Recommendations for Performance Assessment of Automatic Sleep Staging Algorithms

Syed Anas Imtiaz and Esther Rodriguez-Villegas

***Abstract*— A number of automatic sleep scoring algorithms have been published in the last few years. These can potentially help save time and reduce costs in sleep monitoring. However, the use of both R&K and AASM classification, different databases and varying performance metrics makes it extremely difficult to compare these algorithms. In this paper, we describe some readily available polysomnography databases and propose a set of recommendations and performance metrics to promote uniform testing and direct comparison of different algorithms. We use two different polysomnography databases with a simple sleep staging algorithm to demonstrate the usage of all recommendations and presentation of performance results. We also illustrate how seemingly similar results using two different databases can have contrasting accuracies in different sleep stages. Finally, we show how selection of different training and test subjects from the same database can alter the final performance results.**

I. Introduction

Human sleep is broadly classified into two distinct oscillatory phases based on the eye movements during sleep, known as Rapid Eye Movement (REM) and Non-Rapid Eye Movement (NREM). The NREM phase is further divided into different stages. According to Rechtschaffen and Kales (R&K) classification of sleep stages [1], published in 1968, NREM is further classified into Stages 1, 2, 3 and 4 known as S1, S2, S3 and S4 respectively. In 2007 the American Academy of Sleep Medicine (AASM) published a more simplified set of guidelines [2] based on which NREM is subdivided into N1, N2 and N3 stages. Both R&K and AASM classification include a Wake (W) stage while the former also includes an additional Movement Time (MT) stage.

In clinical practice, physiological signals from brain (EEG), eyes (EOG), muscle movements (EMG) and respiratory effort are recorded as part of a sleep study known as Polysomnography (PSG). These signals are then segmented into epochs of 30 seconds, analysed and assigned one of the sleep stages based on R&K or AASM rules by sleep experts. Visual analysis of these signals is a costly, tedious and error-prone task. It can take between 2-4 hours to analyse an overnight PSG recording [3] with the scoring agreement between different experts about 82% on average [4]. Therefore, automation of this analysis is desirable not only to save time and costs but also to improve uniformity between different scoring sessions and experts. The approach of automatically analysing and classifying sleep stages is generally referred to as *automatic sleep staging* or *automatic sleep scoring*.

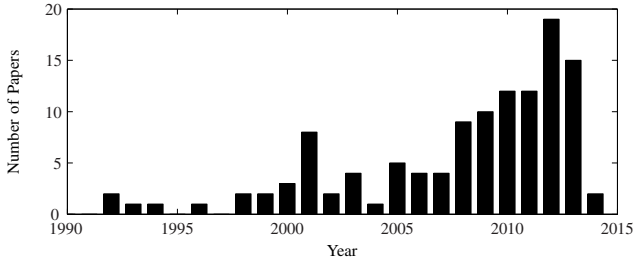

Fig. 1: *Number of automatic sleep staging algorithm (and related features) papers published in IEEEXplore over the last 25 years*

Automatic sleep staging, an actively growing research area, involves using a variety of feature extraction and machine learning methods for the analysis and classification of signals. The recent consumer focus on wearable devices for sleep tracking has also accelerated research in this area resulting in the use of other signals such as heart rate variability, body movements, etc. that are not conventionally used for the classification of sleep stages. Further, there is also a push towards using the least number of sensors for scoring sleep. For example, the use of single channel EEG or EOG and classification based on respiratory signals exclusively have received recent attention. The number of research papers published in this area has increased steadily over the past few years. Fig. 1 charts this rise for papers indexed in *IEEEXplore* that present features and/or methods for automatic classification of at least one stage of sleep.

The direct comparison between sleep staging methods that use a variety of signals for classification is hampered by the lack of standardization as they use different databases to report their performances. Further, although the AASM guidelines were publised in 2007, a number of research papers still use the older R&K classification (for valid reasons, discussed later). The selection of limited or partial test signals from the same database and the disparity in performance metrics used to report results also contribute in making the comparison difficult.

In this paper we look at different polysomnography databases available free of cost and propose a set of recommendations for their usage. We further show how adopting these usage guidelines and uniform performance metrics

S. A. Imtiaz and E. Rodriguez-Villegas are with the Circuits and Systems Group, Electrical and Electronic Engineering Department, Imperial College London, United Kingdom. Email: ({anas.imtiaz,e.rodriguez}@imperial.ac.uk).

The research leading to these results has received funding from the European Research Council under the European Community's 7th Framework Programme (FP7/2007-2013) / ERC grant agreement no. 239749.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

would allow fair comparison of the strengths and weaknesses of different sleep staging algorithms. Section II briefly describes the PSG databases and Section III how each of them should be used. Section IV discusses the performance metrics that should be reported with each algorithm. In Section V we demonstrate the usage of our proposed recommendations with an automatic sleep staging algorithm and illustrate how its performance can vary with the choice of database and the recordings within the same database.

II. Polysomnography Databases

In this section we list and briefly explain four PSG databases that can be used to develop and test sleep staging algorithms.

A. PhysioNet Sleep EDF Database [5]

The PhsyioNet Sleep EDF database [6] was made available online over 10 years ago and many algorithms have reported their detection performance using certain sections of this database. It consists of PSG recordings from 8 subjects, of which four were recorded overnight (cases starting with *st**) and the rest during a 24-hour period (starting with *sc**). All the recordings in this database include hypnograms scored using R&K classification.

B. PhysioNet Sleep EDF Expanded Database [7]

This is the superset of the previously described database which has recently been published in full. It consists of 61 subjects, some with overnight recordings and others with up to 24 hours of recordings, scored using R&K classification.

C. DREAMS Subjects Database [8]

This database from University of MONS - TCTS Laboratory and Université Libre de Bruxelles - CHU de Charleroi Sleep Laboratory consists of overnight PSG recordings of 20 subjects. It contains two hypnograms for every subject scored using R&K and AASM classification of sleep stages.

D. DREAMS Patients Database [8]

This dataset, also from the same source as above, has 27 PSG recordings of subjects with various sleep disorders including insomnia, PLMS and others. It also contains hypnograms that have been scored using both R&K and AASM classification of sleep stages.

III. Recommendations for using PSG Databases

An algorithm's performance can be reported on either of the databases listed above as they are available online free of cost. However, enough details about how the database has been used should be provided so that the results could be reproduced.

We propose a set of recommendations to be followed in conjunction with the publicly available databases (including those not listed above). These would simplify the comparison and reproduction of results leading to improvement in algorithms already published.

A. Classification: AASM and R&K

The AASM classification of sleep stages was published in 2007 and until then all sleep staging algorithms, naturally, reported their performance using the R&K classification. The adoption of R&K classification is so widespread that it is still in use in many clinics as well as some recent research publications. The major reason for publications still using the R&K instead of the AASM classification is that the PSG database they have is scored before 2007 using the former classification. We recommend using the AASM classification as it is the newer standard and also overcomes some of the limitations in the R&K classification [9], [10].

PhysioNet Sleep EDF databases include hypnograms with the R&K classification only. To report results using these databases according to the AASM classification, care must be taken not to ignore epochs from stages which are not part of the AASM classification. In most publications MT stage (movement) of R&K is ignored when using the AASM classification to present results. This can lead to incorrect or biased results since major body movements commonly transitions to wakefulness [10]. To roughly *convert* a R&K hypnogram to AASM, S3 and S4 stages should be marked as N3 while Wake and MT together should be marked as Wake (as shown in Table I). If using either of the two DREAMS databases, the accompanying AASM hypnogram should be used without any need of conversion from R&K.

TABLE I: *Conversion from R&K to AASM classification*

R&K	S1	S2	S3 S4	REM	Wake MT
AASM	N1	N2	N3	REM	Wake

B. Epoch size and signal duration

The standard epoch size for scoring of sleep stages according to both R&K and AASM classifications is 30 seconds. Some scorers and algorithms have also used different epoch sizes in the past. PhysioNet database includes hypnograms with standard 30 s epoch size while the two DREAMS databases listed here have been scored at a non-standard interval of 5 s. If the DREAMS databases are used then we recommend converting the hypnogram to 30 s epoch size using the following method. Starting from time zero, each 30 s epoch will have six scores in the original hypnogram for every block of 5 s. We recommend using the modal value of these six scores and assign it as sleep stage of the 30 s epoch. For epochs where there are ties between two stages, the previous value should be assigned. In cases where the last scored epoch has a duration of less than 30 s it should be removed. In other words, partial epochs towards the end of recording should not be analysed and the total signal duration should be a multiple of 30 (epoch size).

C. Selecting data from long term recordings

Some subjects in the PhysioNet database were recorded for a duration of up to 24 hours. To use these cases, most of the wake sections during the day is usually removed to

select only overnight sleep data. However, this selection of data is not consistent as some groups use data from the start of sleep removing all of the pre-sleep wake sections while others include greater periods of wake.

We recommend using the *lights off* time as the start time for these longer recordings. In cases where this is not available, 15 minutes of wake period prior to the first scored sleep epoch should be used. Similarly, to mark the end of a recording *lights on* time (if available) or 15 minutes of wake period after the last scored sleep epoch should be used. This selection of data is not required for DREAMS Subjects and Patients databases as they contain only the overnight recordings.

D. Training and test set

Most algorithms split the database into two sets: a training set for learning and a test set for performance validation. However, this split is often not clearly described and can have a big impact on the performance. If a 50/50 split on database is applied then it is important to know which ones were used for training and which ones for testing. It is difficult to reproduce the results of an algorithm without this knowledge, therefore the subjects used in each set should be clearly stated.

E. Unscored Epochs

All the databases listed have some epochs that were not assigned any of the known sleep stages. These epochs are considered unscored and we recommend removing them from the results when reporting the performance. Further, the number of unscored epochs removed should be stated.

IV. Performance Metrics

The overall performance of an algorithm is commonly represented by its *accuracy*, that is, the fraction of epochs correctly classified by the algorithm.

$$Accuracy = \frac{no.\ of\ true\ detections}{total\ no.\ of\ epochs} \tag{1}$$

However, not all stages of sleep occur for similar periods of time and their detection performances may vary considerably. Most papers present a further confusion matrix (or a contingency table) that provides details of the epochs correctly and incorrectly classified. Along with this, the following metrics should also be computed for each sleep stage to give a better understanding of in algorithm's performance.

$$Sensitivity = \frac{no.\ of\ true\ detections\ in\ stage\ X}{no.\ of\ reference\ epochs\ in\ stage\ X} \tag{2}$$

$$Selectivity = \frac{no.\ of\ true\ detections\ in\ stage\ X}{no.\ of\ all\ detections\ in\ stage\ X} \tag{3}$$

Sensitivity represents the fraction of correctly detected epochs in a sleep stage X, where X can be any of the five sleep stages. *Selectivity* refers to the proportion of true detections amongst the epochs classified by the algorithm.

V. Sleep Staging Algorithm

In this section an automatic sleep staging algorithm is presented and its performance is characterised using two PSG databases. The databases are used by following the recommendations in Section III. Three cases are used to illustrate how different databases and different subjects from the same database can affect the performance results.

The algorithm uses data from one EEG (frontal) and one EOG channel which are split into epochs of 30 s. Each epoch is further divided into 2 s blocks and transformed to frequency domain using Fast Fourier Transform (FFT). For each block of 2 s EEG, spectral power in every 2 Hz frequency bin from 0-30 Hz range is calculated i.e. 0-2 Hz, 2-4 Hz, 4-6 Hz and so on. For the corresponding EOG block, spectral power within 0-6 Hz is also calculated similarly for every 2 Hz frequency interval. Subsequently, the average of every feature is calculated within a 30 s epoch. Since each feature is calculated for a 2 s block, there are 15 such values within an epoch to calculate the average. This results in 18 features overall (15 EEG and 3 EOG) computed for an epoch and were classified with a Support Vector Machine (SVM). It was implemented using LIBSVM package [11] in MATLAB(ver. R2010a) with a third degree radial basis kernel function.

A. Case 1: Using DREAMS Subjects Database

In this case PSG data from DREAMS subjects database was used (Fp1-A2 and EOG1). It was partitioned such that subjects 1-10 were used in the training set while subjects 11-20 formed the test set. The database includes the hypnogram scored using the AASM classification with epoch size of 5 s. This is converted into a 30 s scoring interval by using the modal value of the sleep score in every 30 s epoch (as explained in Section III-B). Further, it was ensured that the total duration of recording in each subject contained a whole number of 30 s epochs discarding any remaining seconds at the end that formed an incomplete epoch. As a result, there was a total of 10178 epochs in the training set and 10087 epochs in the test set including 3 and 20 unscored epochs in the training and test sets respectively.

On the training set, accuracy of 82.7% was achieved while the test set resulted in an accuracy of 77%. The confusion matrix for the algorithm performance on the test dataset is shown in Table II. It shows that the sensitivity for stages W and N3 are more than 86% whereas only 17% of N1 epochs are correctly detected. This case illustrates how a high accuracy can easily mask the poor performance of the algorithm in one or more sleep stages.

TABLE II: *Case 1: Results using DREAMS Subjects Database*

		REFERENCE						
		W	**N1**	**N2**	**N3**	**R**	**Sen(%)**	**Sel(%)**
ALGORITHM	**W**	1599	226	112	15	60	87.0	79.5
	N1	52	142	75	0	201	17.2	30.2
	N2	132	326	3340	249	156	82.5	79.5
	N3	9	5	334	1627	1	86.0	82.3
	R	47	126	187	0	1046	71.5	74.4

B. Case 2: Using Sleep-EDF Database

In this case data from all 8 subjects in the PhysioNet Sleep EDF database was used (Fpz-Cz and EOG horizontal). The database consists of two kinds of recordings (described in Section II-A). The *st** recordings were used as is while data from *sc** recordings was selected using the recommendation in Section III-C. The eight subjects were partitioned to include two of each kind of recording in both the training and test dataset. The training set included sc4002, sc4102, st7022, st7121 and the test set included sc4012, sc4112, st7052, st7132. In total there were 8905 epochs (4650 in training and 4295 in test set) of which 1133 were unscored (589 in training and 544 in test set).

The accuracies achieved on the training and test sets were 79.5% and 73.4% respectively. This result is similar to Case 1 however, the confusion matrix shown in Table III shows that the sensitivity in each sleep stage is actually quite different. In particular, the results show improved sensitivities in REM, N1 and N2 stages, reduction in N3 sensitivity while it fails to classify any of the Wake epochs.

TABLE III: *Case 2: Results using PhysioNet Sleep EDF Database*

		REFERENCE						
		W	**N1**	**N2**	**N3**	**R**	**Sen(%)**	**Sel(%)**
ALGORITHM	**W**	0	0	0	0	0	0	0
	N1	117	98	30	7	17	30.6	36.4
	N2	36	55	1562	83	25	85.1	88.7
	N3	72	9	166	392	24	80.0	59.1
	R	111	158	78	8	703	91.4	66.5

C. Case 3: Using Sleep-EDF Database with different training and test set

In this case the same database as in Case 2 is used with the difference that all four *sc** recordings were part of the training set (3929 epochs with no unscored epochs) while the other four *st** recordings were part of the test set (5016 epochs including 1133 unscored epochs). An accuracy of 86.6% was achieved for the training set while the test set resulted in an overall accuracy of only 61.6%. The confusion matrix and individual sleep stage performances are shown in Table IV. In contrast to Case 2, the sensitivity in Wake stage is now close to 80% while in REM stage it has gone down from 91% to 19%. The overall accuracy is also less than that achieved in Case 2. This illustrates how using a different selection of training and test cases from the same database can result in a vastly different performance result.

TABLE IV: *Case 3: Results using PhysioNet Sleep EDF Database with a different training and test set*

		REFERENCE						
		W	**N1**	**N2**	**N3**	**R**	**Sen(%)**	**Sel(%)**
ALGORITHM	**W**	265	178	14	4	134	79.6	44.5
	N1	1	17	2	0	61	5.4	21.0
	N2	43	102	1293	117	427	81.6	65.2
	N3	24	16	276	649	82	84.3	62.0
	R	0	5	0	0	164	18.9	97.0

VI. Discussion & Conclusion

In this paper we have proposed a set of guidelines and recommendations for using the common PSG databases freely avaialble on the internet. We have listed four databases but the recommendations apply equally to other databases that will become available in future. In particular, we proposed using the AASM classification in all future work and explained how to roughly convert the hypnograms scored with R&K classification. We also described a method to convert non-standard epoch size hypnograms to the standard 30 s scoring interval.

We used a sleep staging algorithm based on spectral features and SVM classifier to demonstrate how different databases can alter its performance. We showed that even if the classification accuracy using different databases is similar, the results of detection in each sleep stage can be very different. We also showed how the results can easily change by using a different set of taining and test subjects from the same database.

The algorithm in this paper is not intended to be a high performing sleep staging method. It is used only to show the usage of the proposed recommendations and the effect of databases on results. We hope that the recommendations in this paper will allow researchers to fairly compare different methods subsequently leading to improvements in already existing automatic sleep staging algorithms.

References

[1] A. Rechtschaffen and A. Kales, Eds., *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Washington D.C.: Public Health Service, U.S. Government Printing Office, 1968.

[2] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, Eds., *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. Westchester, IL: American Academy of Sleep Medicine, 2007.

[3] M. Ronzhina, O. Janousek, J. Kolarova, M. Novakova, P. Honzik, and I. Provaznik, "Sleep scoring using artificial neural networks," *Sleep Med. Rev.*, vol. 16, no. 3, pp. 251–63, 2012.

[4] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, M. Boeck, H. Dorn, G. Gruber, E. Heller, E. Loretz, D. Moser, S. Parapatics, B. Saletu, A. Schmidt, and G. Dorffner, "Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard," *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, 2009.

[5] PhysioNet. (2013) Sleep-edf database. [Online]. Available: http://www.physionet.org/physiobank/database/sleep-edf/.

[6] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[7] PhysioNet. (2014) Sleep-edf database [expanded]. [Online]. Available: http://www.physionet.org/physiobank/database/sleep-edfx/.

[8] University of MONS - TCTS Laboratory. (2014) The DREAMS Databases. [Online]. Available: http://www.tcts.fpms.ac.be/~devuyst/#Databases/.

[9] S.-L. Himanen and J. Hasan, "Limitations of rechtschaffen and kales," *Sleep Med. Rev.*, vol. 4, no. 2, pp. 149–167, 2000.

[10] M. Silber, S. Ancoli-Israel, M. Bonnet, S. Chokroverty, M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. Keenan, M. Kryger, T. Penzel, M. Pressman, and C. Iber, "The visual scoring of sleep in adults," *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 121–31, 2007.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.