Kungliga Tekniska Högskolan

CM2013
Signal Processing and Data Analytics in
Biomedical Engineering

EEG-Based Sleep Staging Project

Students:
François Dejaegere
Teodor Pstrusinski
Lasse Stahnke

Teacher:
Farhad Abtahi

June 7, 2022

Table of Contents

1 Introduction

Sleep is important part of every individual's life. Although all the underlying mechanisms of sleep are not fully understood, the effects of sleep deprivation or sleep disorders have been largely studied. Sleep disorders, such as sleep apnea can have an impact on a person's sleep, and hereby affect their quality of life.

In order to asses sleep, it can be divided into different stages. The conventionally used stages are wake, N1, N2, N3, and R [1]. Stages N1 to N3 correspond to non-rapid eye movement speed stages, and each of the stages is considered to be a deeper sleep [1]. REM sleep (R) is the sleep stage associated to dreaming, with activity being similar to the one of an awake person [1].

This project aims at obtaining a sleep staging classifier based on machine learning algorithms. Data from different modalities is first analysed and cleaned in order to be used for extracting the features used in the algorithms. The project therefore covers the full scope from data analysis, data cleaning and filtering, and lastly implementation and evaluation of machine learning algorithms.

2 Methods

In this section, the methods used for the project are described. First, the dataset used is detailed and the way the it was split into train, validation, and test sets is explained. Next, the pre-processing of some of the data is described, and features extraction is also presented. Further, the different classification algorithms are detailed and lastly the evaluation metrics used to assess the performance of those algorithms are presented.

2.1 Dataset

The present dataset includes sleep data from 10 different patients. The data is part of the Sleep Heart Health Study (SHHS) [2, 3]. For each patient the dataset contains different modalities. These modalities include two different electroencephalogram (EEG) channels, two electrooculography (EOG) channels (left and right eye), electrocardiogram (ECG), arterial blood saturation (SaO2), heart rate (HR), electromyogram (EMG), abdomen and thorax movement, location of the extremities and sensor data to sense external light. The data was sampled with frequencies ranging from 1 Hz (e.g. HR) to 125 Hz (e.g. EEG). For EEG, ECG, EOG and EMG a hardware implementation of a high-pass was used in the study, thus there is no need to further filter out baseline wander or a DC offset voltage.

The sleep data can be divided into epochs. The length of one epoch is 30 seconds. For each epoch, except for the last one, the sleep-stage has been determined. The provided dataset has been labelled using 6-classes: Wake, N1, N2, N3, N4, REM. In order to comply with recommendations of American Academy of Sleep Medicine (AASM) [4], stages N4 and N3 have been merged together into one stage N3 resulting in a 5-class dataset. These annotations have been used to train the methods to determine the sleep stage for unseen epochs.

Table 1: Distribution of sleep stages in train and test dataset. The values in parentheses denote the values in MATLAB. The number of training samples are given in absolute and relative values for the train and test dataset. It can be seen that the sleep stages are unevenly distributed, within the datasets. The relative number of epochs in a given stage are approximately the same for the train and test dataset.

		Train		**Test**	
Stage		abs	rel [%]	abs	rel [%]
REM	(0)	925	11.38	193	8.9
N3	(2)	1374	16.91	244	11.3
N2	(3)	2843	34.98	965	44.6
N1	(4)	186	2.29	89	4.1
WAKE	(5)	2799	34.44	675	31.2
Sum:		8127		2166	

2.1.1 Train and Test Data

To compare different methods in the present project, the data has been splint into train and test data. For the test data, two patients (patients 1 and 5) have been randomly drawn.

In order to avoid a bias in the estimation of the performance of the different methods on unseen data, the test data has not been used for hyperparameter-optimization nor for parameter-optimization. The test set should represent the unseen data as well as possible. Thus, having epochs from multiple patients in the test set would be ideal. However, since the features for determining the sleep stages are likely to have patient specific properties, it must be ensured that the patients in the test set are not included in hyperparameter- or parameter-optimization. Consequently, it was decided that only 2 patients are included in the test set.

The train data was used in combination with 5-fold cross validation to determine hyperparameters of the models. Once all hyperparameters for the models had been determined, the models were trained on the whole train dataset and tested using the test set.

In total, there are 8127 epochs (79%) and 2166 epochs (21%) in the train and test set, respectively. The distribution of sleep stages in the datasets can be seen in Table 1

2.2 Feature Extraction

Different features were extracted from the different signals provided for each epoch. Depending on the modality, different features could be added. Indeed, temporal features were always extracted for any modality, but frequency features and Hjorth parameters were only extracted for EEG data. The following subsections explain how these features were calculated from the dataset.

2.2.1 Temporal Features

Five temporal features were extracted: the signal mean, variance, amplitude, skewness and kurtosis.

The mean μ and variance s^2 of the signal were quite simply extracted by obtaining the mean and the variance for each epoch. The signal amplitude A corresponds to the maximal value of the signal for each epoch.

The kurtosis and skewness of the signal were also extracted for each epoch. Skewness is a measure of lack of symmetry of the data, and is defined in Equation 1 [5], where s is the standard deviation of the signal on the epoch.

$$\text{skewness} = \frac{\Sigma_{i=1}^{N}(y_i - \mu)^3}{Ns^3} \tag{1}$$

Kurtosis measures whether the data is heavily- or light-tailed relative to a normal distribution [5]. If kurtosis is high, this means that the data is more spread than a normal distribution with the same mean and standard deviation [5]. Kurtosis is defined in Equation 2 [5].

$$\text{kurtosis} = \frac{\Sigma_{i=1}^{N}(y_i - \mu)^4}{Ns^4} \tag{2}$$

To better understand skewness and kurtosis, Figure 1 gives an illustration of these two concepts.

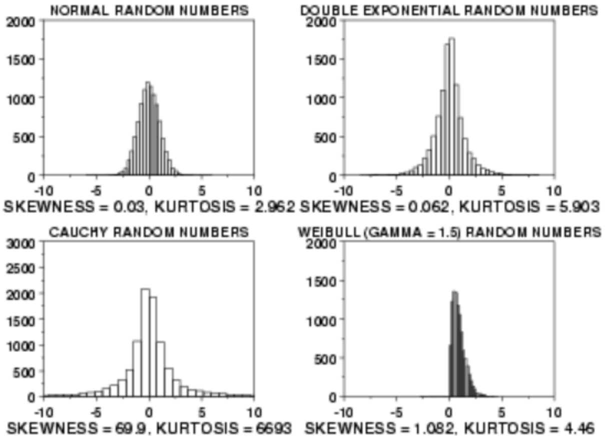


Figure 1: Illustration of skewness and kurtosis of four different distributions. Skewness increases as the dataset loses its symmetry and kurtosis increases as the dataset distribution becomes more heavily-tailed.

2.2.2 Frequency Features

In order to separate out important EEG signal frequency bands (Delta, Theta, Alpha, Beta), a 5 level discrete wavelet decomposition has been carried out using Daubechies-8(db8) mother wavelet. Further, 4 single branches of the signal corresponding to previously mentioned frequency bands have been reconstructed from 1-D wavelet coefficients.

For every epoch from each frequency band, signal energy E_s has been calculated using equation 3.

$$E_s = \Sigma_{n=1}^{N} |x(n)|^2 \tag{3}$$

2.2.3 Hjorth Parameters

Hjorth parameters are features extracted that characterise EEG signals. The Hjorth parameters are activity, mobility and complexity, their respective definitions are given in equations 4, 5 and 6 [6].

$$activity(y(t)) = s^2(y(t)) \tag{4}$$

$$mobility(y(t)) = \sqrt{\frac{s^2(\frac{dy(t)}{dt})}{s^2(y(t))}} \tag{5}$$

$$complexity(y(t)) = \frac{mobility(\frac{dy(t)}{dt})}{mobility(y(t))} \tag{6}$$

As the Hjorth activity is equal to the variance of the signal, already extracted in the temporal features, it is not extracted again and only used to calculate the mobility and complexity of the signal.

2.3 Classification Algorithms

The next paragraphs introduce the three different types of classification algorithms selected for this project: support vector machine, fully connected neural network, and convolutional neural network.

2.3.1 Support Vector Machine

Support vector machines (SVM) are one of the models used in supervised machine learning. They rely of statistical theory to perform classification and regression. The main purpose of SVM is to divide data in the feature space with a surface (known as hyperplane) that would maximise the margin between the classes. In most real life cases our data is not linearly separable however using a kernel to transform the data into a higher dimension in which our data becomes linearly separable. Examples of the most used kernel functions are: linear kernel, polynomial kernel, gaussian kernel and sigmoid kernel. While SVMs are a good tool for solving binary classification problems they add complexity if used for multi-class problems what requires us to convert them into multiple binary classification problems. There are two algorithms allowing us to solve multiclass classification: "one versus one" and "one versus all". The first one constructs a hyperplane for each possible pair of classes while the latter one creates a i-th hyperplane between i-th class and rest of the data.

Multi-class SVM model selection

In order to determine what type of kernel should be used for classification of sleep stages based on features extracted from 2-channel EEG and EOG signals, multiple SVM models have been trained using different hyperparameters. Hyperarameters being manipulated were kernel functions and kernel scale. Principal component analysis has also been used to optimise the speed of the classification algorithm.

Most of the models have trouble correctly classifying the N1 stage what could be caused by a small number of epochs labelled as such in our dataset (Figure 1). Therefore, the SVM model using a cubic kernel was selected because it correctly classified the most N1 labels despite not having the highest validation accuracy (78.0% compared to 80.7% for a quadratic kernel and 78.2% for a Gaussian kernel with scale 5.7). Enabling PCA resulted in 17 out of 32 features explaining 95% variance.

2.3.2 Fully Connected Neural Network

Fully connected neural networks (FCNN) were used for classification of sleep stages. Such a network has a first layer that is connected to the network input, which are the features. Afterwards, each subsequent fully-connected layers has as input the output of the previous layer multiplied by a weight matrix. Finally, after the last layer, a softmax activation function is used to produce the FCNN's outputs which are the predicted labels. The default structure of a FCNN in MATLAB is given in Figure 2.

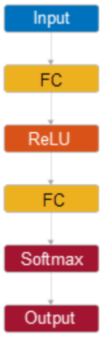


Figure 2: Structure of default MATLAB FCNN. Contains two fully-connected layers (FC), with a ReLU activation layer between both, and a Softmax activation layer after the second FC, giving the output.

The MATLAB *Optimizable Neural Network* function was used in order to obtain the best FCNN hyperparameters for a set of features. The hyperparameters optimised using this tool are:

- The number of fully connected layers: 1-3
- The activation functions between layers: ReLU, Tanh, Sigmoid, or None
- The size of each layer: 1-300

The optimizer options used were a Bayesian optimization (with acquisition function: *Expected improvement per second plus*), with 30 iterations and no time limit.

2.3.3 Convolutional Neural Network

In addition to the traditional machine learning methods, a deep learning method, in particular a Convolutional Neural Network (CNN) has been used to classify the sleep stages using the EEG and EOG data. CNNs have been proven to being able to outperform methods that require handcrafted features [7, 8]. Furthermore, the development of such an approach can be more efficient, since the network learns the features of interest itself and there is no need for a domain expert to actually design and choose those features.

It has been shown that the activity in the EEG is dependent on the stage of sleep that the patient is in [9]. Furthermore, the sleep stage 'rapid eye movement' is characterised by the movement of the eyes. Thus, both EEG and EOG channels have been used to train the CNN to classify sleep stages.

Architecture and Optimizer

A visualisation of the used CNN can be seen in Figure 3. The network contains of two 1-D Convolutional layers that are each followed by ReLu-activation functions, a MaxPooling layer and a Normalization layer. Subsequently, a Global-Averaging-Pooling layer is used in combination with a Fully-Connected layer and a Softmax-classifier to classify the sleep stages.

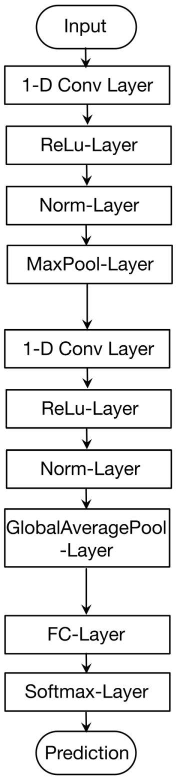


Figure 3: Visualization of the CNN that was used for the classification of sleep stages. It can be seen that the CNN consists of two 1D convolutional layers with ReLu-activation functions. Additionally a Normalization- and Max-Pooling-layer was used. Finally, after a global averaging layer, a fully-connected layer in combination with a Softmax-classifier is used to classify the sleep stages.

To train the network, the adam optimiser [10] was used with a minibatch-size of 50 and 100 epochs.

2.4 Evaluation Metrics

The upcoming paragraphs introduce the different metrics used to evaluate classification algorithms. First, the accuracy is presented, followed by the true positive rate and the positive predicted value.

2.4.1 Accuracy

To compare the performance of different methods, the accuracy metric was used. Accuracy is defined as the fraction of the number of correctly classified cases on the total number of cases. The formal definition can seen in Equation 7.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \begin{cases} \text{TP} = \text{True Positive} \\ \text{TN} = \text{True Negative} \\ \text{FP} = \text{False Positive} \\ \text{FN} = \text{False Negative} \end{cases} \tag{7}$$

2.4.2 True Positive Rate and Positive Predicted Value

The accuracy is strongly dependent on the distribution of the classes in the test dataset i.e. it increases if there are a lot of easy-to-classify cases (e.g. Wake stage) and decreases if there are more difficult-to-classify cases. Thus the accuracy is not sufficient to evaluate the performance of a method. Consequently, the True Positive Rate (TPR) and Positive Predicted Value (PPV) have been used additionally.

The True Positive Rate describes the rate of correct classifications of positive cases relative to the number of True Positive cases. It is defined as seen in Equation 8.

$$TPR = \frac{TP}{TP + FN}, \begin{cases} \text{TP} = \text{True Positive} \\ \text{FN} = \text{False Negative} \end{cases} \tag{8}$$

The Positive Prediction Value is a measure that compares the number of correctly classified cases within one class to the total number of predicted cases in that class. The definition for PPV can be seen in Equation 9.

$$PPV = \frac{TP}{TP + FP}, \begin{cases} \text{TP} = \text{True Positive} \\ \text{FP} = \text{False Positive} \end{cases} \tag{9}$$

3 Results

The classification algorithms presented in section 2.3 were trained and tested on multiple combinations of modalities present in the dataset. In this section, summary results of the classification algorithms on different sets of modalities are presented and detailed results are provided for the performances of the algorithms on the best-performing modalities set.

3.1 Comparison of Algorithms on Different Modalities Sets

Table 2 shows the obtained accuracy of the different algorithms used for different sets of modalities. It can be seen that the CNN outperforms the other methods for all set of modalities on the validation set. It can be also seen that, overall, the set including only EEG, EEGsec, EOGL, EOGR and EMG performed slightly better than the other sets.

Table 2: Validation accuracies of different algorithms on different sets of modalities.

		Val. Accuracy [%]		
Modalities	**N. Features**	**SVM**	**FCNN**	**CNN**
EEG	11	59.7	77.2	79.0
EEG, EEGsec	22	77.3	79.6	82.1
EEG, EEGsec, EOGL, EOGR	32	78.2	80.0	84.3
EEG, EEGsec, EOGL, EOGR, EMG	37	79.5	83.3	85.5
EEG, EEGsec, EOGL, EOGR, EMG, ECG	42	80.0	82.8	84.9

3.2 SVM

The confusion matrix for the SVM model using EEG and EOG with principal component analysis that has been described in Section 2.3.1 can be seen in Figure 4. It can be seen that the classification performance on the test set varies between different sleep stages. The True Positive Rate of the classes is 68.9%, 85.7%, 53.5%, 7.9% and 67.7% for REM, N3, N2, N1 and Wake, respectively. Overall the accuracy of the classification on the test set is 61%.

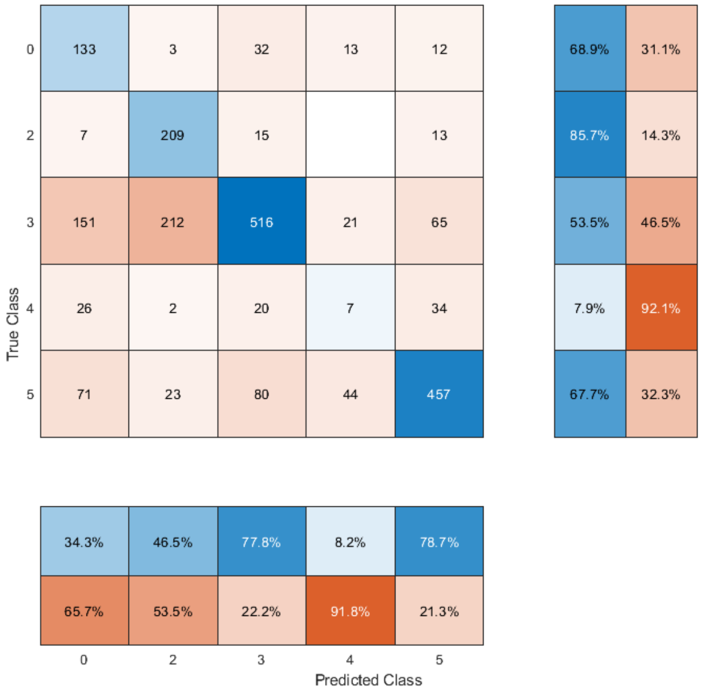
3.3 FCNN

FCNNs have been trained in MATLAB using the *optimizable neural network* feature of the machine learning toolbox. Different results were obtained using different sets of features. Indeed, it was observed that using the secondary EEG channel improved the accuracy, that using the EOG channels also improved the accuracy, that the ECG features did not improve the accuracy but that the EMG features did improve the accuracy, as can be observed in Table 2. Therefore, the selected FCNN model was the one with the highest accuracy on validation data (using 5-fold cross-validation). This model included the following modalities: EEG, EEGsec, EOGL, EOGR, EMG.

The hyperparameters of the obtained model were the following:

- Number of fully connected layers: 3;
- Size of layers: 17, 54, 293;
- Activation function: Sigmoid.

The model was trained with the training data, and the results of the predictions on the test data are presented in the confusion matrix of figure 5.



True Class	0	2	3	4	5		
0	133	3	32	13	12	68.9%	31.1%
2	7	209	15		13	85.7%	14.3%
3	151	212	516	21	65	53.5%	46.5%
4	26	2	20	7	34	7.9%	92.1%
5	71	23	80	44	457	67.7%	32.3%
	34.3%	46.5%	77.8%	8.2%	78.7%		
	65.7%	53.5%	22.2%	91.8%	21.3%		

Figure 4: Confusion matrix for a SVM that was trained to classify sleep stages in sleep data based on EEG and EOG data. The classes 0,2,3,4,5 correspond to the stages REM, N3, N2, N1 and Wake, respectively. Two bottom rows represent Positive Predictive Values (PPV) and False Discovery Rates (FDR) respectively. Two far right columns represent True Positive Rates (TPR) and False Positive Rates (FPR) respectively.

The obtained accuracy on the validation set was 83.3%, but the accuracy obtained on the test set is 70.2%. The TPR of the classes in the test set is 80.8%, 79.1%, 64.4%, 0.0%, and 81.6% for REM, N3, N2, N1 and Wake, respectively. It is observed that the FCNN has most difficulties with classifying the N1 sleep stage correctly. The stage with the best TPR is the wake stage, and the stage with the best PPV is also wake (82.4%).

Let us note that all results for the FCNNs were obtained without using a specific feature extraction or dimensionality reduction algorithm. Indeed, after observing for the same set of modalities and features that the results were poorer (lower overall accuracy) if PCA was used, it was decided to not use this dimensionality reduction method.

3.4 CNN

The confusion matrix for the trained CNN that has been described in Section 2.3.3 can be seen in Figure 6. It can be seen that the classification performance on the test set varies between different sleep stages. The True Positive Rate of the classes is 50.8%, 37.7%, 62.1%, 0% and 90.7% for REM, N3, N2, N1 and Wake, respectively. Overall the accuracy of the classification on the test set is 64.7%.

It is noteworthy, that the classification of the N1 stage is substantially lower than the

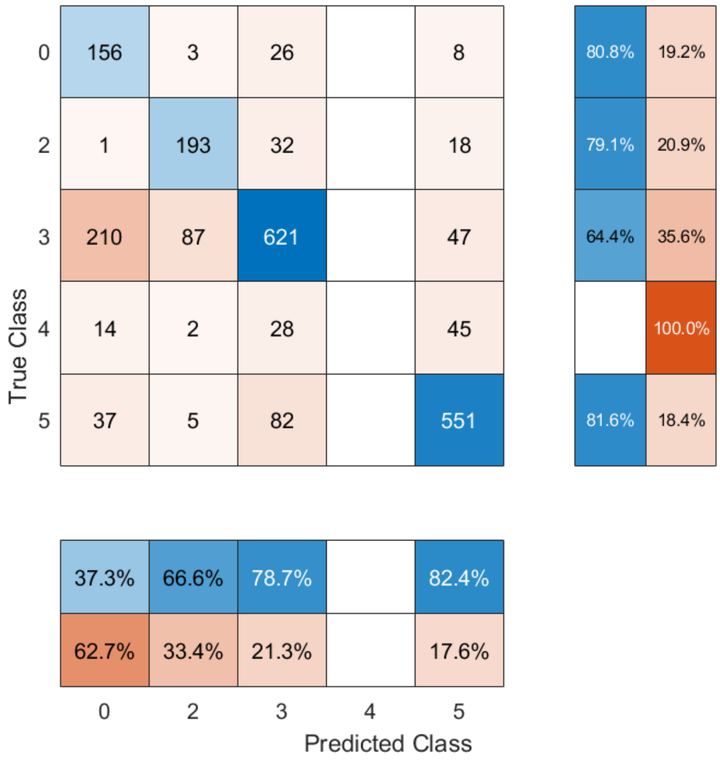


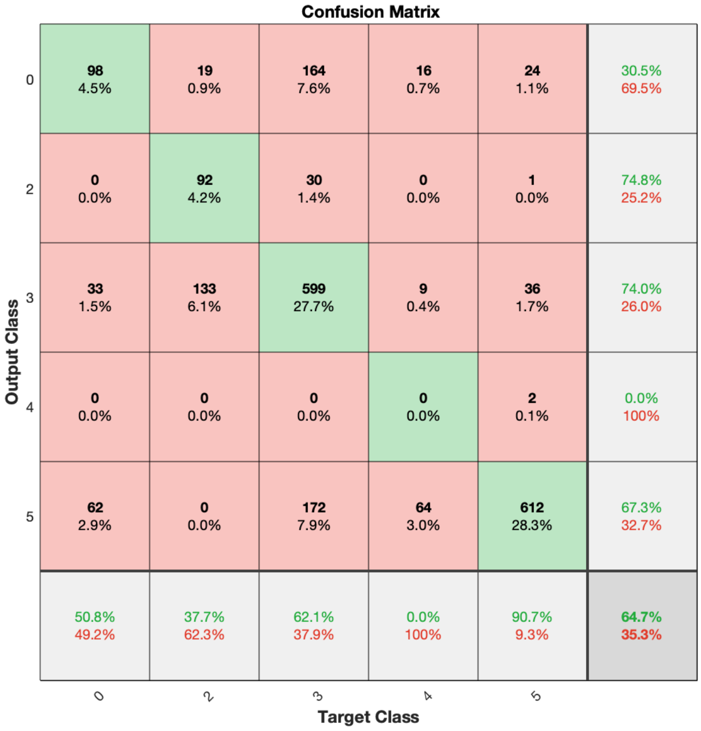
Figure 5: Confusion matrix for a FCNN that was trained to classify sleep stages in sleep data based on EEG, EOG, and EMG data. The classes 0,2,3,4,5 correspond to the stages REM, N3, N2, N1 and Wake, respectively. Two bottom rows represent Positive Predictive Values (PPV) and False Discovery Rates (FDR) respectively. Two far right columns represent True Positive Rates (TPR) and False Positive Rates (FPR) respectively.

classification of other stages. Out of all cases that have been classified as N1 stage, 0% were correctly classified, thus the CNN failed to classify all N1 cases. It can be also seen that the sleeping stage WAKE achieved the best TPR, while the N3 stage achieved the best PPV.

In Table 3, a comparison between the classification performance of the CNN in the train and test set can be seen. It can be seen that the classification performance is substantially better and that the network fails to classify the epochs in sleep stage N1 in the test set. This large difference suggests overfitting of the model.

3.5 Comparison of Methods

In Table 4 a comparison between all previously proposed methods can be seen. It can be seen that in comparison to the other methods, the FCNN has the best classification result for REM, N3 and N2. SVM achieved the best result in N1 classification compared to the other methods. However, less only 7.9% of N1 epochs have been correctly classified. The best result for the Wake phase was achieved by the CNN.


Output Class
Target Class

Confusion Matrix

	0	2	3	4	5	
0	98 4.5%	19 0.9%	164 7.6%	16 0.7%	24 1.1%	30.5% 69.5%
2	0 0.0%	92 4.2%	30 1.4%	0 0.0%	1 0.0%	74.8% 25.2%
3	33 1.5%	133 6.1%	599 27.7%	9 0.4%	36 1.7%	74.0% 26.0%
4	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.1%	0.0% 100%
5	62 2.9%	0 0.0%	172 7.9%	64 3.0%	612 28.3%	67.3% 32.7%
	50.8% 49.2%	37.7% 62.3%	62.1% 37.9%	0.0% 100%	90.7% 9.3%	64.7% 35.3%

Figure 6: Confusion matrix for a CNN that was trained to classify sleep stages in sleep data based on EEG and EOG data. The classes 0,2,3,4,5 correspond to the stages REM, N3, N2, N1 and Wake, respectively. It can be seen that the accuracy between stages varies substantially. The classification performance of stage N1 is the worst, while the performance of detecting stage N2 is the best.

It was found that the overall accuracy on the whole test set was 61%, 70.2% and 64.7% for the SVM, FCNN and CNN, respectively.

Table 3: Comparison of test and train classification performance of the CNN per sleep stage. It can be seen that the difference is substantial, which indicates overfitting. Furthermore it can be seen that some of the epochs with sleep stage N1 are correctly classified in the training set, while none of the epochs with N1 are correctly classified in the test set.

	CNN			
	Train		**Test**	
Stage	TPR	PPV	TPR	PPV
REM	0.874	0.870	0.508	0.305
N3	0.890	0.866	0.377	0.748
N2	0.880	0.866	0.621	0.740
N1	0.161	0.500	0.000	0.000
Wake	0.935	0.922	0.907	0.673

Table 4: Comparison of TPR and PPV per sleep stage across different classification methods i.e. SVM, FCNN and CNN.

	Test Classification Performance					
	SVM		**FCNN**		**CNN**	
Stage	TPR	PPV	TPR	PPV	TPR	PPV
REM	0.689	0.343	0.808	0.373	0.508	0.305
N3	0.857	0.465	0.791	0.666	0.377	0.748
N2	0.535	0.778	0.644	0.787	0.621	0.740
N1	0.079	0.082	0.000	0.000	0.000	0.000
Wake	0.677	0.787	0.816	0.824	0.907	0.673

4 Discussion

4.1 Preprocessing and Baseline Wander

According to the documentation, the signals that were used in this project were filtered using hardware high-pass filters. Thus, no additional baseline wander has been performed. In order to improve the robustness of the introduced algorithms, it could be beneficial to apply more strict noise filtering to the signals.

4.2 Classification of N1

As can be seen in Table 4 for all used methods, N1 stage detection accuracy repeatedly has the lowest values out of all sleep stages. The reason for this might be that our learning dataset has a very low number of epochs that are being classified as N1 stage relatively to other stages which is visible in Table 1. That follows the pattern shown in the article published in 2018 [11], where the detection accuracy of N1 stage is shown to be lower than other stages for all previous studies and that most databases have a heavy imbalance towards the Wake stage. The article suggests to use both N1 and overall accuracy to assess the quality of the model.

4.3 Effect of Imbalanced Dataset on Accuracy

As mentioned in Section 2.1.1, the dataset is unbalanced, meaning that the number of epochs per stage is not uniformly distributed. This causes two main problems. If the distance of the clusters (i.e. sleeping phase) in the feature-space are unevenly distributed in a way that one or two classes are easier to detect and if those classes correspond to the classes that have the most epochs in the dataset, the accuracy will be substantially better. If more hard-to-classify epochs are in the dataset, the accuracy will be worse. Thus, the accuracy is highly dependent on the test set. As seen in Table 4, the Wake and REM phase seem to be easily detectable by all methods. Especially the wake phase is prone to be over-represented in the datasets, since it is highly dependent on the start- and stop-time of the data acquisition.

The second problem that might arise from an unbalanced dataset is that the training optimizes the overall accuracy instead taking into account the accuracies per phase. Thus, it is likely that sleep phases with a small appearance in the dataset are ignored by the optimizer, thus leading to a bias in the classification towards stages that are dominantly present in the dataset. This phenomenon can be seen for the classifications of the N1-sleep-stage. Therefore, it might be reasonable to artificially balance out the dataset.

4.4 Inclusion of more Modalities

It can be seen in Table 2 that the classification performance increased with the inclusion of a second EEG channel and the inclusion of EOG data, which is sensible, since the sleep phases have an effect on the brain waves as well as the eye movement. Nonetheless this increase could be caused by chance. Thus, it might be reasonable to analyse the results per epoch using a Paired Student's t-test. However, this detailed analysis is out of the scope of this report. Furthermore it can be seen, that the accuracy did not increase substantially when including ECG data. Thus, it is likely that ECG does not carry a lot of information that is useful for sleep classifications. To further improve the classification performance of SVM as well as FCNN models, more statistical features could be obtained from EOG after discrete wavelet decomposition using Daubechies-4 (db4) wavelet as mentioned in [11] where sleep stages were classified using only single-channel EOG.

4.5 Difference Between Validation and Test Accuracy

As previously seen the CNN seems to perform the best on the validation set compared to the other methods, yet the test accuracy is substantially lower. The difference between validation and test accuracy is the highest for the CNN. This could be explained by a higher sensitivity to patient specific features compared to the other methods, since the validation set included epochs from patients that were included in the train set as well. Thus, the CNN would overfit the data. This could be overcome by using regularisation methods such as dropout. Furthermore, one could consider changing the architecture, as the presented architecture was a proof of concept and was not empirically optimized. Varying the filter sizes or changing the architecture to an Long-Short-Term-Memory (LSTM) might be an option for future work.

4.6 Generalizability of Results

Due to the small dataset, a datasplit was chosen to exclude only two patients from the hyperparameter-optimization and training. Thus, there is a risk that these two patients are outliers and in fact do not represent the general population. In order to evaluate how similar these two patients to the other ones are, the whole experiment, including all hyperparameter-optimizations could be redone on other sets of the data. Due to time limits, this was not done in this project.

4.7 Best Method

In this project, it has been found that the FCNN performs the best on the present dataset, thus it is recommended to use this method. However, all methods failed to accurately classify N1. Therefore, it should be studied how to improve the classification of that stage. Given that CNNs have outperformed other methods in analysis of sleep data [7, 8], it might be also sensible to further study different CNN architectures to improve the classification performance.

5 Conclusion

Throughout this project, the classification of sleep stages was studied and their importance in the sleep rating understood. Different modalities were extracted from the dataset and pre-processed, before extracting features from these data. Thereafter, different algorithms to classify sleep stages were introduced, trained, and their results were compared.

First, pre-processing was applied in the form of wavelet filtering of the EEG data. Thereafter, feature extraction was implemented in order to summarise data over epochs in different features. To do so, temporal, frequency and Hjorth features were extracted. Further, different classification algorithms were introduced.

An cubic-kernel SVM algorithm was trained on 8 patients using 5-fold cross-validation. The data used was the features extracted for every epoch. The validation accuracy obtained was 78.2% using the EEG and EOG modalities but yielded only 61% accuracy on the test data consisting of two other patients. The TPR and PPV on the N1 stage were nevertheless respectively 7.9% and 8.2%.

Further, a FCNN was optimized using the same training method as the SVM, which using EEG, EOG and EMG data, yielded accuracies of 83.3% on validation and 70.2% on test set. However, the FCNN never predicted N1 stages.

Lastly, a CNN was also optimised, this time without data extraction but by applying successive convolutions to the chosen modalities. This algorithm yielded on EEG and

EOG data an accuracy of 84.3% on validation data and 64.7% on test data. The CNN also failed at predicting N1 stages.

It is therefore concluded that higher accuracies should be achieved both on the N1 stage and globally. The different algorithms used yielded either lower overall accuracies with higher N1 accuracy (SVM) or the opposite, high general accuracy, but null accuracy for N1. Methods to compensate class unbalance of N1 could be introduced to aim at reducing this unbalance. Also further investigation in the splitting of data between train-validation and test could be done, by repeating the hyperparameters optimisation on different sets of validation patients, and comparing results.

References

[1] Aakash K. Patel, Vamsi Reddy, and John F. Araujo. "Physiology, Sleep Stages". eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2022. URL: http://www.ncbi.nlm.nih.gov/books/NBK526132/.

[2] Guo-Qiang Zhang et al. "The National Sleep Research Resource: towards a sleep data commons". In: *Journal of the American Medical Informatics Association: JAMIA* 25.10 (Oct. 1, 2018), pp. 1351–1358. ISSN: 1527-974X. DOI: 10.1093/jamia/ocy064.

[3] S. F. Quan et al. "The Sleep Heart Health Study: design, rationale, and methods". In: *Sleep* 20.12 (Dec. 1997), pp. 1077–1085. ISSN: 0161-8105.

[4] R B Berry, S F Quan, and A R Abreu. *The AASM Manual dor the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2.6*. Darien, IL: American Academy of Sleep Medicine.

[5] *1.3.5.11. Measures of Skewness and Kurtosis*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm.

[6] M. Vourkas, Micheloyannis Sifis, and Giorgos Papadourakis. "Use of ANN and Hjorth parameters in mental-task discrimination". In: Feb. 2000, pp. 327–332. ISBN: 0-85296-728-4. DOI: 10.1049/cp:20000356.

[7] "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification". In: *Ieee Transactions on Bio-Medical Engineering* 66.5 (Oct. 22, 2018), pp. 1285–1296. ISSN: 0018-9294. DOI: 10.1109/TBME.2018.2872652. pmid: 30346277. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6487915/ (visited on 04/28/2022).

[8] Yu Xue, Ziming Yuan, and Adam Slowik. "A Novel Sleep Stage Classification Using CNN Generated by an Efficient Neural Architecture Search with a New Data Processing Trick". Mar. 19, 2022. arXiv: 2110.15277 [cs, eess]. URL: http://arxiv.org/abs/2110.15277 (visited on 04/28/2022).

[9] Dale Purves et al. "Stages of Sleep". In: *Neuroscience. 2nd edition* (2001). Publisher: Sinauer Associates. URL: https://www.ncbi.nlm.nih.gov/books/NBK10996/ (visited on 05/04/2022).

[10] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[11] Md Mosheyur Rahman, Mohammed Imamul Hassan Bhuiyan, and Ahnaf Rashik Hassan. "Sleep stage classification using single-channel EOG". In: *Computers in Biology and Medicine* 102 (2018), pp. 211–220. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2018.08.022. URL: https://www.sciencedirect.com/science/article/pii/S0010482518302427.