Towards a user-friendly sleep staging system for polysomnography part I: Automatic classification based on medical knowledge

Jade Vanbuis, Mathieu Feuilloy, Guillaume Baffet, Nicole Meslier, Frédéric Gagnadoux, Jean-Marc Girault

Version of Record: https://www.sciencedirect.com/science/article/pii/S2352914820306043
Manuscript_1d8e990e1d5a45e341e9aa299c9908ca

Towards a user-friendly sleep staging system for polysomnography Part II: patient-dependent features extraction using the SATUD system

Jade Vanbuis[a,b,*], Mathieu Feuilloy[a,b], Lucile Riaboff[a,b], Guillaume Baffet, Alain Le Duff[a,b,1], Nicole Meslier[c,d], Frédéric Gagnadoux[c,d] and Jean-Marc Girault[a,b]

[a]*ESEO, Angers, France*
[b]*LAUM, UMR CNRS 6613, Le Mans, France*
[c]*Angers sleep laboratory, University Hospital, Angers, France*
[d]*INSERM UMR 1063, University of Angers, Angers, France*

ARTICLE INFO

Keywords:
The SATUD system
Data-dependent features
Unsupervised thresholding
Expert knowledge
Interpretable classifier
Sleep scoring

ABSTRACT

Manual sleep stages scoring is time-consuming, complex and requires specific medical knowledge. Automatic sleep stages classification, usually based on supervised methods of machine learning, is the object of researchers interest. However, it remains challenging because of the high variability among patients which is not considered with such algorithms. This paper presents a method to extract patient-dependent qualitative features from electrophysiological signals, preceding a supervised machine learning classifier. Instead of using fixed thresholds, the developed method called "Self-Adaptative Thresholding Using Descriptors" (SATUD), proposes an unsupervised self-adjusting thresholding. Thresholds are automatically adjusted to maximize both the similarity within a same sleep stage and the dissimilarity between different ones. This method is evaluated using manual sleep stages scoring from 60 patients with various pathologies to ensure high variability. The SATUD shows a better adaptation to the patient specificities, compared with two other thresholding methods implemented in this study. Indeed, the number of 30-seconds recording segments respecting all their sleep stage properties increased by more than 80 % with the use of the SATUD, compared to other thresholding techniques. It was also proved robust to noise and sweat artifacts. The SATUD thereby provides patient-dependent qualitative features which can be used for automatic sleep stages scoring using a machine learning method. This last point was presented in the companion paper.

1. Introduction

Sleep-related disorders nearly affect the third of the population and are increasingly recognized as real public health problems [1]. The need of sleep diagnosis has then increased during the last few years [2, 3].
Gold-standard procedure for sleep diagnosis consists of electrophysiological and respiratory signals recording with the use of a polysomnograph. With all those signals, medical staff manually scores both sleep events and sleep stages. Sleep events, as apneas and hypopneas, are detected through respiratory signals. On the other hand, sleep stages are scored based on the electrophysiological signals. Among them, electroencephalograms (EEG), electrooculograms (EOG) and electromyograms (EMG) are used for the measurement of cerebral, ocular and muscular activities, respectively. Sleep scoring consists on the classification of wakefulness, stage N1 and stage N2 (both light sleep), stage N3 (deep sleep) and REM sleep (Rapid Eye Movement, also called R stage or paradoxical sleep: stage with active brain but reduced muscle tone) in 30-second sections, also called epochs. Besides being a time-consuming task, sleep scoring requires specific medical knowledge. A manual of recommendations published by the American Academy of Sleep Medicine (AASM) in 2007 [4] describes the temporal and spectral contents of each sleep stage, as well as sleep patterns that can be recognized by the scorer and the possible transitions from one sleep stage to another.

There was a growing number of automatic sleep scoring algorithms developed those last few years, spread out into three categories: deep learning [5, 6], machine learning [7–13] and hybrid approaches [14–16]. Despite the increasing number of models based on artificial intelligence (AI), only a few are routinely employed by sleep specialists. Several reasons for that have been identified and detailed in the companion paper [REF] and in [17]. One of them is the lack of transparency of the developed approaches. Indeed, deep learning approaches, which often reach the best scores, are opaque and their lack of transparency raises skepticism among physicians. With opaque approaches, medical practitioners must accept to loose their control upon the task that is realized by the algorithm. It could prevent them to properly react when dealing with an unusual pathology (that was not necessarily represented when training the model). For this reason, some researchers attempt to improve the interpretability of their models [18], and provide some concrete elements for the practitioner to relate to. Hybrid approaches [14–16] were then designed to overcome several identified limitations, including the lack of transparency of the implemented method.

The present article is the second part of a two-part paper. In the companion paper [REF], a novel hybrid approach replicating the steps of a manual scoring was developed and

*Corresponding author
jade.vanbuis@eseo.fr
ORCID(s): 0000-0001-6437-1597 (J. Vanbuis)
[1]Present address: Olympus NDT Canada, Québec, Canada

presented. This hybrid system is composed of several functions, including one dedicated to features extraction, necessary for classification. Called SATUD, this features extractor will be fully detailed in the following section, since it is the core of the present paper. Extracted features describe each epoch. They are qualitative and represent concrete elements mentioned in the AASM guidelines. However, the extraction of such features can be problematic for classification if not correctly carried out. Indeed, there is a strong variability within subjects. This variability is one of the major difficulties encountered by sleep specialists while manually scoring sleep. To score sleep stages appropriately, medical practitioners generally need to get familiar with the recording specificities. To do so, they start with a quick visualization of the entire recording, before doing the scoring. However, amongst the many automatic scoring algorithms developed the last few years [5–16], only a few were adapted to the variability between subjects [14–16]. Furthermore, adaptation to each recording specificities implies having a sufficient number of recordings from patients with various pathologies, which is rarely the case.

This study presents the Self-Adaptative Thresholding Using Descriptors (SATUD) method. Developed for the extraction of subject-dependent features, the SATUD reduces the impact of subjects variability. Features values are adjusted to each patient, as sleep specialists naturally do when scoring sleep. The SATUD principle is explained in the following section, and its ability to correctly adjust thresholds is then evaluated and compared with other thresholding methods.

2. Material and method

In this section, the data used for the implementation of the SATUD is first presented (2.1). The chosen approach is then detailed in Section 2.2.

2.1. Data

Patients with various sleep pathologies underwent one-night polysomnography (PSG) in Angers sleep laboratory (University Hospital, INSERM UMR1063, Angers, FRANCE). PSG is a sleep diagnostic device acquiring electrophysiological signals, respiratory signals, body movements and position. Recordings were part of the "Institut de Recherche en Santé Respiratoire des Pays de la Loire" [IRSR] sleep cohort. Approval was obtained from the University of Angers ethics committee and from the "Comité Consultative sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé" [CCTIRS] (07.207bis). All patients included in the IRSR sleep cohort have given their written informed consent. Recordings were anonymous. Sleep stages and events were recorded and scored following the AASM recommendations [4] using CID102L8D polysomnographs. Besides those standard polysomnographic signals, tracheal sounds were recorded using a PneaVoX® device for enhanced ventilatory events recognition [19]. A total of 60 anonymous patients recordings was scored by sleep specialists (three sleep specialists were involved in the study, but each recording was scored by a single scorer). In this study, the automatic algorithm only employed EEG, EOG and EMG signals used by physicians to score sleep stages. The sequence of wake, N1, N2, N3 and REM sleep in epochs of 30 seconds is called hypnogram and constitutes our reference.

2.2. Method

The hypothesis made in this paper is that concrete features can help overcome the model lack of transparency. Qualitative (ordinal) features allow close translation of medical knowledge (AASM recommendations). Their estimation requires the discretization of quantitative features. For example, the quantitative feature *amplitudeEEG*[1] is discretized into qualitative features as *amplitudeEEGHigh* and *amplitudeEEGLow*, employing a certain number of thresholds needing to be estimated. For this example, the problem is to know the threshold for which we can consider having a low or high EEG amplitude. The SATUD algorithm aims to adjust those thresholds automatically, for each recording. The integration of the SATUD in the sleep staging system implemented in this study is schematized in Figure 1. The architecture is composed of several main functions: **F1**, **F2**, **F3** and **F3.A**. A detailed example can be found in Appendix A.

2.2.1. SYSTEM

The proposed system aims to provide a set of patient-specific qualitative features. As described in Figure 1, it was composed of three main functions (*F1*, *F2*, *F3*). For applications concerning sleep staging, the system inputs were:

- **a priori knowledge** giving information about sleep scoring through AASM manual. For example, this manual indicates that N3 sleep stage can be recognized using low-frequency (0.5-2 Hz) high-amplitude ($> 75\,\mu V$) EEG frontal waves. A high proportion of those waves called 'Slow wave activity', indicates N3;
- **electrophysiological signals** obtained from the recorded polysomnographies, and more precisely three electroencephalograms (EEG), two electrooculograms (EOG) and the chin electromyogram (EMG).

2.2.2. F1 - Sleep stages description

F1 aims to build one list of properties for each sleep stage. The a priori knowledge from AASM manual (as detailed in SYSTEM) was the input of the first function F1. In order to build the list, we carefully studied the AASM manual for all sleep stages, and translated subsequently into lists of properties which represent time and frequency information used to differentiate sleep stages. The upper part of the Figure 2 shows an example of two properties that were used to describe sleep stage N3. The lists of properties associated with each sleep stage are essential for the SATUD proper functioning, as they replace labelled data.

[1] Note the signals were recorded with a bit-depth of at least 8 bits.

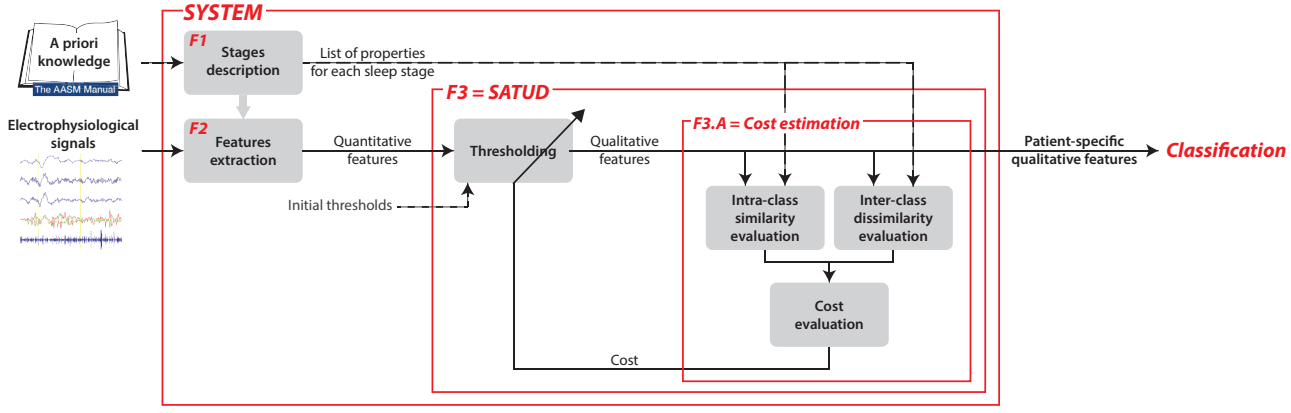
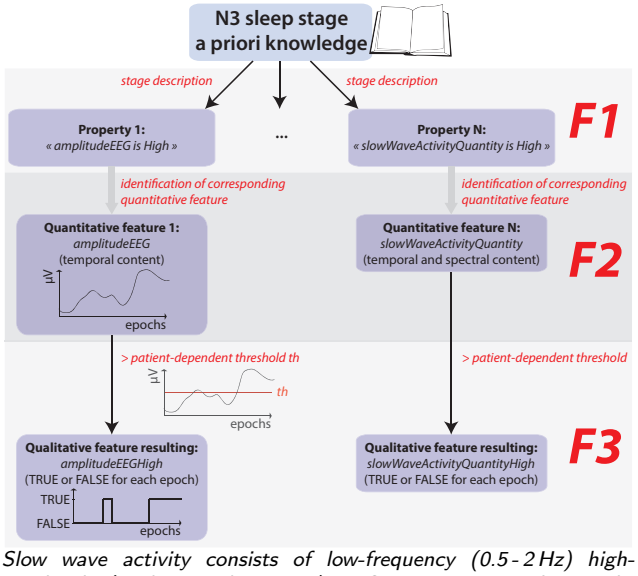


Figure 1: Functional architecture composed of three main functions (F1, F2 and F3). Medical knowledge from the AASM manual and electrophysiological signals are used as inputs. The output is a set of qualitative features with less vulnerability to patients specificities.



Slow wave activity consists of low-frequency (0.5 - 2 Hz) high-amplitude (peak-to-peak > 75 µV) EEG waves measured over the frontal region of the brain.

Figure 2: Simplified example illustrating the links between sleep stage N3 properties (defined in F1), quantitative features (defined in F2) and qualitative features (defined in F3).

2.2.3. F2 - Features extraction

F2 aims to extract the quantitative features corresponding to the properties identified in function F1. Using the electrophysiological channels split into epochs, 13 quantitative features were extracted. Those quantitative features are listed in the 1st column of Table 1. Those quantitative features can reflect temporal or spectral content, but also a combination of both temporal and spectral content. They were identified in the AASM guidelines as being required for sleep scoring. In the middle part of Figure 2 are presented two quantitative features identified thanks to the previous properties.

2.2.4. F3 - SATUD

The SATUD aims to deduce patient-specific information from quantitative features using specific knowledge. It decreases subjects variability with the estimation of patient-specific thresholds. Thresholds were applied on the 13 quantitative features obtained in F2 to generate 41 qualitative features. The qualitative features and the associated number of thresholds were chosen in agreement with the properties described in F1. To do so, for each quantitative feature from F2, and depending on the properties from F1 (which translate the AASM guidelines), the appropriate qualitative features were extracted[2]. Those qualitative features are listed in the 3rd column of Table 1.

Different ways of defining initial thresholds were tested. In this application, initial thresholds were chosen using a statistical approach (percentiles). Thresholds values were then adjusted in order to minimize the cost function described in F3.A. To do so, several meta-heuristics were tested: global search algorithms such as simulated annealing [20] and genetic algorithms [21, 22] and also local algorithms such as gradient descent methods [23, 24]. For our application, the use of local search algorithms alone was ineffective due to the number of thresholds to adjust. We thus chose to use a global search algorithm to initialize the search zone, but still combined it with a final local search algorithm for better precision. As indicated in the pseudo-code reported in Appendix B, the method chosen for this study was the combination of simulated annealing followed by a gradient descent.

As there is a high variability between subjects, optimal thresholds values could be very different from one patient to another. Thresholds adaptation to each patient is explained in the following section.

[2]Only qualitative features which correspond to sleep stages properties described in the AASM guidelines were computed, even if the number of thresholds allowed the estimation of more features. For example, 2 thresholds (Low → Mid and Mid → High) can generate 6 qualitative features: low, low or mid, mid, mid or high, high and low or mid or high. However, the study of the AASM guidelines rarely indicates there is a need for all those features. Only few of them can be sufficient to translate the guidelines.

Table 1
List of the 41 qualitative features extracted for sleep scoring, depending on the 13 quantitative features extracted in F2 (ranging from EEG amplitude to Substracted EOG instability) and the number of associated thresholds.

Quantitative features	Th[a]	Qualitative features used	N[b]
EEG amplitude	2	Low | Low or Mid | Mid or High | High	4
EEG instability	1	No | Yes	2
Slow wave activity quantity	2	Low | Low or Mid | High	3
Alpha waves quantity	2	Low | Low or Mid | Mid | Mid or High | High	5
Beta waves quantity	2	Low | Mid | Mid or High	3
Delta waves quantity	2	Low | Mid or High	2
Theta waves quantity	2	Low | Low or Mid | Mid or High	3
Chin level	2	Low | Low or Mid | Mid or High | High	4
Chin instability	2	Low | Low or Mid | Mid or High | High	4
Summed EOG level	2	Low | Low or Mid | Mid or High	3
Summed EOG instability	2	Low | Low or Mid | Mid or High	3
Substracted EOG level	2	Low | Mid or High	2
Substracted EOG instability	2	Low | Mid or High | High	3
		Total	41

[a] Number of Thresholds employed.
[b] Number of qualitative features used for each quantitative feature.

2.2.5. F3.A - Cost estimation

The final cost was defined as a weighted sum of the costs associated to each class (sleep stage in this study):

$$finalCost = \sum_{c=1}^{L} w_c \times cost_c \quad (1)$$

where L is the number of classes, and w_c and $cost_c$ are the weight and cost associated to the c^{th} class respectively. Weights are optional. In our application, they have been chosen to promote sleep stages that are difficult to identify or demote the ones that occur rarely (N1 sleep stage only represents approximately 5 % of the night).
$cost_c$ was defined as:

$$cost_c = \frac{1}{conc(R_{P1}, R_{P2}, ..., R_{PN}) \times std(antiScore_c)} \quad (2)$$

with:

$$antiScore_c = \sum_{i=1}^{N} v_i \times \overline{R_{Pi}} \quad (3)$$

where N is the number of properties of class c, R_{Pi} is a binary variable representing the respect of the i^{th} property of class c (as explained hereafter), and v_i is the weight associated to the i^{th} property of class c. This time, weights have been chosen to better translate the AASM guidelines. Indeed, some properties are clearly indicated as being more important for sleep staging.
The cost function is minimized for each recording individually. Defined as above, it is equivalent to i) maximize the similarity between the epochs of a same sleep stage of the same patient, hereinafter referred to as intra-class similarity and ii) maximize the differences between the epochs belonging to different sleep stages of the same patient, hereinafter referred to as inter-class dissimilarity.

i) **maximize intra-class similarity**: intra-class similarity is optimized by maximizing $conc(R_{P1}, R_{P2}, ..., R_{PN})$. This term represents the concordance between the respect of the properties used to describe class c. The respect of a property is assessed using the corresponding qualitative feature. For our application, the concordance between the respect of the properties was estimated using the Fleiss' Kappa [25]. This is a statistical measure used to assess the reliability of agreement between several categorical vectors. Indeed, we can consider that when a patient is in a specific sleep stage, most of the properties related to this sleep stage are respected. When the sleep stage changes, a certain number of those properties will no longer be respected. If we focus on the properties related to a particular sleep stage, the transitions to and from this sleep stage will be highlighted by a simultaneous change of the respect of its properties. Those simultaneous changes can be evaluated by assessing its properties respects interdependency. Therefore, a high interdependency will indicate a general concordance between thresholds. Thresholds were adjusted unsupervisingly, until they agreed and confirmed each other, leading to a higher concordance measure (Fleiss' Kappa). This is illustrated in Figure 3.

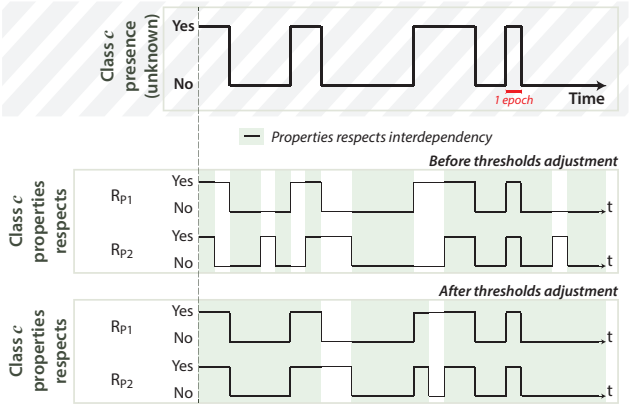


Figure 3: Simplified example of the properties respects behaviour before and after thresholds adjustment. Class c is described by only two properties ($P1$ and $P2$). The interdependency between properties respect R_{P1} and R_{P2} increases when thresholds are being adjusted.

In this example, that has a length of 26 epochs, class c presence is unknown for thresholds adjustment (unsupervised functioning). The better the thresholds are, the more properties are respected (R_{P1} and R_{P2}) when and only when the patient is in class c. Before thresholds adjustment, 18 epochs

among the 26 were in concordance with each other (highlighted areas). This number increased from 18 to 23 after thresholds adjustment, leading to a better Fleiss' Kappa.

ii) **maximize inter-class dissimilarity**: inter-class dissimilarity is optimized by maximizing $std(antiScore_c)$. The anti-score function is defined as a weighted sum of the no-respect of a class properties (Equation 3). Weights are optional. For our application, they were empirically chosen to translate the degree of importance according to the AASM manual.
$std(antiScore_c)$ represents the fluctuation (using standard deviation value) of class c anti-score function. The anti-score function of a particular sleep stage varies from 0, when all the properties associated to this sleep stage are respected, to 1 when none of the properties are respected. When thresholds are being adjusted, the anti-score values are getting closer to its limits (0 when in the sleep stage and 1 when in another sleep stage), with fewer values in between. Consequently, the anti-score standard deviation on the entire recording increases. This is illustrated in Figure 4, which pursues the example presented in Figure 3.

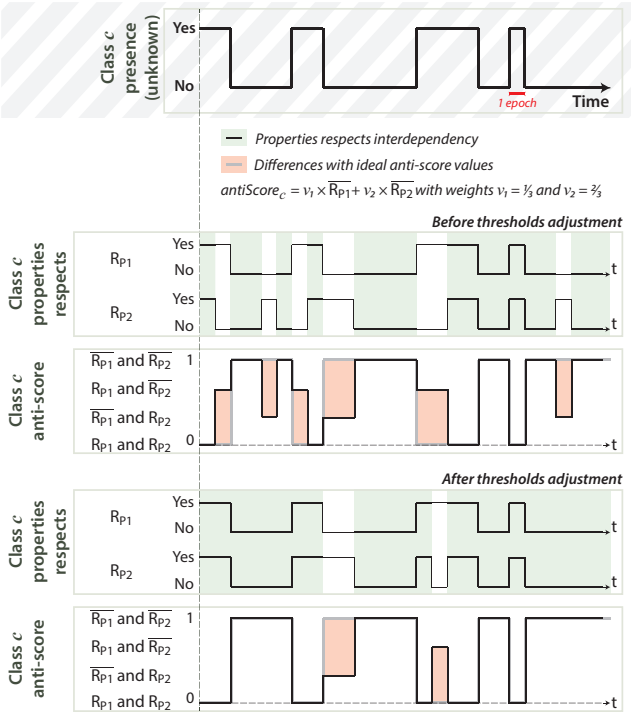


Figure 4: Simplified example of the anti-score behaviour before and after thresholds adjustment. Class c is described by only two properties ($P1$ and $P2$). The anti-score is a weighted sum of the respect of each property (R_{P1} and R_{P2}). Its standard deviation value increases when thresholds are being adjusted.

In this example, the anti-score standard deviation increased from 0.41 to 0.47 with thresholds adjustment.

To summarize, the cost depends on the intra-class similarity and inter-class dissimilarity, both associated with the respect of each class properties. The respect of each class properties are evaluated from the qualitative features values. Those values are directly related to the thresholds. For each recording individually, and without using the manual scoring as the reference, the thresholds are thus adjusted by minimizing the cost. At the end of the process, the patient-specific qualitative features are extracted (using the final thresholds) and can be used for classification.

3. SATUD evaluation methods

This section is dedicated to the assessment of the SATUD performance, which was evaluated from the totality of the 60 recordings (since the method is unsupervised). The SATUD was employed to extract qualitative features that are understandable and represent concrete information from the AASM guidelines. It was thus compared with other thresholding methods employed the same way. The complete system reported in Figure 1 was replaced by GST (General Statistical Thresholding) and IST (Individual Statistical Thresholding) as described thereafter.

- *General Statistical Thresholding (GST)*: with this method, thresholds were adjusted using statistical information of the entire database, as percentiles (thresholds were not patient-dependent).
- *Individual Statistical Thresholding (IST)*: with this method, thresholds were adjusted for each patient depending on statistical information (thresholds were patient-dependent).

Impact on the obtained qualitative features was first estimated (3.1). The SATUD behaviour in presence of noise or artifacts was then tested (3.2) and, eventually, the impact on classification was evaluated (3.3).

3.1. SATUD impact on qualitative features

Using the lists describing each sleep stage, we quantified the agreement between each epoch qualitative features and the properties of the associated sleep stage. For example if there were 10 properties in the list describing sleep stage Wake, then a Wake epoch with qualitative features respecting only 5 of them had a global respect value R of 50 % with its sleep stage properties.
Outcomes will be presented in Section 4.1.

3.2. Robustness test

To evaluate the robustness of the SATUD, we assessed the quantity of epochs highly respecting the properties associated to their sleep stage under several situations:

- *noise amplification*: white Gaussian noise was added to all raw electrophysiological signals. Raw signals presented a native Signal-to-Noise Ratio (SNR) of approximately 46 dB. Several SNR levels were tested, ranging from 30 dB to 0 dB, in 10 dB intervals.
- *artifacts addition*: several artifact types are usually present on electrophysiological signals. All artifacts were kept in our recordings to evaluate our algorithm under real-life

conditions. To further test the robustness of the method, we added artificial artifacts to the signals. In this study, we chose sweat artifacts, considered as the most disruptive ones. Indeed, during such artifacts, both temporal and spectral information are compromised or even lost. Artificial sweat artifacts were created using random slopes and durations within intervals defined after visualization of several natural sweat artifacts. A duration limit was used to prevent an entire 30-seconds segment to be exclusively composed of sweat artifacts. Natural sweat artifacts occur more often in some sleep stages compared to others and their quantity is generally limited [26, 27]. For this reason, we tested the results with the addition of artificial artifacts on 0 % (raw signals) to 100 % of the epochs composing recordings. Figure 5 presents natural (5a) and artificial (5b) sweat artifacts examples.

Outcomes will be presented in Section 4.2.

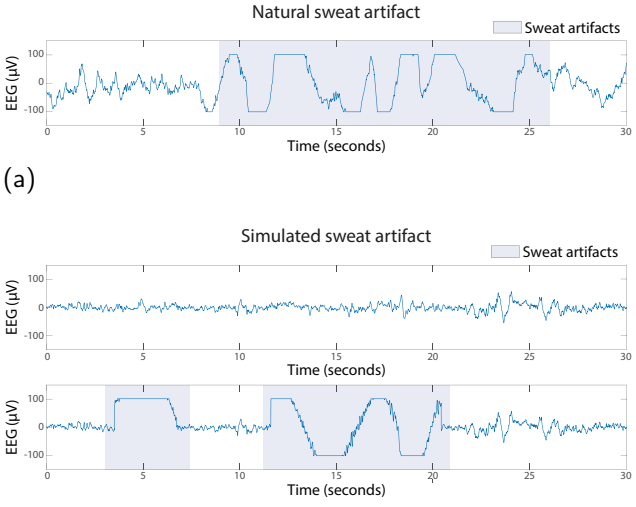


Figure 5: Examples of EEG signals during 30-seconds epochs: (a) an epoch with natural sweat artifacts and (b) an epoch with added simulated sweat artifacts.

3.3. SATUD impact on classification

Patient-dependent qualitative features described in the previous section are used for sleep stages classification. To properly estimate the impact of the SATUD, a first simple classification model was tested in this paper and compared with manual sleep scoring. The implementation of an advanced classifier was presented in the companion paper [REF]. However, it is necessary to assess the SATUD efficiency by avoiding any bias that could be linked to the use of powerful classification models. The classifier described in the current paper was built to be relatively transparent and easy to understand. It did not required training and testing steps. Indeed, using the adjusted thresholds, percentages of agreement with each sleep stage list of properties were evaluated for each epoch. The sleep stage with the higher agreement rate between the properties and its patient-dependent qualitative features was the chosen one. In the current paper, results were simply expressed in term of accuracy rate with the manual scoring, named *Acc*:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

where *TP*, *FN*, *FP* and *TN* represent the number of true positives, false negatives, false positives and true negatives, respectively.
Outcomes will be presented in Section 4.3.

4. Results

The SATUD performance was estimated through the qualitative features generated (see Section 4.1) and compared with the two other thresholding methods: GST and IST. Robustness to noise or artifacts tests results were then reported in Section 4.2. The impact on classification were evaluated in Section 4.3.

4.1. SATUD impact on qualitative features

Figure 6 shows the number of epochs that respect at least 0 % to 100 % of their sleep stage properties, according to the thresholding method used. Of course, all epochs respected at least 0 % of their sleep stage properties whatever the method used. With the SATUD, the number of epochs respecting at least 60 % to 100 % of the properties associated with its sleep stage were higher than with GST and IST. It means that qualitative features obtained with the SATUD better respect the properties expected for their sleep stage. There were indeed relative increases of 81 % ($\frac{8135-4506}{4506}$) and 89 % ($\frac{8135-4306}{4306}$) of the number of epochs that perfectly respect the properties associated with their sleep stage when using the SATUD compared to GST and IST, respectively.

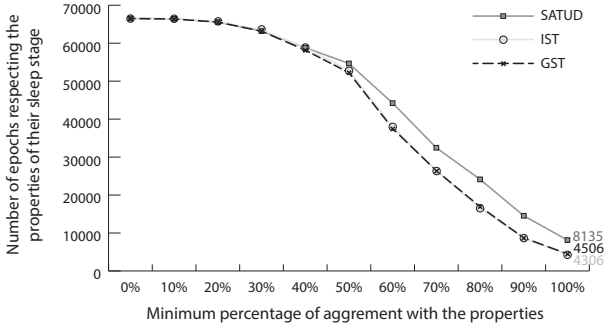


Figure 6: Evolution of the number of epochs respecting at least 0 % (the total number of epochs is obtained) to 100 % of the properties associated to their sleep stage, depending on the method used.

When focusing on epochs that highly ($60\,\% < R \leq 80\,\%$) or almost perfectly ($R > 80\,\%$) respect the properties associated with their sleep stage, we compared the different methods depending on each sleep stage. Figure 7 shows the impact of the SATUD on those epochs compared to GST and IST. When comparing the SATUD with GST and IST, we registered improvements for sleep stages W, N2, N3 and REM sleep but not for N1 sleep stage.

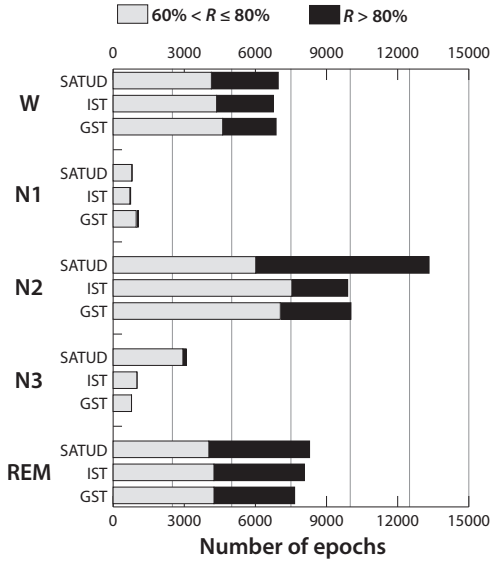


Figure 7: For each sleep stage, quantity of epochs highly ($60\,\% < R \leq 80\,\%$) or almost perfectly ($R > 80\,\%$) respecting the properties associated with their sleep stage, according to the method used.

4.2. Robustness test

Robustness tests described in Section 3.2 were conducted on GST, IST and the SATUD methods.

We first studied noise and sweat artifacts impact on qualitative features. The quantity of epochs that almost perfectly respect the properties associated with their sleep stage ($R > 80\,\%$) were evaluated. Results were compared for different levels of noise and artifacts for GST, IST and the SATUD (Figure 8). The SATUD behaviour to noise was globally similar than GST and IST, but with higher scores. Almost no N1 and N3 epochs respected almost perfectly the N1 and N3 properties. The quantity of epochs that almost perfectly respect the properties associated with their sleep stage dropped from a SNR level of 10 dB for all sleep stages except for Wake, where it remained quite constant with GST and IST, and slowly decreased for the SATUD.

Considering the addition of artificial sweat artifacts, the use of the SATUD once again positively impacts the respect of sleep stages with their properties. However this time, it behaved differently than GST and IST. For those last ones, artifacts had a very small impact, with a slight increase for Wake and N2 epochs and a progressive decrease for REM sleep. Using the SATUD, N2 and N3 epochs underwent an improvement whereas adding sweat artifacts until they are applied on 30 % - 40 % of the recording epochs. Afterwards, it decreased progressively.

The SATUD obtained better results than GST and IST in terms of quantity of epochs respecting almost perfectly properties associated with their sleep stage. While remaining superior, its behaviour in noise and artifact situations was globally similar to GST and IST thresholding methods.

Table 2
Accuracy rate with the manual scoring according to the different thresholding methods while testing robustnes.

	SATUD	IST	GST
	Robustness to noise test results		
Raw signals	55 %	46 %	45 %
30 dB SNR	55 %	46 %	45 %
20 dB SNR	54 %	46 %	45 %
10 dB SNR	44 %	34 %	34 %
00 dB SNR	43 %	32 %	33 %
	Robustness to sweat artifacts test results		
Raw signals	55 %	46 %	45 %
20 %	55 %	45 %	41 %
40 %	55 %	45 %	37 %
60 %	54 %	44 %	36 %
80 %	54 %	44 %	35 %
100 %	54 %	44 %	33 %

4.3. SATUD impact on classification

Classification global accuracy rates were estimated for all three methods. For raw signals, the SATUD obtained the best agreement with the reference with $Acc = 55\,\%$, versus $Acc = 45\,\%$ and $Acc = 46\,\%$ for GST and IST respectively (Table 2). Confusion matrices are shown in Figure 9. The SATUD confusion matrix (Figure 9a) registered significant improvements for sleep stages N2 and N3, if compared to IST and GST (Figure 9b and Figure 9c). Results for different levels of noise and artifacts are reported in Table 2. The SATUD obtained the best results for all levels of signals degradation. For IST and the SATUD, sweat artifacts did not have an important impact on the classification scores.

5. Discussion

The method developed in our study, called SATUD (Self-Adaptative Thresholding Using Descriptors algorithm), aims to facilitate data-dependent features obtainment to be used for classification. It automatically and unsupervisingly adjust thresholds by minimizing a cost function. The cost function was determined based on mathematical reasoning and translation of the expert knowledge. For our application, the SATUD was employed to reduce the impact of subjects variability before automatic sleep staging. It reproduces the first screening done by experts when manually scoring sleep.

This pre-processing step proved its value since the SATUD showed an improvement of epoch agreement with the properties associated with their sleep stage, compared to two other thresholding methods. Compared to GST (General Statistical Thresholding) and IST (Individual Statistical Thresholding), the quantity of epochs almost perfectly respecting their sleep stage properties increased from less than 3000 to more than 7300 epochs with the use of the SATUD for sleep stage N2. This stage had the best classification improvement. We can consider that the SATUD highlighted this sleep stage, probably because it represents almost 50 % of a night. On

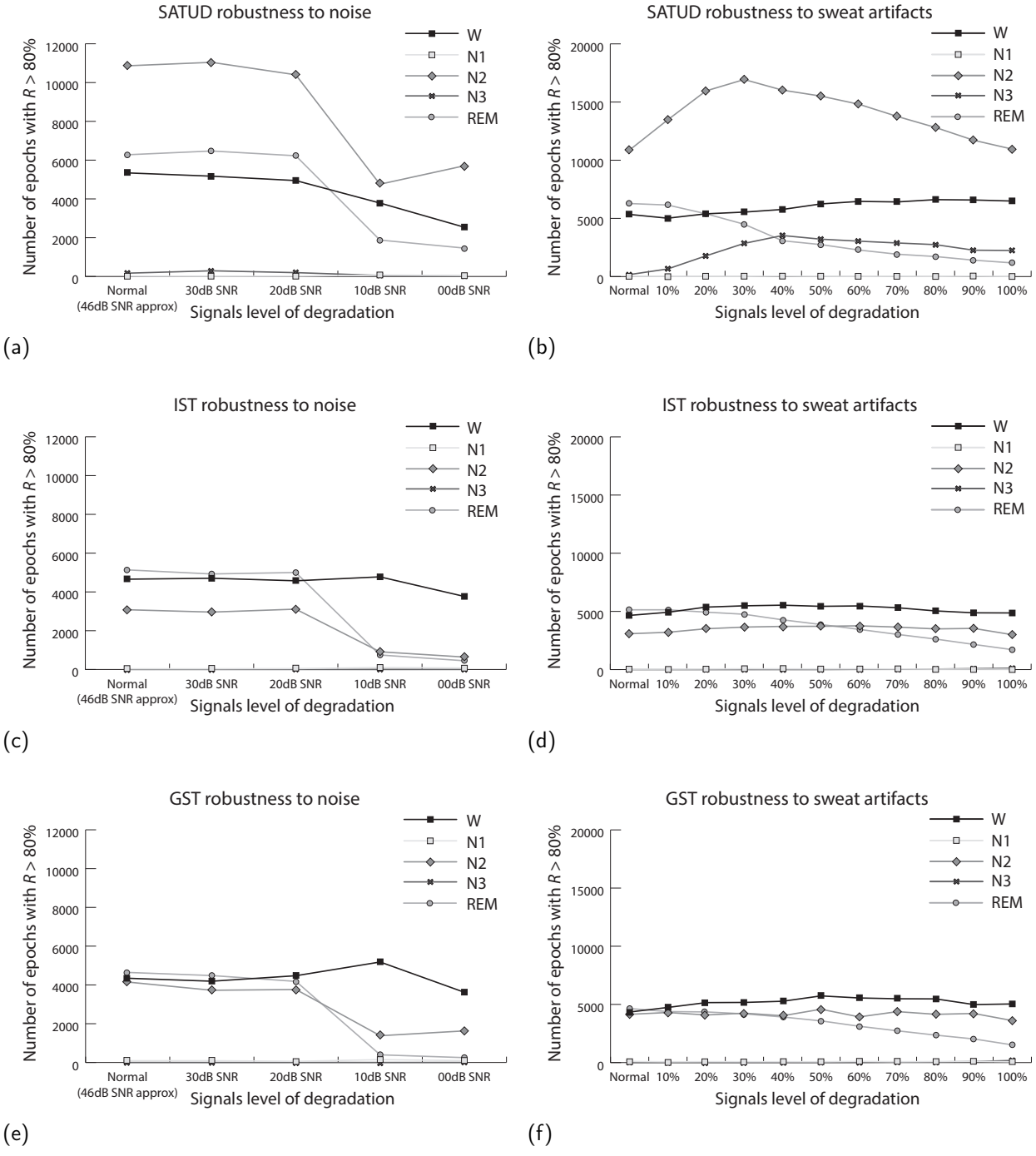


Figure 8: Evolution of the number of epochs with an almost perfect agreement ($R > 80\,\%$) with the properties associated to their sleep stage for: (a) the SATUD thresholding tested with increasing noise, (b) the SATUD thresholding tested with increasing number of artifacts, (c) IST thresholding tested with increasing noise, (d) IST thresholding tested with increasing number of artifacts, (e) GST thresholding tested with increasing noise and (f) GST thresholding tested with increasing number of artifacts.

the opposite, and for all methods, N1 was the sleep stage with the worst results. This could be explained by the fact that N1 is a transitional stage that occurs during only 5 % of the night. This limitation does not have a significant impact on the classification performance because N1 appears to be rare, and N1 errors are common in literature, even within manual scorers [28]. The primary classification tool tested also showed higher results for the SATUD than GST and IST. Improvements were more important for the N3 sleep stage. The proposed method also showed a good performance regarding noise and artifacts. The SATUD was not sensitive to additional white noise until the SNR level of 10 dB was reached. Note that this is an important noise level that would be difficult to reach using sensors as used in polysomnog-

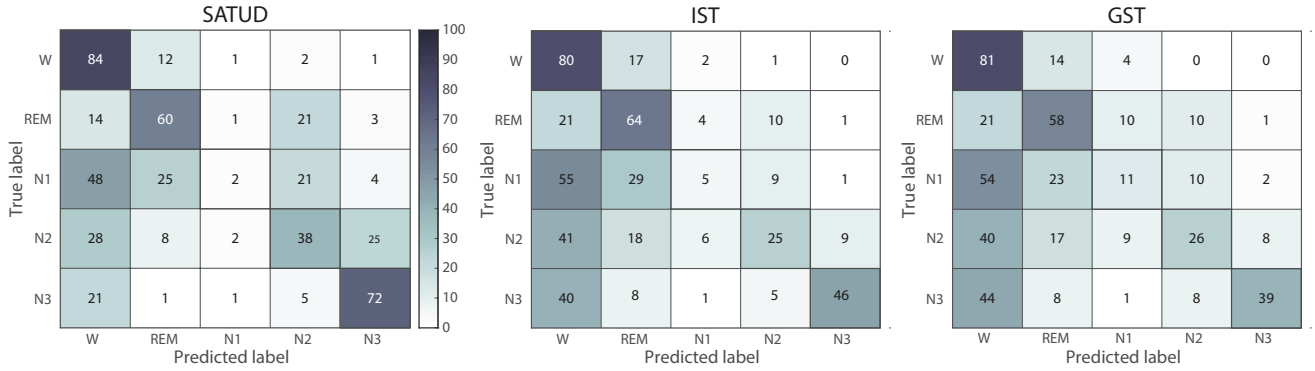


Figure 9: Classification agreement rate (percentage) between the predicted and true labels for: (a) the SATUD thresholding method, (b) IST thresholding method and (c) GST thresholding method.

raphy. Moreover such noise would also make the manual scoring a lot more complicated, leading to a probable invalidation of the recording. Artificial sweat artifacts did not have an important impact on the SATUD and IST classification agreement rate. Regarding the SATUD, they however tended to increase the number of epochs almost perfectly respecting the properties associated with their sleep stage for sleep stages N2 and N3, until a limit around 30 %-40%. It could be explained by the fact that one property of those sleep stages is the high amplitude, more often respected with the addition of sweat artifacts. Also, the number of Wake epochs almost perfectly respecting Wake properties increases with sweat artifacts when using the SATUD. The reason would be that agitated wake can make saturation appear on signals, like in presence of sweat artifacts.
The proposed method showed its efficiency, especially since the database used is composed of many patients with various pathologies.

Based on medical knowledge, this method was built to contribute to the development of a user-friendly sleep staging system. As discussed in the companion paper [REF], very few sleep staging systems are considered by physicians because of several limitations. One of them is the lack of transparency of models, often considered as black boxes. To overcome this limitation, a methodology replicating the manual scoring process was implemented in the companion paper [REF]. To give the medical practitioner concrete elements to relate to, qualitative features were chosen as the input of the classifier. However, such features can be problematic if not adjusted to each patient. The SATUD was designed to extract patient-dependent qualitative features, without the need of previous partial scoring by a sleep specialist. The SATUD method is potentially transposable to other applications. It however requires a good knowledge of the specific field, with the identification of a maximum of rules describing each class. Those numerous rules are key points for the SATUD functioning, enabling the use without the need of knowledge on the classification output. For this reason, this algorithm can run under real time conditions.

6. Conclusion

This paper presents an approach to extract features from a dataset. This approach, called SATUD, is at the core of the companion paper [REF], devoted to an automatic sleep staging for polysomnographic recordings.

The SATUD method was the outcome of an approach combining mathematical reasoning and medical knowledge. Compared to other thresholding method, the SATUD allowed a better generalization of features depending on each sleep stage. The performance proved that resulting features are significantly in agreement with the expected sleep stage properties. A straightforward classification model was also developed (to test the SATUD without being biased by the abilities of a complex classifier). Reported results were better for the SATUD, compared to the other thresholding methods. This performance, obtained on 60 patients with various sleep pathologies, confirmed that this approach is suitable for sleep staging. Tests with noise and artifacts showed this algorithm had a sufficient robustness for the application.

The companion paper [REF] presented an entire and user-friendly classification model based on the extracted features. A detailed analysis of obtained classification impact on each patient diagnostic is also worthy of investigation.

Funding

This work was supported by grants from the Institut de Recherche en Santé Respiratoire des Pays de La Loire.

Acknowledgment

The authors would like to thank Christelle Gosselin and Jean-Louis Racineux, from the Institut de Recherche en Santé Respiratoire des Pays de La Loire. We thank Julien Godey, Laetitia Moreno and Marion Vincent, sleep technicians in the Department of Respiratory and Sleep Medicine of Angers University Hospital. We thank Roberto Longo for its contribution.

A. Simplified example of the SATUD employment

This appendix presents a simplified version of the use of the SATUD for sleep stage classification. We thus consider only two patients, Patient1 and Patient2 and we admit they had a similar night. For a better understanding, we merged sleep stages N1, N2 and N3 into the so-called NREM sleep. We also considered in this appendix that wake was only composed of calm wake with eyes closed. The simplified hypnogram is represented in Figure A.1.

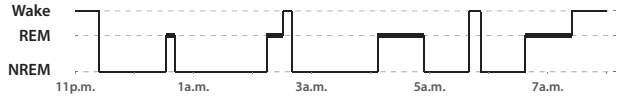


Figure A.1: Simplified hypnogram presenting the nights of patients Patient1 and Patient2.

The goal is then to retrieve the hypnogram using each patient signals, knowing that there is a variability between each patient. A highly simplified version of sleep stages description (**F1**) is presented in Table A.1. For this example, only three properties were used to describe wake, REM sleep and NREM sleep.

The SATUD aims to estimate the thresholds following the properties indicated in Table A.1. Then, a classification tool will be able to rebuild the hypnogram. For both Patient1 and Patient2, thresholds have to be adjusted until *H - L - L* $\leftrightarrow$ *Wake*, *M - MH - M* $\leftrightarrow$ *REM* and *L - L - H* $\leftrightarrow$ *NREM*.

If thresholds are well adjusted, each patient properties will be respected, as in Figure A.2. Because of the rapid eye movements that is Low during wake but also during NREM, we can see that the properties respects interdependency is not at 100 % for Wake and NREM sleep.
If thresholds are not correctly settled, and for example the rapid eye movements threshold is too high, the property *rapid eye movements is Mid or High* associated to REM sleep will never be respected. The result on the properties respects interdependency of REM sleep would then be deteriorated, as shown in Figure A.3a. For each sleep stage, the properties respects interdependency has to be maximized to find the best thresholds. However thresholds that are too far from correct values (highly erroneous) can generate a situation where the properties respects interdependency would be maximum as in Figure A.3b. To prevent this situation, fluctuation has been considered. Indeed, we can assume that all sleep stages will appear during a whole night. To estimated fluctuation, anti-scores were created using a weighted sum of the disrespect of each property.

Table A.1
Simplified sleep stage description.

	Alpha waves	Rapid eye movements	EEG amplitude
Wake[a]	High	Low	Low
REM	Mid	Mid or High	Mid
NREM	Low	Low	High

[a] Only wake with eyes closed was considered for this example.

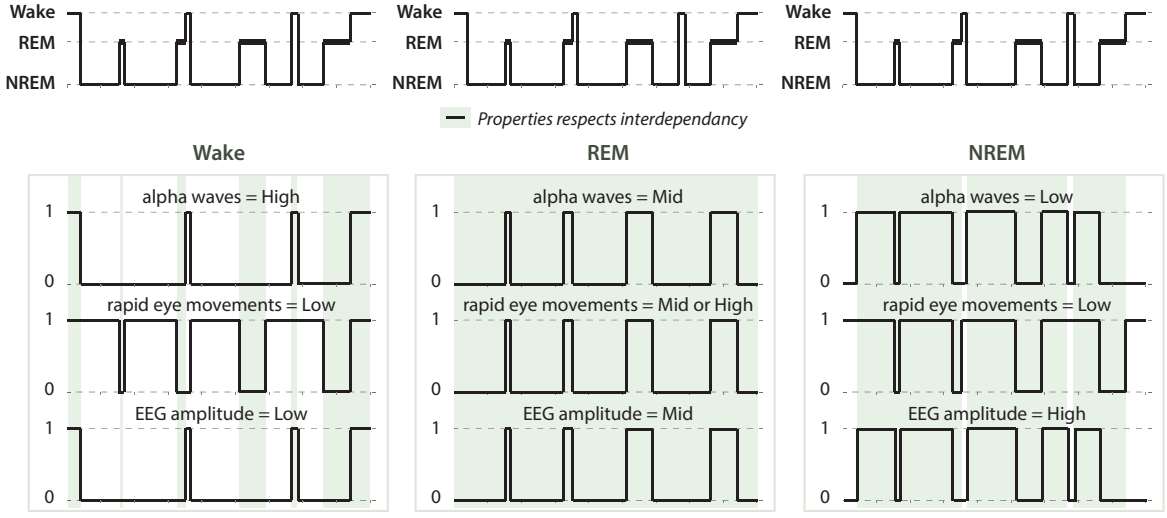


Figure A.2: Example of properties respects for each sleep stage, when thresholds are perfectly adjusted. Using Table A.1, we understand that even with perfectly adjusted thresholds, there is not always 100 % interdependency between properties respects. This is due to the possibility of some properties to be respected in different sleep stages.

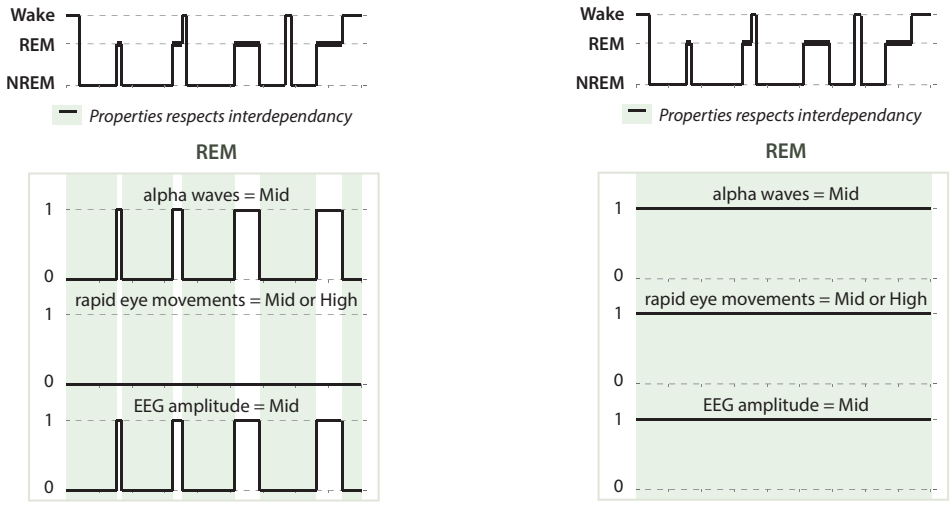


Figure A.3: Example of properties respects for REM sleep: (a) when the threshold associated with rapid eye movements is settled too high and (b) when all thresholds are highly erroneous.

Here are this example anti-scores:

$$antiScore_{Wake} = 0.4 \times \overline{alphaH} + 0.4 \times \overline{rapEyeMovL} + 0.2 \times \overline{EEGampL} \quad (A.4)$$

$$antiScore_{REM} = 0.3 \times \overline{alphaM} + 0.4 \times \overline{rapEyeMovMH} + 0.3 \times \overline{EEGampM} \quad (A.5)$$

$$antiScore_{NREM} = 0.3 \times \overline{alphaL} + 0.4 \times \overline{rapEyeMovL} + 0.3 \times \overline{EEGampH} \quad (A.6)$$

The behaviour of REM sleep anti-score in different scenarios presented in Figure A.2 and Figure A.3 is shown in Figure A.4. We can see that anti-score fluctuation is higher when thresholds are well adjusted, and null when thresholds are highly erroneous. For each sleep stage, the anti-score fluctuation has to be maximized to find the best thresholds. In order to maximize properties respects interdependency and anti-score fluctuation for each sleep stage, costs have been evaluated as:

$$cost_{Wake} = \frac{1}{conc(alphaH, rapEyeMovL, EEGampL)} \times \frac{1}{std(antiScore_{Wake})} \quad (A.7)$$

$$cost_{REM} = \frac{1}{conc(alphaM, rapEyeMovMH, EEGampM)} \times \frac{1}{std(antiScore_{REM})} \quad (A.8)$$

$$cost_{NREM} = \frac{1}{conc(alphaL, rapEyeMovL, EEGampH)} \times \frac{1}{std(antiScore_{NREM})} \quad (A.9)$$

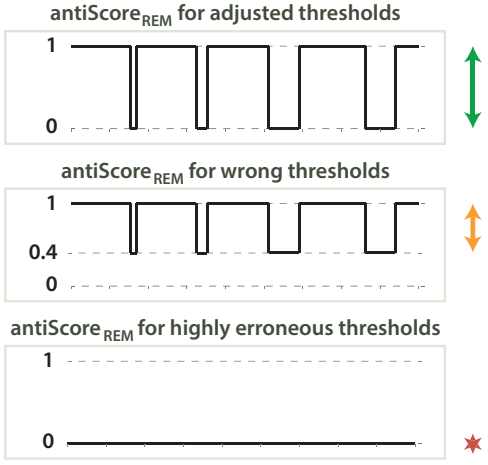


Figure A.4: REM sleep antiScore behaviour when thresholds are well adjusted, when the threshold associated with rapid eye movements is settled too high and when all thresholds are highly erroneous.

In this example, the total cost could have been defined as:

$$finalCost = 0.35 \times cost_{Wake} + 0.35 \times cost_{REM} + 0.3 \times cost_{NREM} \quad (A.10)$$

In this example, the weight associated to wake and REM sleep are higher than the weight associated to NREM sleep. Indeed, we estimated that wake and REM sleep were more complicated to detect and decided to bring them more emphasis. Equation A.10 was then minimized using several global and local search algorithms as explained in F3.

B. Pseudo-code of the SATUD

The SATUD pseudo-code is presented in Algorithm B.1. It uses here a global search algorithm (l.3-7) followed by a

local search algorithm (l.8-12). 'FinalCost' function is evaluated as described in Equation 1, with weights w_c defined empirically.

Data:

quantFeat: quantitative features

propLists: properties lists associated to each class

Result:

thresholds: patient-dependent thresholds adjusted with the SATUD

qualFeat: qualitative features

1 **Function** SATUD(*quantFeat, propLists*)**:**
2 *thresholds* ← *Stats*(*quantFeat*)
3 **while** *global stopping criterion* not respected* **do**
4 *thresholds* ← *thresholds* + *changes**
5 *qualFeat* ← *Thresholding*(*quantFeat*, *thresholds*)
6 *cost* ← *FinalCost*(*propLists*, *qualFeat*)
7 **end**
8 **while** *local stopping criterion† not respected* **do**
9 *thresholds* ← *thresholds* + *changes†*
10 *qualFeat* ← *Thresholding*(*quantFeat*, *thresholds*)
11 *cost* ← *FinalCost*(*propLists*, *qualFeat*)
12 **end**
13 **return** *thresholds* ***and*** *qualFeat*

* Depends on the global search algorithm used.
† Depends on the local search algorithm used.

Algorithm B.1: Pseudo-code of the SATUD.

References

[1] J. B. Croft, CDC's Public Health Surveillance of Sleep Health, 2017.

[2] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, K. M. Hla, Increased Prevalence of Sleep-Disordered Breathing in Adults, American Journal of Epidemiology 177 (2013) 1006–1014. doi:10.1093/aje/kws342.

[3] K. A. Franklin, E. Lindberg, Obstructive sleep apnea is a common disorder in the population- a review on the epidemiology of sleep apnea, Journal of Thoracic Disease 7 (2015) 1311–1322. doi:10.3978/j.issn.2072-1439.2015.06.11.

[4] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, R. M. lloyd, S. F. Quan, M. M. Troester, B. V. Vaughn, The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, number 2.4 in American Academy of Sleep Medicine, Darien IL, 2017.

[5] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, M. T. Bianchi, Expert-level sleep scoring with deep neural networks, Journal of the American Medical Informatics Association 25 (2018) 1643–1650. doi:10.1093/jamia/ocy131.

[6] L. Zhang, D. Fabbri, R. Upender, D. Kent, Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks, Sleep 42 (2019). doi:10.1093/sleep/zsz159.

[7] S. Enshaeifar, S. Kouchaki, C. C. Took, S. Sanei, Quaternion Singular Spectrum Analysis of Electroencephalogram With Application in Sleep Analysis, IEEE Transactions on Neural Systems and Rehabilitation Engineering 24 (2016) 57–67. doi:10.1109/TNSRE.2015.2465177.

[8] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, K. Jerbi, Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines, Journal of Neuroscience Methods 250 (2015) 94–105. doi:10.1016/j.jneumeth.2015.01.022.

[9] S. Mahvash Mohammadi, S. Kouchaki, M. Ghavami, S. Sanei, Improving time-frequency domain sleep EEG classification via singular spectrum analysis, Journal of Neuroscience Methods 273 (2016) 96–106. doi:10.1016/j.jneumeth.2016.08.008.

[10] S. Charbonnier, L. Zoubek, S. Lesecq, F. Chapotot, Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging, Computers in Biology and Medicine 41 (2011) 380–389. doi:10.1016/j.compbiomed.2011.04.001.

[11] G. Garcia-Molina, F. Abtahi, M. Lagares-Lemos, Automated NREM sleep staging using the Electro-oculogram: A pilot study, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, pp. 2255–2258. URL: http://ieeexplore.ieee.org/abstract/document/6346411/.

[12] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, H. Dickhaus, Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier, Computer Methods and Programs in Biomedicine 108 (2012) 10–19. doi:10.1016/j.cmpb.2011.11.005.

[13] M. Zokaeinikoo, Automatic Sleep Stages Classification, Ph.D. thesis, 2016.

[14] A. Ugon, Fusion Symbolique et Données Polysomnographiques, Ph.D. thesis, 2015.

[15] C. Chen, An e-health system for personalized automatic sleep stages classification, Ph.D. thesis, Université Pierre et Marie Curie - Paris VI, 2016.

[16] C. Chen, A. Ugon, C. Sun, W. Chen, C. Philippe, A. Pinna, Towards a Hybrid Expert System Based on Sleep Event's Threshold Dependencies for Automated Personalized Sleep Staging by Combining Symbolic Fusion and Differential Evolution Algorithm, IEEE Access 7 (2019) 1775–1792. doi:10.1109/ACCESS.2018.2887082.

[17] L. Fiorillo, A. Puiatti, M. Papandrea, P.-L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, F. D. Faraci, Automated sleep scoring: A review of the latest approaches, Sleep Medicine Reviews 48 (2019). doi:10.1016/j.smrv.2019.07.007.

[18] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, arXiv:1702.08608 [cs, stat] (2017). ArXiv: 1702.08608.

[19] T. Penzel, A. Sabil, The use of tracheal sounds for the diagnosis of sleep apnoea, Breathe 13 (2017) e37–e45. doi:10.1183/20734735.008817.

[20] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by Simulated Annealing, Science 220 (1983) 671–680. doi:10.1126/science.220.4598.671.

[21] J. H. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, MIT Press, 1992. Google-Books-ID: 5EgGaBkwvWcC.

[22] D. E. Goldberg, J. H. Holland, Genetic Algorithms and Machine Learning, Machine Learning 3 (1988) 95–99. doi:10.1023/A:1022602019183.

[23] R. H. Byrd, M. E. Hribar, J. Nocedal, An Interior Point Algorithm for Large-Scale Nonlinear Programming, SIAM Journal on Optimization 9 (1999) 877–900. doi:10.1137/S1052623497325107.

[24] R. Waltz, J. Morales, J. Nocedal, D. Orban, An interior algorithm for nonlinear optimization that combines line search and trust region steps, Mathematical Programming 107 (2006) 391–408. doi:10.1007/s10107-004-0560-5.

[25] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological Bulletin 76 (1971) 378–382. doi:10.1037/h0031619.

[26] E. S. Arnardottir, B. Thorleifsdottir, E. Svanborg, I. Olafsson, T. Gislason, Sleep-related sweating in obstructive sleep apnoea: association with sleep stages and blood pressure, Journal of Sleep Research 19 (2010) 122–130. doi:10.1111/j.1365-2869.2009.00743.x.

[27] R. Broughton, R. Poiré, C. Tassinari, The electrodermogram (Tarchanoff effect) during sleep, Electroencephalography and Clinical Neurophysiology 18 (1965) 691–708. doi:10.1016/0013-4694(65)90113-6.

[28] R. S. Rosenberg, S. Van Hout, The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring, Journal of Clinical Sleep Medicine (2013). doi:10.5664/jcsm.2350.