Signal Processing and Data Analytics in Biomedical Engineering
Feature extraction for sleep scoring

KTH Royal Institute of Technology
School of Engineering Sciences in Chemistry, Biotechnology and Health

Elin Edblom
Salvar A. Jóhannsson
Vittorio Spadolini

Introduction

This project will focus on the study of various biological signals for the purpose of sleep scoring. The signals are included in 10 EDF (European Data Format) files, which are taken from the Sleep Heart Health Study (SHHS) The Sleep Heart Health Study (SHHS) [1]. The signals are analyzed along with any artifacts and noise they may include, such as movement artifacts and powerline noise. A segmentation process provided the division of the signals into epochs, where each epoch lasted 30 seconds. Proper signal processing methods are applied in order to remove disturbs from the signal and extract both frequency and time features. The features are then compared statistically to the different sleep stages. The features and signals are used to create a dataset to implement a machine learning (ML) approach for detecting sleep stages, this can be used to evaluate and improve sleep for clinical and consumer applications [2][3].

The dataset used in this project is taken from 10 different patients, containing different measurements such as: Heart rate (HR) [BPM], Oxygen saturation ($Sa02$) [%], Electrooculogram (EOG) [μV], Electromyography (EMG) [μV], Electroencephalogram (EEG) [μV], Body position and respiratory signals.

The dataset also includes the sleep stages of the patients. The sleep stages are split into 5 different categories. Descending order from awake to deep sleep:

- Wake
- Sleep stage 1 (N1)
- Sleep stage 2 (N2)
- Sleep stage 3 (N3)
- Sleep stage 4 (N4)
- REM (Rapid Eye Movements)

The signals

The project does not make use of all the signals included in the dataset, but is rather focused on select biological signals that are expected to characterise the different sleep stages, in addition, the American Academy of Sleep Medicine (AASM) guidelines are used as the main reference to detect the sleep stages [4]. This project focuses more in detail on EEG and EOG.

EEG is the method used to detect the electrical activity of the brain. EEG is widely used in sleep research. During EEG measurement, electrodes are usually placed on the scalp according to an international 10-20 system. This tool is often used to detect important features from brain activity, these can distinguish the sleep stages as highlighted in previous literature [5].

EOG measures the standing potential of the eye's cornea, relative to the posterior of the eye [6]. The dataset contains EOG measurements for both the left (EOL) and right (EOR) eye. It is expected to be helpful to detect sleep stages due to the characteristic of the REM sleep stage, after which it is named, of rapid eye movements. It could also be helpful for detecting wakefulness once blink detection algorithms have been implemented.

Method

In order to analyse and process the signals fundamental steps have been followed in order to obtain appropriate results:

- **Pre-processing**, in this phase the signals are prepared and cleaned to remove all possible undesirable information. The artefacts included can be respiration and muscular movements, electrode displacement or poor positioning. In addition, all the signals were filtered with a bandpass filter to remove powerline noise at 60 Hz.

- **Feature extraction**, this phase aim to extract the relevant information from the data in order to distinguish particular patterns and the different sleep stages. It can be conducted both in time (temporal features) and frequency domain (spectral features). For the purpose of extracting spectral features two method are proposed, one parametric and one non-parametric. According to the AASM manual, multiple features are of interest when evaluating sleep stages. A more comprehensive description of this stage is presented in the Feature Extraction section.

- **Feature classification**, the last phase involves the use of a ML approach to make a scorer that is able to automatically classify sleep stages. The k-nearest neighbours algorithm also known as KNN was used, which is a supervised learning classifier that leverages the proximity of data points to classify or predict their grouping. It utilizes the distances between data points to determine the class or category to which an individual data point belongs. In order to train the learning algorithm (LA) the dataset was split into training (70%) and test (30%) sets.

Signal processing

The process began by studying the artifacts in the signal. This was done by plotting the signal, histograms, and the Fourier transform of the main signals. The sleep stages were represented by a integer from 0-5, where each integer represented a sleep stage or wakefulness. Most of the patients had only 4 sleep stages (N1, N2, N3, REM) and the wakefulness stage, there was one exception found in patient 2, which had an extra sleep stage (assumed to be N4). Patient 2 was also the only patient to begin asleep, and was only for the first third of the measurements and awake for the rest. The patients HR also dropped to zero around that time, which could mean, among other things, that the electrodes were improperly attached and fell off, problem with devices or they were removed. The signals themselves were sampled at different frequencies, with the EEG and EMG being sampled at 125 Hz and the EOG (EOG will be used to refer to both channels) at 50 Hz. The EOG data was zero padded in order to match the size of the signals with higher frequency. In order to account for that the first $Fs_{EOG} \times length(EOG)/Fs_{EEG/EMG}$ of the EOG signals were extracted.

In the frequency domain of the higher sampled EEG signals there was observed a spike at 60 Hz. This was most likely due to noise from powerlines which is at 50Hz or 60Hz depending on the location. This was not noticed in the lower sampling frequency EOG signal since it was not sampled at high enough rate to capture this noise, with a Nyquist frequency of only 22.5Hz. The EEG and EOG were filtered for this using a 2nd order lowpass Butterworth filter at 56Hz. Other filters were tried, such as FIR Equiripple and Chebyshev, but it was decided to use the Butterworth due to the flat passband of the Butterworth filter. Other artifacts that were observed were baseline fluctuations, this was particularly noticeable for the EEG of patient 2. Several methods were used for removing the baseline fluctuations, the first one tried was a simple linear detrending method, which removed the best fitting line from the signal, which did not work particularly well for patient 2 which did not have a linear baseline fluctuation. The other more effective methods for fixing the baseline were the use of a highpass Butterworth filter on the signal, and the subtraction of the lowpassed filtered signal from itself. The method which was used was the subtraction of the lowpassed filtered signal from itself was used as it didn't require a highpass filter on the signal itself which would introduce more phase shifting and possible loss of information. For good measure, unwanted trends or biases from the data were removed by linearly detrending each epoch, allowing for a clearer analysis of the underlying patterns or signals. Mean removal subtracted the mean value of the signal from each data point. The mean represents the central tendency of the EEG, and by removing it, any constant bias or offset in the data where eliminated. After an accurate analysis, it was decided to not normalize the signal, since some features required the voltage. This procedure was also performed for the EMG, with the exception of using a notched filter at 50Hz since the signal had a spike at 50Hz compared to the spike at 60Hz for the EEG. The EOG was treated in the same way, with the exception of filtering noise due to powerlines, since the sampling frequency of the signal made that unnecessary. A median absolute deviation (MAD) was considered to remove possible outliers, however, it was not applied because it could have removed important information.

Feature Extraction

The feature extraction allowed to reduce the amount of data into a limited number of features that are able to represent the relevant information; thus creating a group of input variables that can be efficiently inserted in an algorithm. When evaluating sleep stages, the AASM (American Academy of Sleep Medicine) manual highlights the following features as important considerations for EEG and EOG:

- Percent of Alpha Activity (8-13 Hz)
- Percent of Delta Activity (0-4 Hz)
- Percent of Slow waves
- Blinks
- Slow eye movement
- K-Complexes (0.5-3 Hz)

Some additional features were added to the previous list such as:

- Percent of Theta Activity (4-8 Hz)
- Percent of Beta Activity (13-30 Hz)
- Rapid eye movement
- Spectral mean
- Spectral variance
- Spectral flux: $SF = \sqrt{\sum[epoch(n) - epoch(n-1)]^2}$
- Spectral flatness: $SFM = \frac{\sqrt[n]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{0}^{N-1} x(n)}{N}}$
- Centroid: $C = \frac{\sum_{n=0}^{N-1} x(n) f(n)}{\sum_{n=0}^{N-1} x(n)}$
- Spectral Roll-Off

The preceding list of features is considered to represent the most relevant ones, however, a complete list used in the project can be found in the Appendix (see Features for full list).

All these features where chosen to assign class labels to the observed data. A brief description of each one will follow. The percent of each waves was calculated as the sum of spectral power at each frequency band and then divided by the total power. For Blinks, Slow eye movement, K-complexes, Rapid eye movement opportune algorithm was created. Spectral mean was computed by summing the product of each frequency component and its corresponding power value, and then dividing by the total power. Spectral variance was computed by calculating the variance of the frequencies around the spectral mean. These measures, spectral mean and spectral variance, provided summary statistics that helped describing the central tendency and spread of power in signal's frequency spectrum. Spectral flux is defined as the change in the spectrum between consecutive windows using the Euclidian distance. Spectral flatness quantifies the relative distribution of power across frequencies in the signal's power spectrum. It provided information about the shape of the spectral content. The centroid can be considered as the "center of mass" of the frequency distribution. Spectral roll (SR) off is defined as the frequency where 85% of the energy in the spectrum is below this point.

For the power spectrum estimation, two methods were evaluated.

1. **Parametric method**: assume a specific mathematical model for the signal's spectral. In the project, it was selected an autoregressive (AR) model combined with the Burg method, where a window of 1 second and 50% overlapping sequences was used. The decision was supported by the fact that in EEG the observations are highly correlated with the past values.

2. **Non-parametric method**: do not assume a specific model for the signal. Instead, it estimates the power spectrum directly from the data without assuming a specific structure. In the project, it was used Welch's method combined with a 1 second Hamming window with a 50% overlap. This method with the relative window permitted overlapping sequences and a good trade-off between spectral resolution and spectral masking [7].

Statistics

When machine learning is developed, it is important to investigate which parameters affect the trained model. When tracking sleep, it is fundamental to know which different features are important for the different stages of sleep. Using a statistical analysis we found those that are crucial to distinguish different stages and which ones may be in conflict with each other due to correlation. This procedure is called dimensionality reduction and it is conducted to create a model that is the most representative. Moreover, a high number of extracted features can lead to an overfitted model, meaning the model has learned the training data too well. To determine which features should be included or excluded, tests with both ANOVA and Kruskal Wallis have been done, these provide insight to which feature the model is sensible too. The features that are shown to be least important for the detection of different stages have been excluded. Wrapper based feature selection was then done with the remaining features. A forward selection was done by adding the features one by one and evaluate if the feature made an improvement on the model. This was done to further develop the model. In this project, a statistical analysis has been carried out to distinguish which features are most important, in addition, mean, standard deviation and variance were applied as statistical descriptors to summarize and outline the data.

The selected features:

- KComplex,
- EEG variance,
- EEG SFM,
- EEG centroid,
- EEG SR,
- EEG SF,
- EEG power variance
- beta mean,
- beta percent,
- beta variance,
- theta mean,
- theta percent,
- theta variance,
- alpha mean,
- alpha percent,
- alpha variance,
- delta mean,
- delta percent,
- delta variance,
- difference (left and right EOG) variance,
- difference mean,
- rapid eye movement,
- blinks,
- slow wave percent.

Machine Learning

When it comes to ML, there is the necessity to choose the way how the machine will learn. There are three main options:

1. Unsupervised learning where the algorithm learns patterns and structures from unlabeled data, there are no predefined feedback or outputs in the training data. The algorithm aims to discover hidden patterns, relationships, or clusters within the data.

2. Supervised learning where an algorithm learns from labeled training data, in order to make predictions or classify new data points.

3. Reinforcement learning focuses on an agent learning how to make sequential decisions in an environment to maximize cumulative rewards. The agent learns through a trial-and-error process, interacting with the environment and receiving feedback in the form of rewards or penalties.

Each of these learning approaches has its own strengths and applications. Supervised learning is useful for tasks where labeled training data is available and prediction is the primary objective. Unsupervised learning helps uncover patterns and insights from unlabeled data. Reinforcement learning is suitable for sequential decision-making problems with a focus on optimizing long-term rewards. In the project, a particular supervised algorithm called KNN was used, due to its simplicity and non-parametric characteristic, meaning it does not make any assumption about the distribution of the data. Moreover, it was considered the most suitable since we had a large number of labelled data.

The ML approach produced important graphical metrics that were used to examine the results. The confusion matrix is a tabular representation that provides a comprehensive view of the performance of the KNN classification. The matrix is able to illustrate the percentage of the predicted stages against the true stages. A useful parameter that can be extracted from this is the accuracy of the classification, defined as the proportion of correct predicted stages among all the predictions made. However the classification accuracy can be better evaluated with another metric obtained, the receiver operating characteristic (ROC) curve, which illustrates the performance of the KNN classification model across various classification thresholds. It displays the trade-off between the true positive rate (TPR), also known as sensitivity or recall, and the false positive rate (FPR) at different threshold settings. The threshold is influenced by sensitivity and specificity, in particular, they have an inversely proportional relationship, therefore choosing a lower threshold brings an increase in sensitivity and a decrease in specificity. On the other hand, choosing a higher threshold brings a decrease in sensitivity and an increase in specificity. Moreover, the area under the ROC curve (AUROC) is calculated as a single metric to summarize the overall performance of the model.

Results

The different sleep stages in the result are labelled as numbers: 0 is REM, 1 is N3, 2 is N2, 3 is N1 and 4 is awake. Figure 1 shows a confusion matrix of the model when all the 39 features were included. The accuracy of the model was 50.4%. The model used was weighted KNN.

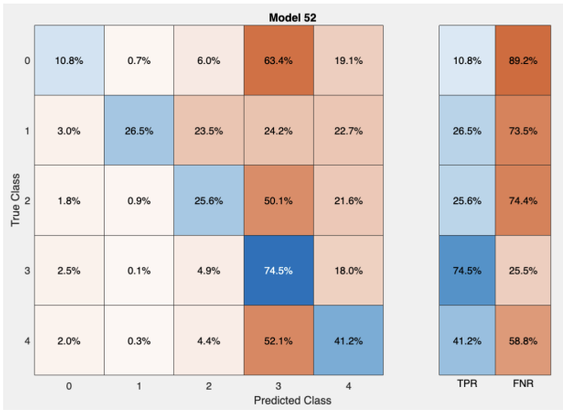

Figure 1: Confusion matrix using all the features

Figure 2 shows a confusion matrix of the model when 24 out of 39 features were included. The accuracy of the model was 72.2%. The model used was Subspace KNN. It was tested on 20% of the sleep dataset, and used 24 features.

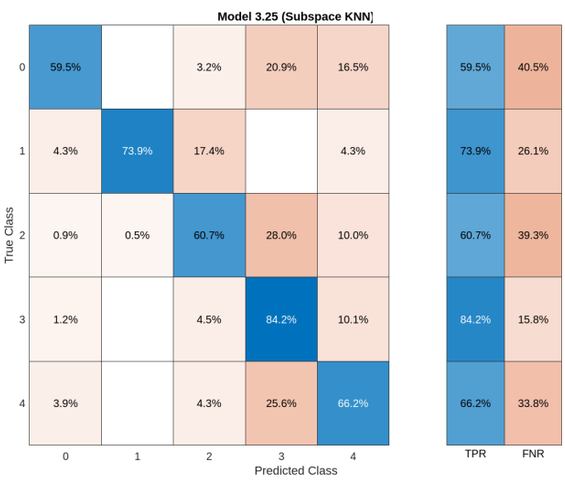

Figure 2: Confusion Matrix using 24 features, showing the overall better performance of this model.

Figure 3 presents the ROC curve for the model using all the features. It shows that N3 stage approaches the top left corner, suggesting high sensitivity and low false positive rate, with a high value of the $AUROC \simeq 0.88$. The other 4 stages (awake, N1 ,N2, REM) are not as good as the the previous, they are much far away from the top left corner and they all have an AUROC under 0.7 suggesting that the model's overall performance in distinguishing between these true or false stages is moderate.

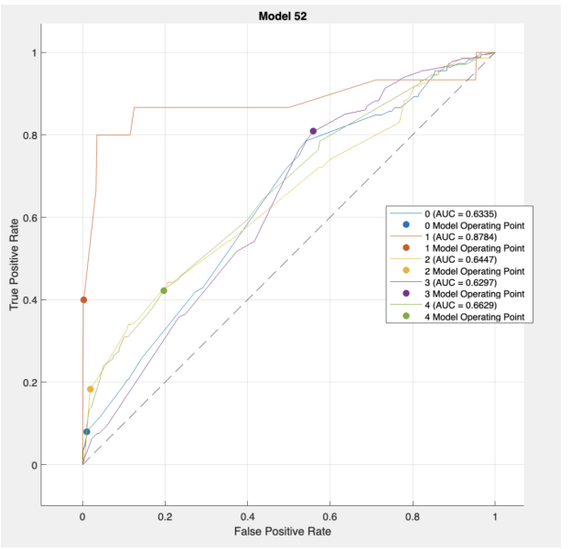

Figure 3: ROC curve using 39 features

Figure 4 present the ROC curve for the improved model which used fewer features. As the first model the N3 stage shows the highest sensitivity with a low positive rate. The AUROC is almost ideal $AUROC \simeq 0.99$, therefore the improved model can discriminate perfectly the N3 stage. Also REM stage has an optimal trend with $AUROC \simeq 0.92$. For N1, N2 and awake stages the curves are significantly better compared tho the first model, resulting in $0.86 \leq AUROC \leq 0.87$.

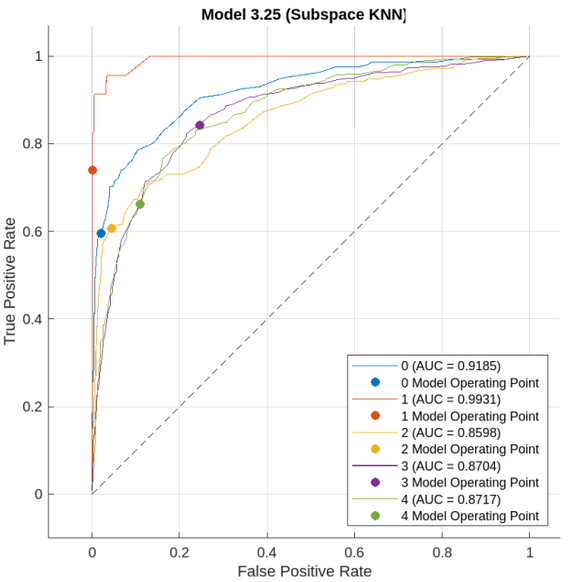

Figure 4: ROC curve using 24 features

Altogeher the improved model showed better results, including a better behaviouur in discriminating the sleep stages.

Discussion

The signals were studied for artifacts, and there were some inconsistencies that were interesting, such as the varying powerline frequency between the EEG and EMG. There were performed various signal preprocessing techniques and the ones that were considered to be more suited to the problem or provided better results according to a visual inspection.

The feature extraction was performed using methods found in the course material. While some of

the features had an easy to understand instructions for extracting from the signals, other were more unclear. While trying various methods for extracting the features, it proved to be difficult to know when the feature was correctly extracted. With a lack of a way to validate the results of features such as rapid eye movement, k-complexes and blinks, with data. The resulting features could therefore only be verified on the basis of the course material and could not be validated on anything but feeling. This caused it to be hard to know whether the feature that was being extracted was truly the one attempted to extract. This also meant that the instructions were followed blindly for things like thresholds which might be different depending on the signal.

The machine learning was performed alongside continuing efforts to fetch increasingly more features from the signals. Increasing the features turned out to be able to have a negative impact on the performance of the model, and the best performing model was one that had retroactively selected features. Attempts to improve the feature extraction seemed to improve the model performance in some situations, but increasing the features was not always a good thing.

The machine learning models that performed the best were good at distinguishing between REM, N3 and N2. The model does however not perform as well when identifying N1 and wake, while it has high True Positive Rate for both of them, it seems to wrongly classify other stages as NR1 and wake, quite often. This could be due to a bias in the dataset where these sleep stages are more common, or a lack of features that can accurately identify these stages. This might be solved with oversampling the less represented sleep stages, finding or fixing features that can help identify the stages. Another method that might improve a machine learning model is using autoregression since sleep seems to follow a pattern, so the previous predicted sleep stage might be a good indicator of the current stage. This would have to be implemented with precaution.

Conclusion

Using EEG, EOG and EMG features for identifying sleep stages with machine learning proved to be somewhat successful. There is however room for improvement, especially for preventing error and bias for NR1 and wake. Limiting the bias in the dataset by means such as oversampling and increase the amount of useful features by being able to validate the features extracted for improved feature extraction algorithm, could lead to significantly improved results.

Acknowledgements

The project was organized with different objectives in order to obtain the most effective and reliable results.

- Literature research and background
- MATLAB CODING
- Analysis of the result including validation and verification
- Report writing

Each team member had a valuable contribution to the project and report work.

- Salvar A. Jóhannsson: Worked on signal processing, feature extraction, assisted in writing report and presenting.
- Vittorio Spadolini: Literature research and background, assessed the project best approach, preprocessed the EEG, report creation and writing.
- Elin Edblom: Literature research, statistical analysis, selection of features, write report.

Inclusive communication encouraged us to assist each other whenever needed. We found that by working together and offering our expertise, we collectively surpassed challenges that arose during the project.

References

[1] National Sleep Research Resource. *Sleep Heart Health Study (SHHS)*. 2014. DOI: 10.25822/GHY8-KS59. URL: https://sleepdata.org/datasets/shhs.

[2] Gary Garcia-Molina, Farhad Abtahi, and Miguel Lagares-Lemos. "Automated NREM sleep staging using the Electro-oculogram: A pilot study". In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2012, pp. 2255–2258. DOI: 10.1109/EMBC.2012.6346411.

[3] Manish Sharma, Jainendra Tiwari, and U Rajendra Acharya. "Automatic sleep-stage scoring in healthy and sleep disorder patients using optimal wavelet filter bank technique with EEG signals". In: *International journal of environmental research and public health* 18.6 (2021), p. 3087.

[4] Richard B Berry et al. "The AASM manual for the scoring of sleep and associated events". In: *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* 176 (2012), p. 2012.

[5] A Roebuck et al. "A review of signals used in sleep analysis". In: *Physiological measurement* 35.1 (2013), R1.

[6] Donnell J. Creel. "The electrooculogram". In: *Clinical Neurophysiology: Basis and Technical Aspects*. Elsevier, 2019, pp. 495–499. DOI: 10.1016/b978-0-444-64032-1.00033-3. URL: https://doi.org/10.1016/b978-0-444-64032-1.00033-3.

[7] Mohd Ammar Bin Hayat and Md Belal Bin Heyat. "Hamming Window are used in the Prognostic of Insomnia". In: ().

Appendix

Features

- Blinks
- Alpha mean
- Alpha mean ratio
- Alpha variance
- Alpha variance ratio
- Beta mean
- Beta mean ratio
- Beta variance
- Beta variance ratio
- Delta mean
- Delta mean ratio
- Delta variance
- Delta variance ratio
- EEG mean
- EEG variance
- Theta mean
- Theta mean ratio
- Theta variance
- Theta variance ratio
- Alpha Delta
- Alpha Slow
- Alpha percent
- Beta percent
- Delta Slow
- Delta percent
- EEG Power mean
- EEG Power total
- EEG Power variance
- EEG Spectral lux
- EEG Spectral Flatness
- EEG Spectral Roll-Off
- EEG centroid
- Slow percent
- Theta percent
- EMG Power mean
- EMG Power total
- EMG Power variance
- EMG Spectral lux
- EMG Spectral Flatness
- EMG Spectral Roll-Off
- EMG centroid
- EMG High percent
- EMG Low percent
- High percent / Low percent
- EMG mean
- EMG skew
- EMG variance
- Left High percent
- Left Low percent
- Left Power mean
- Left Power total
- Left Power variance
- Left Spectral lux
- Left Spectral Flatness
- Left Spectral Roll-Off
- Left centroid
- Right High percent
- Right Low percent
- Right Power mean
- Right Power total
- Right Power variance
- Right Spectral lux
- Right Spectral Flatness
- Right S
- Right centroid
- Left Right high
- Left Right low
- Difference (EOGL & EOGR) mean
- Difference (EOGL & EOGR) variance
- Left Right mean
- Left Right variance
- Left mean
- Left variance
- Right mean
- Right variance
- K complex
- Rapid eye movement