Sleep Stage Classification

PDSB 2021 — Group 6
Afonso Ferreira, 86689 | Ana Matoso, 89787 | Inês Arana, 89805

***Abstract*—Sleep is one of the most important components of our lives. During sleep, our body induces repair and rest and should that not happen we are at risk of getting sleep disorders like insomnia and sleep apnea, and multiple health consequences possibly leading to serious comorbidities. Therefore, the study of sleep stages is increasingly important in diagnosing these kind of diseases. In this paper, the effect of preprocessing in the classification of sleep stages is studied, more specifically the application of Independent Component Analysis (ICA) and conventional filters such as high-pass, low-pass and notch filters. It was demonstrated that by applying just ICA the classification of sleep stages yields the best results in testing more specifically, 77.04% of accuracy and a Cohen's Kappa of 0.686. Nevertheless, using unfiltered signals or just conventionally filtered signals yields very similar values and both show substantial agreement with the ground-truth. Thus, it can be concluded that the classification of sleep is slightly improved by just applying ICA, whereas using just filters slightly worsens the results. On the contrary, using filters and ICA together has a diminished Cohen's Kappa (0.428) and accuracy (59.94%). As such, for this case it was tested a model with fewer features which improved results though it did not reach the values of the other models. On the whole, this paper was a great opportunity to study about the mechanics of sleep and to develop our signal processing skills.**

***Index Terms*—Polysomnography, EEG, ECG, EMG, sleep stages, REM**

1. PROBLEM AND MOTIVATION

It cannot be denied that sleep is a crucial part of life, whether by diminishing tiredness or by inducing repair among organs. In fact, it is estimated that people spend a third of their life sleeping [1]. Nevertheless, in the fast paced world we live in, many people underestimate the importance of sleeping, thus increasing the probability of getting sleeping disorders such as insomnia, sleep apnea syndrome and narcolepsy [2].

With regard to sleep duration, apart from intra- and inter-individual variation, the American Academy of Sleep Medicine (AASM) and Sleep Research Society defined 7 hours of sleep as the threshold below which there is a strong correlation with health conditions in several domains [3]. In addition, in this same study, the range of 7–9 hours of sleep was agreed to be the optimal physiological sleep duration among various experts. Therefore, generally, people who sleep less than 7 hours per night in the long term are at risk of later dealing with health consequences related to cardiovascular health, metabolic health, mental health, immunologic health, cognitive performance and pain. Health conditions affecting these domains may co-occur, resulting in aggravated comorbidities.

The diagnosis of sleep disorders is often done with a Polysomnography (PSG) test which is a multi-modal monitoring conducted overnight whilst the patient is asleep that measures several biosignals in the patient. Usually, the set of signals consisting the PSG include Electroencephalograms (EEGs), Electrooculographies (EOGs), Electromyographies (EMGs) and Electrocardiographies (ECGs), among others.

According to the AASM scoring rules, physiological sleep can be broken down in 5 stages if we include the state of full awareness prior to sleep: W (awake), N1, N2, N3 and R.

Each sleep phase is characterized by specific behaviours, electrophysiological activity and physiological changes, which help experts performing manual segmentation for sleep quality assessment. First, behaviours such as body posture changes and responsiveness to stimuli can be assessed using EMGs, position sensors and infrared video cameras [4]. Second, electrophysiological activity can be assessed from the EEGs, and EOGs, as well as EMG signals. Third, the physiological changes can be evaluated from the heart-rate, body temperature and breathing motion.

As for the specifics of each sleep stage, they are characterized by different frequency bands and time domain features.

Stage W usually progresses from an active mind with opening and closing of the eyes to a state of drowsiness where the eyes remain close. This progression is also characterized by a decrease in beta brain waves (18-25Hz) and an increase in alpha waves (8-13Hz) in the EEGs. Submental EMGs are likely to contain high-amplitude motions as chin muscles are not fully relaxed at this point, and EOGs often show eye blinking and fast motions that tend to fade with time. Additionally, the subject may temporally alternate between stages W and N1 before entering the latter. An epoch is considered W if more than 50% of the brain waves are alpha. [5]

After this, the individual enters stage N1, also known as light sleep, which is usually short in duration (i.e. 1 to 7 minutes [5]) and characterized by variable but generally low-voltage patterns. This includes vertex sharp waves, mostly occurring at the end of this stage. The EMG expresses less activity compared to stage W and is associated with a constant tone as opposed to contractions that alternate with relaxation. At this point the heart rate stabilizes along with breathing. An epoch is considered N1 if more than 50% of the brain waves are theta (4-8Hz). [5]

Stage N2, also referred as intermediate sleep, then follows the previous one. It usually lasts at least 20 minutes and is characterized by predominant theta activity and fast activity features in the EEG. It is only at this phase that K-complexes and sleep spindles become noticeable, and they tend to appear as episodic events. K-complexes can be identified as sharp slow waves consisting of a negative deflection followed by a positive one, standing out from other lower amplitude EEG features. Sleep spindles are characterized by sinusoidal-like waveforms that wane quickly, also co-occurring with other slower waves. Both K-complexes and sleep spindles need to last longer than half a second in order to be identified as such. For a sleep stage to be scored as N2, it needs to contain at least 20% of delta activity (0.5-2Hz) in given epoch. With regard to EMG and EOG, there are no determinants features for scoring an epoch as N2. However, heart rate as well as responsiveness to stimuli tend to further decrease with time. [5]

Henceforth, the individual enters a state of deep sleep. First, they undergo stage N3, which is regarded as the most effective phase at contributing to restoring freshness. As for its specific features, the EEG shows with high-amplitude slow waves, whereas there are no determinant features for EMG and EOG.

Finally, if the sleep further progresses, stage R, also referred as Rapid Eye Movement (REM) sleep, is achieved. This phase usually first occurs roughly 90 to 120 minutes after the beginning of sleep (i.e. latency onset) and tends to cycle on this same time range throughout the night. During this stage, EEGs show low-amplitude waves with mixed frequencies: both theta waves and alpha waves slightly lower in frequency tend to occur. Despite being the furthest sleep stage achieved, it shows higher physiological activity compared to stage N3. For instance, heart rate, among other biosignals, is increased, and it is even common for individuals to undergo sexual arousal. In addition, the EOGs show significant bursts of activity, the REMs, for which the phase was named after. This phase, like the aforementioned ones, can also include K-complexes and sleep spindles. Its scoring is thus based on low-amplitude mixed-frequency EEGs, as well as low submental EMG tone and the presence of REMs on the EOG. [5]

Apart from the AASM scoring rules, there are other scoring variants such as the Rechtschaffen & Kales (R&K), which was mostly substituted by the former. Although both models do not yield the same scoring for all sleep stages because of different scoring criteria, the main difference is that stage N3 is divided into two stages, yielding the following for the R&K model: W, S1, S2, S3, S4 and R [6]. As a result, for instance, for the R&K rules, measuring delta wave content is performed using central electrode to score stages S3 and S4, whereas for the AASM rules, this wave band is measured using frontal leads to help scoring stage N3.

Usually, given a PSG, the sleep stages of its epochs are manually annotated by a doctor, which can be an extremely laborious process, especially time-wise. As such, this process can be automated by implementing a machine learning algorithm that automatically classifies sleep stages. By doing this, time can be saved in diagnosing the patient and thus facilitating the assessment of sleep disorders.

Therefore, this project aims at comparing preprocessing approaches and their performances when classifying sleep stages.

2. BACKGROUND AND RELATED WORK

Sleep stages scoring, as it was mentioned before, is usually done by hand by a doctor in a hospital setting. However, it can be a rather subjective task. In fact, a study [7] demonstrated that doctors from different sleep centers can disagree in scoring sleep stages (especially between N2-N3, W-N1, and N1-N2 having a discrepancy ratio of 22.09%, 19.68%, and 18.75% respectively). Therefore, it is necessary to streamline the process of sleep stage scoring in order to have a better accuracy and agreement. As a matter of fact, there are already algorithms based on machine learning that outperform human scorers [8].

In the literature, the articles that aim at classifying sleep stages through PSG signals do all of the processes (preprocessing, feature selection/extraction and classification) separately, thus an analysis on the state-of-the-art methods of each of these processes was made.

Regarding the preprocessing method, ICA is a method useful in separating inherent characteristics of a given set of signals within a given modality, different modalities or both, as long as there is prior knowledge of possible cross-contamination arising from different sources. For instance, applying ICA on a set of EEGs for artefact removal, yields de-correlated brain source signals. This preprocessing procedure is often used in combination with high-pass filering for the removal of EOG artefacts from EEG signals, but can also be used alone as a preprocessing method. In one study, roughly 90% agreement with human expert could be achieved in sleep stage classification by de-correlating all EEG signals using ICA, without using any other filtering methods [9]. Another study concluded that applying high-pass filtering followed by ICA separation of EEG signal components improved the accuracy of artefact classification from $84.2 \pm 1.0\%$ (no preprocessing) up to $85.7 \pm 0.7\%$ [10]. Finally, several studies [11][12][13] classified sleep stages using unfiltered data of EEGs and showed no significant decrease in accuracy and some even show improvements.

Concerning feature selection, there is one study [14] that used principal component analysis (PCA) to determine how well the features extracted separate the sleep stages. Another study [15], used support vector machine (SVM) to classify the sleep stages using spectral features. Similarly, entropy based features together with SVM were also used [16][17]. It is also reported that using non-linear features such as the fractal dimension paired up with spectral features yields better results than previous studies [18][19]. Additionally, one study [20] used heart-rate variability features paired with long short-term memory neural networks and reported an accuracy of 77±8.9%.

About classification methods, one study [21] used decision trees for automated sleep scoring which yielded 82% and 79% accuracy in training and testing, respectively. Lately, with the development of various machine learning algorithms, many deep learning classification methods are being applied [22]. One of the deep learning techniques that has been emerging is convolutional neural networks (CNN) [11]. In fact, one study [23] used CNN to classify 20 healthy subjects on single EEG channel data achieving 74% accuracy. In addition, there is one study [24] that used a 11 layer 2D CNN model that showed similar performance in the classification of sleep stages when limiting the number of EEG channels (from 20 to 6).

Based on the reviewed research articles, in this project it was opted to test various conventional filtering methods (high-pass, low-pass and notch filters) and ICA for the removal of EOG artefacts on the signals contained in the PSG dataset. As for the features, fractal dimension and spectral features were among the features included given the possibility of improving the classification performance. Finally, given the nature of this academic research, in order to avoid a high computational cost and memory, an SVM classifier was used. The next section will describe these steps of the pipeline in a more detailed manner.

3. APPROACH AND UNIQUENESS

A. *Material*

Initially, 28 European Data Format (EDF) files were provided, each corresponding to a PSG of a unique patient. Some of these files corresponded to patients with specific diseases or disabilities somehow related to sleeping (e.g. bruxism and snoring), whereas 5 of them corresponded to disease-free PSGs patient recordings. Furthermore, these files contained different combinations of recorded signal modalities or transducers. Given the great amount of storage and processing power required for all the 28 datasets, as well as the variability in the available signal types, a smaller set of patients within the 28 was selected. The 5 "normal" patients' PSGs were considered for this purpose (i.e. the files with "n" as a prefix: n1.edf, n2.edf, n3.edf, n5.edf and n11.edf). Afterwards, all the available signal types within the normal dataset were cross-referenced, allowing to discard signals of each patient that are not common to all patients. This cross reference yielded 9 signals for each patient file, in which 4 EEGs (C4A1, C4P4, F4C4, and P4O2), a heart rate, an EOG (ROCLOC), an ECG and 2 EMGs (EMG1EMG2 and SX1SX2 that correspond to chin and left leg EMGs, respectively [25]) were included. Additionally, 28 text files containing annotations of the sleep stages of the patients were provided. It is important to note that in these files the sleep stages are classified for each period of 30 seconds, according to the R&K sleep scoring (6 sleep stages). Moreover, each sleep stage was converted to a scale of numbers so that they can be processed, therefore, 0,1,2,3,4,5 correspond to stages R, S4, S3, S2, S1, and W, respectively.

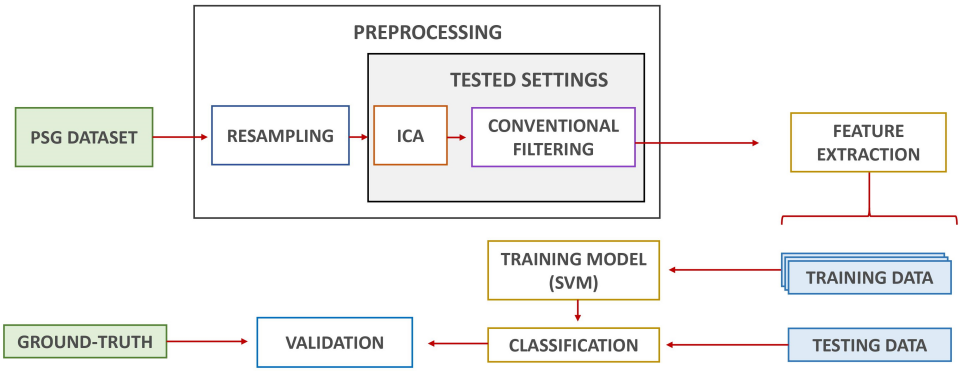

Fig. 1: Sketch of the sleep stage classification pipeline. The leftmost boxes filled in green represent the PSG dataset after data selection and cropping for stage synchronization (PSG dataset) and the manually annotated sleep stages in epochs (ground-truth). The blue box (validation) on the bottom represents the last step, after which data analysis was performed.

Overall, a total of 5×9 recorded signals and 5 ground-truths were used. The number of 30-second epochs and duration of each patient's PSG can be found in Table I.

TABLE I: Summary of the normal PSG dataset used in the pipeline.

Patient/PSG label	Duration	Nº epochs	Effective Duration
n1	9h 37m	1141	9h 30m 30s
n2	8h 47m	1052	8h 46m
n3	12h 15m	999	8h 19m 30s
n5	9h 11m 1s	999	8h 19m 30s
n11	8h 44m 1s	1006	8h 23m

B. *Methods*

After selecting the dataset to be used in the sleep classification, four scenarios of preprocessing techniques were tested: no preprocessing, just ICA, just filters, and ICA together with filters. The process for the latter scenario is described (for the other scenarios the respective steps were skipped).

First, the signals were resampled to the highest sampling frequency found across the 5 patients' signal sets, 512 Hz.

Then, ICA was applied to remove possible sources of ocular motion artefacts from the 4 EEGs. Additionally, highpass and lowpass filters were applied, as well as a notch filter to remove the powerline.

Afterwards, the segmentation of the signals in 30s epochs was made and with the sleep stages annotations a column vector that will work as the response vector of the machine learning algorithm (supervised learning) was built. It is important to note that the number of rows of the sleep stages need to be the same as the number of 30s segments which initially did not happen. One of the reasons why was because the annotations and the EDF files had different start times so it was needed to synchronize the start time. Despite that, they were still different so the data was truncated in the end so that both the sleep stage annotations and the number of epochs matched. Therefore the effective duration, i.e. the one used for classification, can be seen in table I.

The next stage is the feature selection and extraction which is explained in the next section. Some features chosen include the relative power of each EEG band, heart rate, number of K-complexes in the EEG and the fractal dimension. These features are determined and calculated for each epoch and a feature matrix is constructed so that each row corresponds to an epoch and each column to a variable (feature).

After all this, the feature matrix and the response vector was plugged into the "Classification Learner App" in MATLAB and a support vector machine algorithm was applied (since it was the one which was more accurate). For this purpose, PSGs n1, n2, n3 and n5 were used as training data and n11 as testing data.

Overall, the steps of the sleep classification pipeline described here are illustrated in Figure 1.

C. *Proposed solution*

In this section, we will delve deeper into the methods used during the preprocessing and feature extraction steps.

Preprocessing

As for the preprocessing step, ICA was first considered to remove possible sources of ocular motion artefacts from the 4 EEGs. For this purpose, all the signals from all patients were first resampled to the highest sampling frequency found, 512 Hz, using linear interpolation. Table II shows the sampling frequencies found for all signal types across the 5 patients/PSG labels, where the last row gives the highest sampling frequency found for all signal types among the used PSG dataset.

TABLE II: Summary of the sampling frequencies [Hz] found across the various signal types for each PSG label.

PSG label	C4A1	C4P4	ECG1ECG2	EMG1EMG2	F4C4	HR	P4O2	ROCLOC	SX1SX2
n1	512	512	512	256	512	1	512	512	256
n2	512	512	512	512	512	1	512	128	128
n3	512	512	512	512	512	1	512	128	128
n5	512	512	512	512	512	1	512	128	128
n11	512	512	512	512	512	1	512	128	128
Max. S.F.	512								

Before ICA, all PSG signals were normalized to yield mean amplitudes μ of 0, and standard deviations σ of 1, according to the z-score formula in Equation 1.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Now, let the $m \times n$ matrix X_k be the PSG data $k \; \forall \in \{1, 2, 3, 4, 5\}$ where each row corresponds to one of the m signal types with n data points. Equation 2 assumes resampling all m signal types to the same sampling frequency, as previously mentioned.

$$X_k = \begin{bmatrix} x_{11_k} & x_{12_k} & x_{12_k} & \dots & x_{1n_k} \\ x_{21_k} & x_{22_k} & x_{12_k} & \dots & x_{2n_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1_k} & x_{m2_k} & x_{m2_k} & \dots & x_{mn_k} \end{bmatrix} \tag{2}$$

The goal of ICA is to estimate the weights $a_{i,j}$ of the $m \times m$ mixing matrix A (equation 3) and the source vectors of the $m \times n$ matrix S_k that satisfy Equation 4. This estimation is achieved by assuming independence and non-Gaussianity of S.

$$A_k = \begin{bmatrix} a_{11_k} & a_{12_k} & \dots & a_{1n_k} \\ a_{21_k} & a_{22_k} & \dots & a_{2n_k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1_k} & a_{m2_k} & \dots & a_{mn_k} \end{bmatrix} \tag{3}$$

$$X_k = A_k \times S_k \tag{4}$$

In other words, if we consider each signal vector x_{i_k} in the i^{th} row of matrix X_k, then it can be decomposed in a linear combination of m source signals, as in Equation 5.

$$x_{i_k} = a_{i1_k} s_{1_k} + a_{i2_k} s_{2_k} + \dots + a_{im_k} s_{m_k} = \sum_{j=1}^{m} a_{ij_k} s_{j_k} \tag{5}$$

It is then clear that if we apply ICA to find the source signals $s_{j_k}\ \forall j \in \{1, 2, ..., m\}$ and then eliminate a column j in matrix A_k, we can then reapply Equation 4 to obtain all "reconstructed" signals without the contribution of the source j. In practice, the goal in this step of the preprocessing was to eliminate EOG contamination from EEGs only. Therefore, the source signal of the EOG activity was eliminated in all signal types of the PSGs before reconstruction, and then we replaced all non-EEG signals by the original ones. The reason for not removing other source components was that after resampling, various combinations of sources to be removed from the EEGs were tested, and based on this trial and error, only removing the source components of the ROCLOC signal seemed to result in appropriate filtered signals. For instance, the removal the ECG source from the EEGs led to inconsistent filtered signals: either they were previously filtered in this domain or there was no consistency between ECG signal and artefacts produced in the EEG.

After ICA, all PSG signals were converted back to the original distributions, meaning the original μ and σ values of each signal were restored, by substituting the ICA-processed signals in z and the original μ and σ values in Equation 1.

Afterwards, some filters were applied to remove artifacts and unwanted frequencies (referred as "conventional filtering" in Figure 1). The filters applied were highpass, lowpass and notch filters. The highpass filter was a 4^{th} order IIR Butterworth filter with a cutoff frequency of $5\,\text{Hz}$ for EMG and $0.5\,\text{Hz}$ for the rest of the signals. These filters were used to remove lower frequency noise as well as the baseline wander. The lowpass filter was a 4^{th} order IIR Butterworth filter with a cutoff frequency of $200\,\text{Hz}$. Finally, a 2^{nd} order notch filter with a cutoff frequency of $50\,\text{Hz}$ was applied to all signals to remove the powerline interference. [26]

Feature Extraction and Selection

Regarding the feature selection and extraction, several features were chosen in each type of signal.

Some of the most used features in the classification of sleep stages are the band power (or percentage of it) of each type of wave in the EEG. The bands that were considered were the alpha (8-13Hz), beta (18-25Hz), theta (4-8Hz) and delta (0.5-2Hz) [5]. In this project, the percentage of each band in each epoch was calculated. To do this, the "bandpower" function of MATLAB was applied, which uses a modified periodogram to calculate the average power [27].

Moreover, K-complexes, which are negative sharp slow waves that last at least 0.5s, were also detected. In order to detect them the negative and positive peaks are calculated (with empiric thresholds) and then it is checked whether there is a negative peak soon after a positive one and if the amplitude is above a certain threshold (empirically determined by observing the signal) [5].

In addition, sleep spindles were also detected, which occur when the signal has a region of 11Hz to 16Hz that lasts at least 0.5s [5]. This feature was determined by using the "bandpower" function mentioned above.

What's more, the Petrosian's algorithm was used to calculate the fractal dimension (PFD) which is based on the sign changes of the signal [28]. This was calculated using the following formula:

$$PFD = \frac{\log_{10} n}{\log_{10} n + \log_{10}\left(\frac{n}{n+0.4N_\delta}\right)} \tag{6}$$

where n is the number of data points and N_δ is the number of sign changes (zero crossings). This algorithm is one of the simplest and fastest to calculate the fractal dimension.

In addition, the Hjorth parameters were also calculated, which are three quantitative descriptors of the EEG: activity (HA), mobility (HM) and complexity (HC) [28]. These were calculated using the following formula:

$$HA = \sigma_0^2; HM = \sigma_1/\sigma_0; HC = \sqrt{(\sigma_2/\sigma_1)^2 - (\sigma_1/\sigma_0)^2} \tag{7}$$

where σ_i is the variance of the i^{th} derivative.

Finally, some statistical features were also calculated such as the kurtosis and the skewness [29].

Regarding the ECG features, the main one to consider is the heart rate. Although there was one ECG channel, there was also a dedicated heart rate channel that recorded only the heart rate, so that channel was used instead of calculating the heart rate through the ECG.

Additionally, features related to the heart rate variability of each epoch were also extracted: the average, the standard deviation and the difference between the biggest and the smallest time between consecutive beats [20].

Concerning the EMG signal, the maximum peak to peak amplitude and the root mean square of each epoch were calculated [30].

In the EOG, the number of blinks was counted by calculating number of peaks with a threshold of $75\,\mu\text{V}$ and a minimum peak distance of about 4s. These threshold values were empirically determined by looking at the EOG signal [30].

Another feature that was thought of was rapid eye movements (REM), however, since there was only one EOG channel available, it was not possible to calculate it, since it is characterized by conjugate saccades in both EOGs. Similarly, slow eye movements were not possible to detect for the same reason. [5]

Classification

Afterwards, as it was mentioned before, the feature matrix is calculated and is put into the "Classification Learner App" in MATLAB along with the sleep stages (response variable) in order to train a machine learning algorithm, specifically, a medium Gaussian support vector machine. In order to get a better estimate of accuracy, it was chosen to perform a 5-fold cross validation instead of a simple hold-out. For a more visual representation of the model's performance, the confusion matrix of each model has been plotted.

The trained model is then exported in order to be tested and validated on the last patient by calculating the accuracy. Another metric used to evaluate the model was the Cohen's Kappa [31]. This

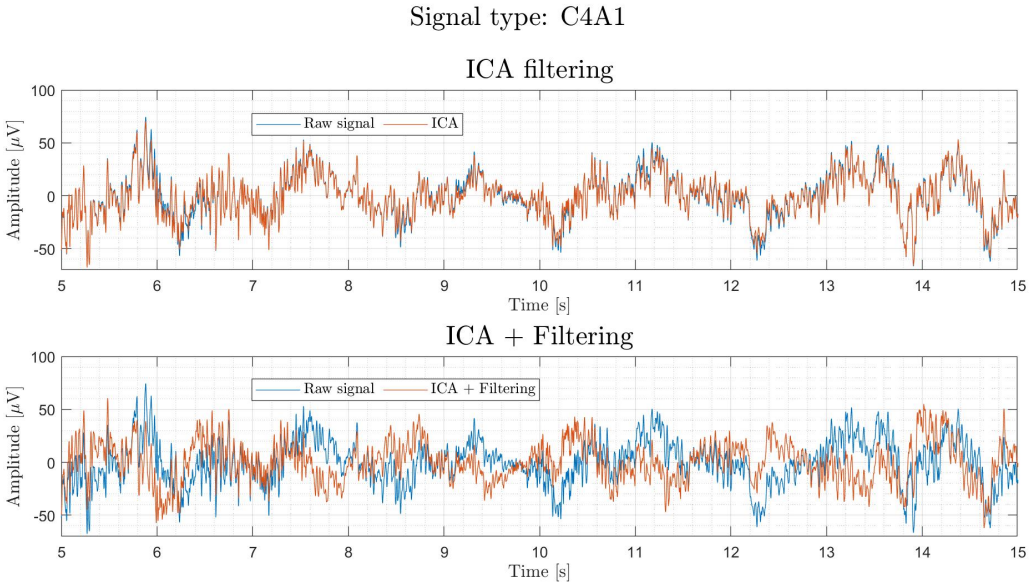

Fig. 2: Comparison of two different preprocessing settings for a 10-second segment of the EEG C4A1 signal of the N1 dataset.

metric is used as a method of measuring inter-rater reliability and it is widely thought to be more robust than the accuracy since it takes into account the possibility of the agreement occurring by chance. Cohen's Kappa values greater than 0.80, between 0.61 and 0.80, 0.41 and 0.60, 0.21 and 0.40 and 0 and 0.20 represent almost perfect, substantial, moderate, fair, and slight agreement respectively.

In the end, a feature analysis was performed, comparing the results obtained with the different models, and analysing the distribution of feature values through the different sleep stages.

4. RESULTS AND CONTRIBUTIONS

Classification

On this section the some of the results of the various steps of the sleep classification pipeline are presented, especially for the classification step considering the different preprocessing settings.

Preprocessing

Figure 2 shows a 15-second segment of the EEG C4A1 signal of the n1 dataset so as to exemplify the extent of artefact attenuation using different preprocessing settings. On the top plot ICA was applied for EOG artefact removal, resulting in minor changes in the EEG segment. On the other hand, the bottom plot shows a high degree of attenuation especially for low-frequencies below 0.5Hz after adding manual filtering to the ICA filtering. From this figure we expected both manual filtering and the use of both preprocessing settings to impact the extracted features more than ICA alone, compared to unpreprocessed (raw) signals.

In figures 3, 4, 5 and 6 the confusion matrices yielded for the training data for the four preprocessing scenarios are represented: no preprocessing, just ICA, just filtering, and filtering together with ICA, respectively. In all confusion matrixes it is verified that the diagonal has the bigger values, which means that the classifier predicts the correct stage frequently.

In addition, it can be seen that, in training, in all scenarios, S2 is the sleep stage that is more accurate. Moreover, the classifier also predicts REM and S4 rather accurately but missclassifies S2 and S4 quite often.

It is also important to note that when the true class is S1, in none of the scenarios it is predicted to be S4 or S3. Similarly, R is never missclassified as S4 in all scenarios.

In table III it is represented the average and standard deviation of the amount of sleep stages present. It can be seen that the most common sleep stage is S3 and the least common is S4.

TABLE III: Number of sleep stages present in the data set

Sleep Stage	Training data	Testing data
W	93.5 ± 53.6	9
S1	187.5 ± 70.6	169
S2	97.8 ± 32.4	134
S3	372.3 ± 101.1	413
S4	57.3 ± 58.9	49
R	239.5 ± 100.4	232

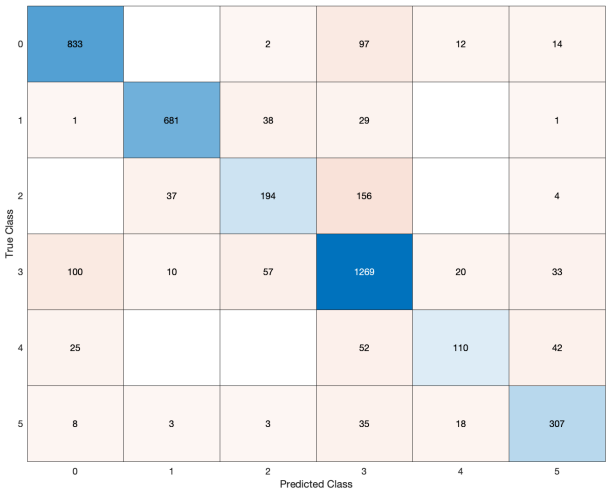

Fig. 3: Confusion matrix of the classification of raw signals

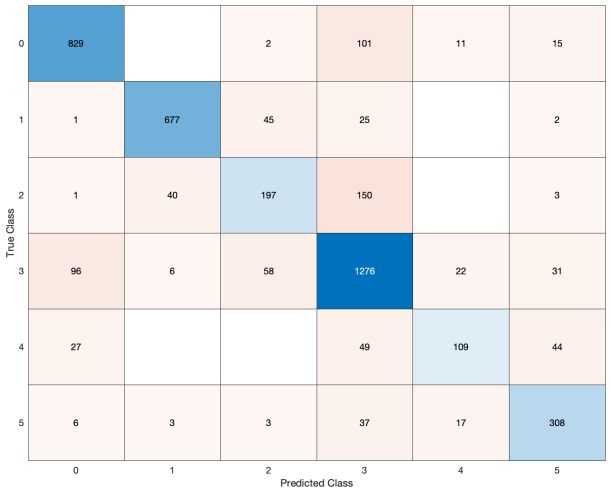

Fig. 4: Confusion matrix of the classification of ICA filtered signals (no filtration)

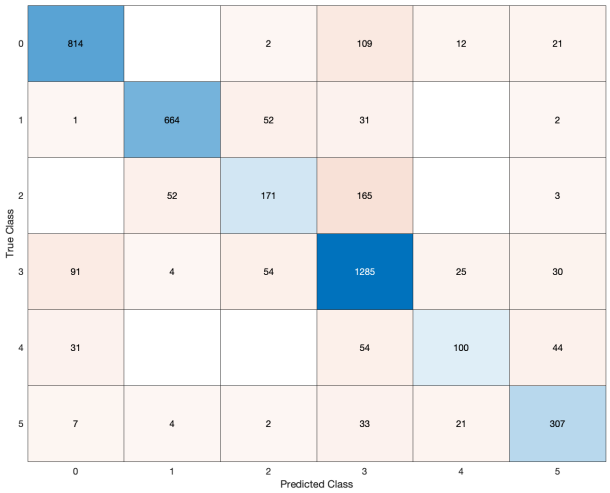

Fig. 5: Confusion matrix of the classification of filtered signals (no ICA)

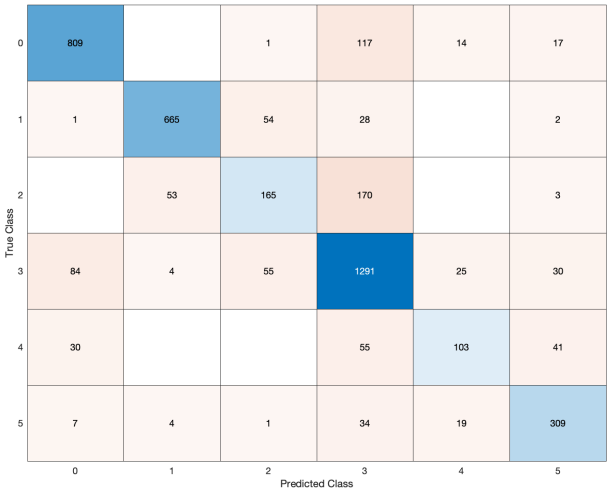

Fig. 6: Confusion matrix of the classification of filtered and ICA filtered signals signals

Figure 7 shows the sensitivities obtained for each sleep stage for the testing signal n11, considering the different preprocessing settings.

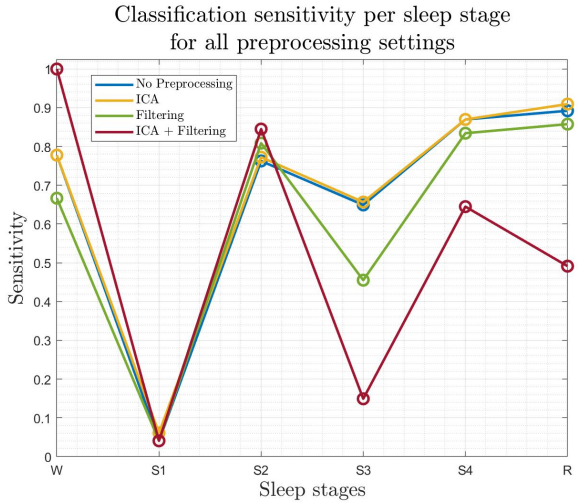

Fig. 7: Sensitivities of each sleep stage in each preprocessing scenario when testing with patient n11

Overall, it can be seen that all four models are not sensitive at all to S1, having all less than 0.1 sensitivity. Sensitivities for no preprocessing and ICA seem to be very similar for all sleep stages. Additionally, manual filtering performed slightly worse than the aforementioned settings, and adding ICA to manual filtering benefited the classification for stage W, but resulted in a decrease for stages S3, S4 and R.

Figures 8 and 9 show the posterior class probabilities of the epochs of the signal n11 for each sleep stage, for correctly labelled epochs, when using no preprocessing and both ICA and manual filtering, respectively. These results show that for stage W, training the model with no preprocessing led to statistically higher posterior probabilities for epochs known to be within this stage, when considering new information in the tested signal. However, when using both ICA and filtering, these probabilities were more robust for the same stage. Epochs known to be within other stages seem be associated with higher probabilities for both ICA and filtering compared to filtering. As for the preprocessing trials using ICA only and filtering only, the statistical distributions of posterior probabilities did not differ significantly from the one for no preprocessing, hence they are not illustrated here.

In table IV it can be found a summary of the Cohen's Kappa and accuracy values in training and in testing with the patient n11.

TABLE IV: Summary of Cohen's Kappa (CK) and accuracy in training and in testing.

Type of processing	Training		Testing	
	Accuracy	CK	Accuracy	CK
Raw Signals	81 %	0.750	76.04 %	0.690
Using just ICA	81 %	0.751	77.04 %	0.686
Using just Filters	79.7 %	0.733	73.86 %	0.638
Using ICA and Filters	79.7 %	0.733	59.94 %	0.428

As it can be seen, looking at the Cohen's Kappa values, all methods show a substantial agreement except the testing phase with ICA and filtration which shows only moderate agreement.

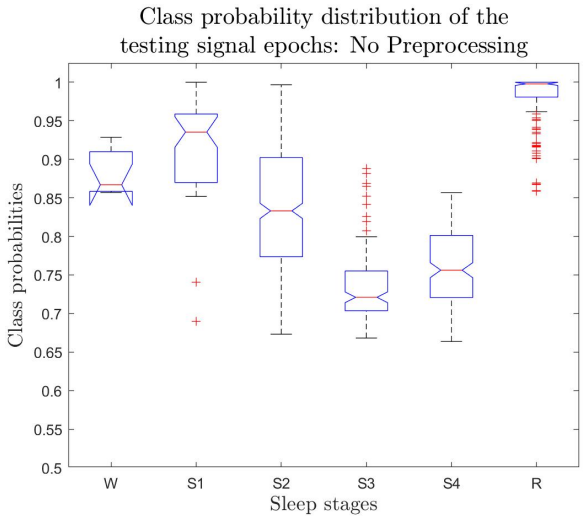

Fig. 8: Posterior probabilities of correctly labeled epochs per sleep stage for all sleep stages, using the model trained with no preprocessing.

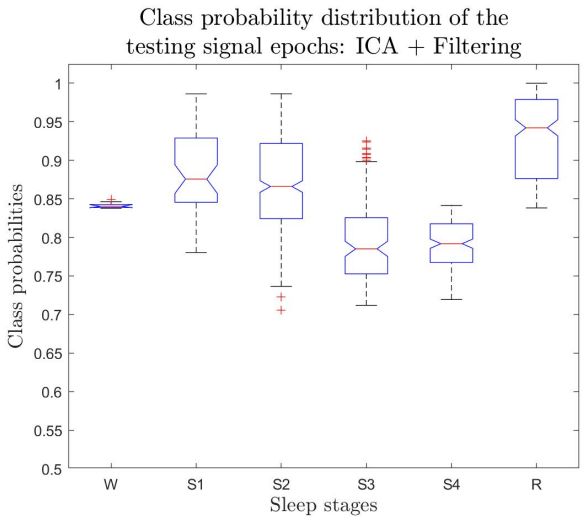

Fig. 9: Posterior probabilities of correctly labeled epochs per sleep stage for all sleep stages, using the model trained with ICA and filtering.

The most accurate scenario/model overall in the training phase as well as in the testing phase was the one in which just ICA was applied to the signals. It can also be seen that the accuracy of all models is similar in training. On the other hand, regarding the Cohen's Kappa, the model where the signals are filtrated and where ICA is applied shows a very low value.

Given all these values, it was chosen the best processing scenario in testing (just applying ICA) and the hypnogram comparing the real and the predicted sleep stages was plotted. In Figure 10 it is present the comparison between the n11's real sleep stages, represented in orange, and the predicted ones that were calculated in the machine learning algorithm (ML) when the signal has no preprocessing at all, represented in blue.

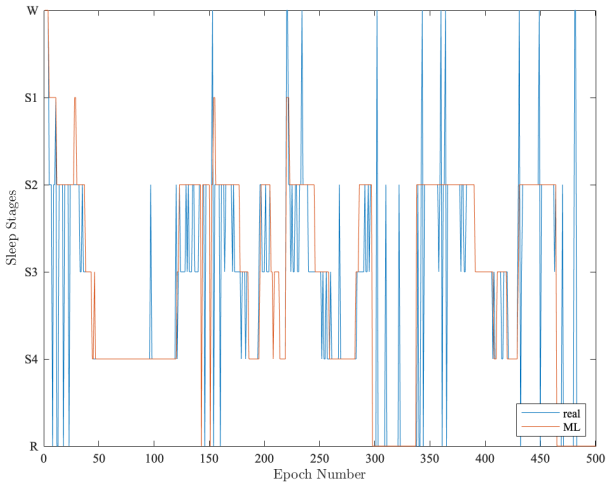

Fig. 10: Comparison of the real Hypnogram with the predicted one for the case of just applying ICA

Feature Analysis

A feature analysis was also performed, using the results from testing the patient n11 with the training model where the signals were processed with both ICA and filters. This was done by analysing the boxplots with the distribution of feature values through the different sleep stages, such as the one in figure 11 regarding sleep spindles in the different EEG channels.

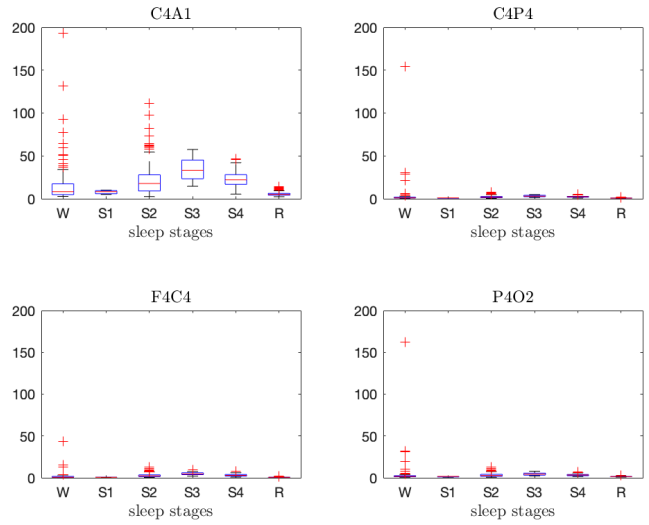

Fig. 11: Comparison of the values of obtained for the sleep spindles feature between different EEG channels

In general, EEG band features such as alpha, beta, theta and delta waves showed good variability between sleep stages, hinting their value in classification, while the skewness of the signals did not since its values remained more or less constant during the different stages. Other features, namely kurtosis, k-complexes and Hjorth activity showed similar values in all stages but one, stage W. The other Hjorth features, mobility and complexity and PFD showed good variability across all stages.

All EMG features computed have distinct values during stage 5 (Wake). For the EOG signal, the only feature computed was the number of blinks, which was higher during stage 5 than the others, as is expected.

Using the information from this feature analysis, another model was trained using a feature matrix where some feature had been removed. For the K-complex, PFD and sleep spindles, only the feature computed from the EEG channel C4-A1 were used, and the skewness feature from all signals was also removed, resulting in 44 features, as opposed to the previous 57. This allowed the training model to be less complex. The results from this analysis are plotted in figure 12, together with the previous sensitivities for other preprocessing settings using the whole initial feature space.

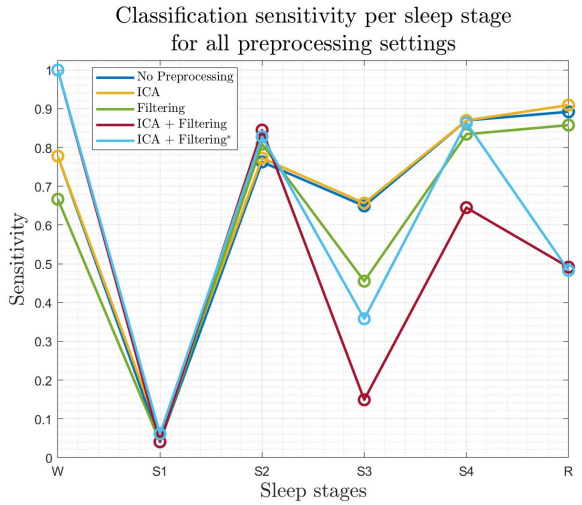

Fig. 12: Comparison of sensitivities of each sleep stage in each preprocessing scenario when testing with patient n11, including the model trained with data preprocessed with ICA and filtering and 44 selected features (represented with an asterisk).

Figure 12 suggests that reducing the number of features in the feature extraction step leads to at least the same sensitivities for classifying the testing data when compared to the whole feature space for the same preprocessing settings: stages W, S1 and R yielded approximately the same sensitivities, whereas for stages S3 and S4 the sensitivities increased more than 10%.

Similarly as with the other models, the accuracy and Cohen's Kappa were computed for this modified model. The training accuracy obtained was 79.8% and the corresponding Cohen's Kappa 0.73 (substantial agreement), whilst the testing accuracy was 65.6% and the Cohen's Kappa 0.51 (moderate agreement).

5. DISCUSSIONS AND CONCLUSIONS

Sleep stages and preprocessing settings

When looking at all confusion matrices it is verified that the diagonal has the bigger values, which means that the classifier predicts the correct stage rather frequently. In addition, the best qualified model in training is S2. This is in accordance with the sensitivities for the testing data, which were relatively high and also very robust among all preprocessing settings. In addition, the proportion of epochs known to belong to stage S2 in the testing data is not within one standard deviation of the number of S2 epochs of the training data, but is very close to this range. The posterior probabilities of epochs labeled as S2 showed the highest variability among all stages and preprocessing settings, which does not seem to have affected the classification sensitivity for this stage.

Next, from all confusion matrices for the training data, stage R was the second best at being correctly predicted. This stage agrees with the results for the testing data. Firstly, the number of R epochs in this signal does not differ significantly to the proportions of the training data, hence we exclude the problem of under or overrepresentedness of this class. Secondly, sensitivities are also relatively high for the testing phase, compared to other sleep stages, except for preprocessing using ICA together with filtering (roughly 0.5 compared to [0.85-0.925] for other sleep stages). Thirdly, posterior class probabilities for this stage were relatively high for all preprocessing settings, apart from the presence of outliers (except of ICA together with filtering) which were still far above the posterior probability interquartile ranges of stages S3 and S4. According to the AASM sleep scoring rules, this stage is characterized by low-amplitude waves with mixed frequencies, which may not be ideal in distinguishing some features such as the band power. Nonetheless, this stage is the only where REMs occur and where both low-amplitude waves on the EEG often co-occur with an increased physiological activity such as heart rate, which may have resulted in generally high classification performances.

Then, stage S4 was the third best one at being correctly classified when using the training data. As with stage R, the proportions of S4 epochs in the testing data are not discrepant from the training data, and the sensitivities for testing were also relatively high and consistent, except for preprocessing with ICA together with filtering. The main difference found was that posterior probabilities were rather low compared to other sleep stages for all preprocessing settings.

Stage W yielded the fourth best classification performance considering the training data, and the highest sensitivity achieved (roughly 1.0 for ICA together with filtering) which may have resulted from the overrepresentation of this stage in the training data compared to the testing data. In other words, because sensitivity measures the rate of true positives among all epochs predicted with the considered stage, then because there are only 9 W epochs present in the testing data, the trained model predicted all of them correctly. However, this measure does not take into account false positives, which may have greatly impacted the performance (including sensitivities) of the other stages. Posterior class probabilities for this stage were relatively high for no preprocessing and slightly lower but very robust for ICA together filtering, meaning feature information from the testing data led to high and a consistent "judgment" of correctly labeled epochs, respectively. Again, it does not imply that these probabilities were much lower for mislabelled epochs. Nonetheless, taking into account that some features of stage W tend to be very unique (i.e. presence of eye blinking and increased heart-rate), we can hypothesize that this clear distinction may have led to improved classification performances for all preprocessing settings. Moreover, the reason why using ICA together with filtering led to increased sensitivities, whereas using filtering alone led to the lowest sensitivity is unclear and could be assessed by checking the impact of ICA on the features in this stage.

Stage S3 was the fifth best classified stage when using testing data. This stage also did not see a discrepancy in S3 epochs between training and testing data, but yielded generally lower sensitivities compared to stages S2, S4 and R. The sensitivities obtained for stage S3 were not the highest among all stages, and both filtering only and ICA together with filtering resulted in poor classification sensitivities. The posterior probabilities for this stage further support this discrepancy, suggesting that the trained models for no preprocessing and both ICA and filtering, may have underestimated how likely stage S3 have occurred for epochs which are true S3 epochs, given the new information of the testing data. The degree of underestimation could be better assessed by comparing these probabilities for epochs wrongly predicted as S3. A possible reason for the low sensitivities for this stage might be the fact that, according to the AASM sleep scoring, stage N3 (which has some degree of correspondence with stages S3 and S4 on the sleep scoring that was used in this work)

has no determinant features with regard to EMG and EOG, and EEG features are less specific compared to the scoring of other stages. It can be hypothesized that the low sensitivities for filtering and ICA together with filtering may have resulted from partially cutting off low frequency wave components from the EEGs, which may have helped the classifier distinguishing the stage when using preprocessing settings without filtering (no preprocessing and ICA only).

Finally, stage S1 yielded the lowest classification performance both when using training and testing data. For the latter, sensitivities were very low, under 0.1, even though no discrepancy in the number of S1 epochs was seen for the two datasets. It can be hypothesized that the alternation between stages W and N1 as an individual falls asleep may lead to a challenging distinction of both classes. This should however be assessed by checking the proportions of epochs of all stages other than S1 within the mislabeled ones.

Regarding the accuracy of the models, it is very similar across all models in the training phase and in the testing phase it diminishes as it was expected since it is tested on unseen data. In fact, the accuracy of the model when the signals are processed both with filters and with ICA decreases substantially. Given the high values of accuracy and Cohen's Kappa in the training phase, overfitting seems to be the reason of this disparity. In fact, this lower value of Cohen's Kappa is an indicator that even if the classifier correctly classifies a stage, it is probably a coincidence (happened by chance).

Feature Analysis

By analysing the distribution of features per sleep stages, it was possible to infer how much each feature contributed to the classification. Features such as kurtosis, k-complexes and Hjorth activity proved to be useful in separating stage W from the others, while skewness did not add any value since its value remained very similar across all stages.

Moreover, the EEG signal corresponding to C4A1 is particularly useful in identifying sleep spindles, which in turn means this feature is better to classify than sleep spindles obtained using other channels [32]. Furthermore, a similar situation occurred with the k-complexes, in which the results showed more variability between stages when using the C4A1 channel. This suggests that using a feature matrix using only sleep spindles and k-complexes from this channel could reduce the complexity of the models without compromising accuracy.

During the awake periods, the movement of the patient increases unlike during sleep when the patient is still. As a result, EMG related features are useful to distinguish between awake and asleep phases of sleep.

Analysing the results regarding the EOG features, the number obtained is higher for the wake stage, but there were some blinks detected during other phases of sleep, which is not realistic. Hence, blink detection might not be done correctly. Nevertheless, the number of detected blinks tends to be higher in the awake stage, which is expected, so it is a useful feature to separate this stage from the others.

The graph in Figure 12 suggests that using less features results in an increase in sensitivity in some stages, not compromising the others. Moreover, for stages W and S4 the increase matches the highest value obtained by the other models.

Additionally, the accuracy and Cohen's Kappa (CK) of the present model and the corresponding one using all features were compared. It is possible to see that while the training accuracy and CK remain very similar, reducing the number of features increases the accuracy during testing as well as CK, suggesting the model with less features is not only simpler but also provides better results.

In the end, it is valid to say that reducing the features resulted in an overall better model, since it is less complex, has better sensitivity in most stages, and reduced randomness during classification, all without compromising accuracy.

Main conclusions

In this paper, four methods of preprocessing are compared in the classification of sleep stages. Moreover, a selective set of features is then selected to determine if the number of features affects the classification.

It is shown that, when training, all types of preprocessing are rather similar. On the contrary, when testing with unseen data, applying just ICA increases slightly the accuracy and the Cohen's Kappa of the classification. By using just filters, the accuracy and Cohen's Kappa decrease but not in a considerable way (less than 10%).

However, in testing, the model where ICA and filters are applied is significantly worse: the accuracy drops to about 60% and the Cohen's Kappa also decreases to 0.428. This means not only that the model is worse in terms of accuracy, but also that even in the cases that it correctly classifies a stage that is probably a coincidence (due to chance).

As such, a model that had the same type of preprocessing method but fewer features was designed and it was determined that with fewer features it yielded better results (more accurate and greater Cohen's Kappa) although it does not reach the performance of the other models.

This type of feature analysis revealed to be of great value since not only it improved the results but also because it simplified the model. Hence, it can also be concluded that it is important to analyse the distribution of each feature throughout the stages to see which ones really are and should be a part of the decision making process.

Future Work

One thing to keep in mind is that in this study it was used 6 sleep stages as it was the number of stages in the txt files, thereby following R&K sleep scoring. However, there are many studies that use a 5 sleep stage classification by merging S4 with S3, hence following the AASM sleep scoring which is more recent. This merger has the potential of increasing accuracy of the classifier [33][34]. Another way of increasing the accuracy, would be to perform a wider study with more subjects/patients or more recordings.

In addition, as several articles cited mention, more complex machine algorithms can be used such as convolutional neural networks. Another way of increasing the complexity and number of the feature extraction is to use Empirical Mode Decomposition (EMD) and calculate features specific to each Intrinsic Mode Function (IMF) or even calculate a Wavelet Decomposition and use the coefficients as features.

Moreover, a down-sampling of the signals to a smaller sampling frequency would be of benefit to save memory and to use less processing power. Nevertheless, it would be interesting to study how that would affect the classification, and to try and find a balance between speed, memory used and performance.

As for the classification performance of each sleep stage, more research needs to be done in order to address the impact of the preprocessing steps used in this pipeline on specific features, especially on the selected features that later proved to improve classification performance. A deeper analysis on the uniqueness of features per sleep stage should be also done, because scoring rules used for manual annotation are rather broad when translating them to a machine-learning based classifier.

Furthermore, other feature selection studies can be performed as a way of not only reducing the complexity of the models used but also reduce the number of signals acquired during sleep. During a PSG study, the patient needs to go through a full night of sleep with several electrodes connected to several parts of the body, all the while trying to sleep as normally and comfortably as possible. This means that the final PSG results will in some way be affected by the measurements, resulting in a unrealistic view of the patients normal sleep. Hence, if it was possible to use for example only EEG signals to study the sleep, the patient's comfort would increase, and perhaps the results would be more reliable.

REFERENCES

[1] M. J. Aminoff, F. Boller, and D. F. Swaab, "Foreword," in *Sleep Disorders Part I*, ser. Handbook of Clinical Neurology, P. Montagna and S. Chokroverty, Eds. Elsevier, 2011, vol. 98, p. vii. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444520067000472

[2] S. Chokroverty, "Overview of sleep & sleep disorders." *The Indian journal of medical research*, vol. 131, pp. 126–40, 2010.

[3] N. Watson, M. Badr, G. Belenky *et al.*, "Joint consensus statement of the american academy of sleep medicine and sleep research society on the recommended amount of sleep for a healthy adult: Methodology and discussion." *Journal of Clinical Sleep Medicine*, vol. 11, pp. 931–952, 2015.

[4] X. L. Xi *et al.*, "Video-Based Actigraphy for Monitoring Wake and Sleep in Healthy Infants: A Laboratory Study." *Sensors (Basel, Switzerland)*, vol. 19, p. 1075, 2019.

[5] R. Malhotra and A. Avidan, *Sleep Stages and Scoring Technique*, 12 2014, pp. 77–99.

[6] D. Moser, P. Anderer, G. Gruber *et al.*, "Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters," *Sleep*, vol. 32, pp. 139–149, 2021.

[7] S. Deng, X. Zhang, Y. Zhang, H. Gao, E. I.-C. Chang, Y. Fan, and Y. Xu, "Interrater agreement between american and chinese sleep centers according to the 2014 aasm standard," *Sleep and Breathing*, vol. 23, no. 2, pp. 719–728, Jun 2019. [Online]. Available: https://doi.org/10.1007/s11325-019-01801-x

[8] A. Guillot, F. Sauvet, E. H. During, and V. Thorey, "Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 9, pp. 1955–1965, 2020.

[9] S. Raiesdana, "Automated sleep staging of OSAs based on ICA pre-processing and consolidation of temporal correlations," *Australasian Physical & Engineering Sciences in Medicine*, vol. 41, pp. 161–176, 2018.

[10] I. Winkler, S. Debener, K.-R. Müller, and M. Tangermann, "On the influence of high-pass filtering on ica-based artifact reduction in eeg-erp," pp. 4101–4105, 2015.

[11] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[12] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J.-F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel eeg," *Biomedical Signal Processing and Control*, vol. 42, pp. 107–114, 2018.

[13] J. Zhang and Y. Wu, "A new method for automatic sleep stage classification," *IEEE transactions on biomedical circuits and systems*, vol. 11, no. 5, pp. 1097–1110, 2017.

[14] C. Vural and M. Yildiz, "Determination of sleep stage separation ability of features extracted from eeg signals using principle component analysis," *Journal of Medical Systems*, vol. 34, no. 1, pp. 83–89, Feb 2010. [Online]. Available: https://doi.org/10.1007/s10916-008-9218-9

[15] C.-S. Huang, C.-L. Lin, L.-W. Ko, S.-Y. Liu, T.-P. Su, and C.-T. Lin, "Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels," *Frontiers in Neuroscience*, vol. 8, p. 263, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2014.00263

[16] T. Nakamura, T. Adjei, Y. Alqurashi, D. Looney, M. J. Morrell, and D. P. Mandic, "Complexity science for sleep stage classification from eeg," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 4387–4394.

[17] J. L. Rodríguez-Sotelo, A. Osorio-Forero, A. Jiménez-Rodríguez, D. Cuesta-Frau, E. Cirugeda-Roldán, and D. Peluffo, "Automatic sleep stages classification using eeg entropy features and unsupervised pattern analysis techniques," *Entropy*, vol. 16, no. 12, pp. 6573–6589, 2014. [Online]. Available: https://www.mdpi.com/1099-4300/16/12/6573

[18] R. Acharya, O. Faust, N. Kannathal, T. Chua, and S. Laxminarayan, "Non-linear analysis of eeg signals at various sleep stages," *Computer methods and programs in biomedicine*, vol. 80, no. 1, pp. 37–45, 2005.

[19] J. Fell, J. Röschke, K. Mann, and C. Schäffner, "Discrimination of sleep stages: a comparison between spectral and nonlinear eeg measures," *Electroencephalography and clinical Neurophysiology*, vol. 98, no. 5, pp. 401–410, 1996.

[20] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *Physiological Measurement*, vol. 36, no. 10, pp. 2027–2040, aug 2015. [Online]. Available: https://doi.org/10.1088/0967-3334/36/10/2027

[21] S. A. Imtiaz and E. Rodriguez-Villegas, "Automatic sleep staging using state machine-controlled decision trees," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 378–381.

[22] O. Yildirim, U. B. Baloglu, and U. R. Acharya, "A deep learning model for automated sleep stages classification using psg signals," *International Journal of Environmental Research and Public Health*, vol. 16, no. 4, 2019. [Online]. Available: https://www.mdpi.com/1660-4601/16/4/599

[23] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Annals of Biomedical Engineering*, vol. 44, no. 5, pp. 1587–1597, May 2016. [Online]. Available: https://doi.org/10.1007/s10439-015-1444-y

[24] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[25] J. Cháberová, "Analysis of sleep polysomnography data using advanced signal processing algorithms," Ph.D. dissertation, 2017. [Online]. Available: https://dspace.cvut.cz/bitstream/handle/10467/67318/F3-DP-2016-Chaberova-Jana-Analysis%20of%20Sleep%20Polysomnography%20Data%20Using%20Advanced%20Signal%20Processing%20Algorithms.pdf

[26] W. Rose, "Electromyogram analysis," *Mathematics and Signal Processing for Biomechanics*, 2019.

[27] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Pearson/Prentice Hall, 2005.

[28] B. Şen and M. Peker, "Novel approaches for automated epileptic diagnosis using fcbf selection and classification algorithms," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 21, no. Sup. 1, pp. 2092–2109, 2013.

[29] Y. Zhang, B. Wang, J. Jin, J. Zhang, J. Zou, and M. Nakamura, "A comparison study on multidomain eeg features for sleep stage classification," *Computational Intelligence and Neuroscience*, vol. 2017, 10 2017.

[30] P. Anderer, G. Gruber, S. Parapatics, M. Woertz, T. Miazhynskaia, G. Klösch, B. Saletu, J. Zeitlhofer, M. Barbanoj, H. Danker-Hopfe, S.-L. Himanen, B. Kemp, T. Penzel, M. Grozinger, D. Kunz, P. Rappelsberger, A. Schlögl, and G. Dorffner, "An e-health solution for automatic sleep classification according to rechtschaffen and kales: Validation study of the somnolyzer 24 × 7 utilizing the siesta database," *Neuropsychobiology*, vol. 51, pp. 115–33, 02 2005.

[31] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[32] C. Yücelbaş, Ş. Yücelbaş, S. Özşen, G. Tezel, S. Küççüktürk, and Ş. Yosunkaya, "Detection of sleep spindles in sleep eeg by using the psd methods," 2016.

[33] P. Van Hese, W. Philips, J. De Koninck, R. Van de Walle, and I. Lemahieu, "Automatic detection of sleep stages using the eeg," in *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, 2001, pp. 1944–1947 vol.2.

[34] O. Yildirim, U. B. Baloglu, and U. R. Acharya, "A deep learning model for automated sleep stages classification using psg signals," *International Journal of Environmental Research and Public Health*, vol. 16, no. 4, 2019. [Online]. Available: https://www.mdpi.com/1660-4601/16/4/599