed Jan 6, 2018).

(7) Pan, S.-T.; Kuo, C.-E.; Zeng, J.-H.; Liang, S.-F. A Transition-Constrained Discrete Hidden Markov Model for Automatic Sleep Staging. *Biomed. Eng. Online* **2012**, *11* (1), 52.

(8) Yılmaz, B.; Asyalı, M. H.; Arıkan, E.; Yetkin, S.; Özgen, F. Sleep Stage and Obstructive Apneaic Epoch Classification Using Single-Lead ECG. *Biomed. Eng. Online* **2010**, *9* (1), 39.

(9) Ebrahimi, farideh; Mikaili, M.; Estrada, E.; Nazeran, H. Automatic Sleep Stage Classification Based on EEG Signals by Using Neural Networks and Wavelet Packet Coefficients. **2008**.