Specializing CNN Models for Sleep Staging Based on Heart Rate

Miriam Goldammer[1], Sebastian Zaunseder[2], Hagen Malberg[1], Felix Gräßer[1]

[1] Institute of Biomedical Engineering, TU Dresden, Dresden, Germany
[2] Department of Information Technology, FH Dortmund, Dortmund, Germany

Abstract

This work aims to classify sleep stages based on tachograms using Convolutional Neural Networks (CNNs) and investigate advantages of specialized classifiers.

The tachograms of 5422 patients were extracted from the Sleep Heart Health Study. A CNN was trained to classify each 30 s epoch into four distinct sleep stages. The patients were divided into four subgroups by Apnoe-Hypopnoe-Index (AHI). From each subgroup, 20 % of patients were held out as test data. One general model was trained on all training patients and four narrowed models were each trained on one subgroup. Furthermore, the general model was retrained on the subgroups, yielding four additional transfer learning models.

Our general model gained an average Cohen's Kappa score of 0.53. The general model outperformed the narrowed models on each test subset. From the narrowed models, training on the subgroup with AHI 5-15 achieved best overall performance. However, a correlation exists between the size of train sets and classification quality. Transfer learning did not improve the results.

CNN models are capable of learning features from tachograms with very good classification performance compared to other works using heart rate only. However, the pursued strategies for specializing classifiers did not yield any advantages over our general model.

1. Introduction

Sleep staging is traditionally done from a polysomnogram (PSG). However, research in the last decade also focused on alternative ways for automatic sleep stage classification. These approaches usually incorporate a reduced set of signals (e. g. electrocardiogram (ECG), radar, acceleration), feature extraction and machine learning. Many successful approaches use cardiorespiratory features, especially heart rate variability and other features from the tachogram. In accordance with the recent developments in machine learning, utilized classifiers in sleep staging have moved from basic approaches such as linear discrimination analysis (e. g. [1] in 2015) to complex models such as long short-term memory (LSTM) networks (e. g. [2] in 2019). The latter approach considers over 100 features related to the cardiorespiratoy system [2]. Nevertheless, there is no consensus on best suited features. Therefore, some recent approaches skip the step of manually designed feature extraction by applying machine learning directly to a signal.

Our model was inspired by Malik et al. (2018) [3] who used a convolutional neural network (CNN) on the tachogramm. Korkalainen et al. (2020) [4] and Sun et al. (2020) [5] adopt a more complex network architecture of extending the CNN by a bi-directional LSTM. However, there are some differences: Korkalainen et al. classified segments from the downsampled photoplethysmogram (PPG) and Sun et al. focused on a parallel architecture, that uses the R-peaks from the ECG and a downsampled respiration signal.

Our work optimizes a mere CNN model on only tachograms and additionally considers the patient characteristic Apnea-Hypopnea-Index (AHI) by training data selection for specialized models. We show that sleep staging from the tachogram by CNNs has a high potential and compare specialized models to a general model.

2. Methods

2.1. Data and Preprocessing

We used only ECGs and metadata data from the full-night PSGs of the Sleep Heart Health Study Visit 1 (SHHS1) [6, 7], containing a wide range of sleep qualities, AHIs and comorbidities. Sleep stages according to Rechtschaffen and Kales are available with the PSGs.

As preprocessing, we extracted a filtered beat-to-beat interval (RRI) time series by QRS detection according to [8, 9] and iterative filtering to remove outliers according to [10]. However, due to poor QRS detection, we excluded 382 of the 5804 patients from SHHS1, leaving 5422 patients for our analysis. To create appropriate input for a CNN, we interpolated the RRIs linearly, resampled at 4 Hz and normalized by subtracting and dividing by the mean value of each night. Afterwards, we created overlapping segments of 300s length, centered around each sleep stag-

ing label. We excluded the first and last 5 min of each night and segments with less than 170 RRIs, as the latter implies a heart rate of less than 34 bpm. This results in a total of 5.1 million segments.

We generated a reduced number of classes for sleep staging: (a) when using 2 classes, we distinguished Wake and Sleep, (b) when using 3 classes, we distinguished Wake, Non-REM Sleep and REM Sleep, (c) when using 4 classes, we distinguished Wake, Light Sleep (S1 and S2), Deep Sleep (S3 and S4) and REM Sleep and (d) when using 5 classes, we distinguished Wake, S1, S2, Deep Sleep and REM Sleep. We generally use 4 classes, other allocation is for comparison to the literature (Table 2).

2.2. Validation

Before training any models, we separated around 20 % of the patients as our designated test set S^{test}_{all}. AHI distribution in S^{test}_{all} is representative of the whole data set and can therefore be separated into four subsets of test data S^{test}_{0-5}, S^{test}_{5-15}, S^{test}_{15-30}, $S^{test}_{>30}$. The subscript index denotes the AHI range in the subset. Consequently, around 80 % of the patients remain as training set S^{train}_{all}, which is split into analog subsets S^{train}_{0-5}, S^{train}_{5-15}, S^{train}_{15-30}, $S^{train}_{>30}$ as needed. The number of patients in these subsets varies strongly (Table 1).

Table 1. Number of patients in sets and subsets.

AHI	all	0-5	5-15	15-30	>30
S^{train}	4420	1318	1903	869	330
S^{test}	1002	294	436	199	73

We expected several problems with using the full train set in the grid search: overfitting, computing time, the later varying subsets sizes (Table 1), and the presumably negative effect of acute cardiovascular diseases on classification of the tachogram. Therefore, we identified 364 patients in S^{train} that we assume to be heart-healthy, according to the available meta data, and used only this subset S^{train}_{heart} for the grid search.

We applied k-fold cross validation (kfcv) for both hyperparameter optimization by means of a grid search and final model evaluation. For evaluation in the grid search, we calculated the mean classification quality on the validation data over all folds. In contrast, for final evaluation on the test data, classification of each segment was a majority vote of all k models.

For further interpretation, we generally evaluated model performance by Cohen's Kappa (κ) [11] rather than accuracy, to accommodate for class imbalances. For mere inter-model comparison, i. e. in the grid search, we used loss.

2.3. CNN Architecture

We started our optimization from the architecture described by Malik et al., which stands out for using convolutional layers with stride 2 instead of pooling layers. This results in convolutional blocks with two convolutional layers that are equivalent in all features, except the stride is 1 for the first layer and 2 for the second layer.

In our architecture, a number of similar convolutional blocks is stacked, flattened and then followed by dense layers and one output layer for classification (Fig. 1). In a grid search, we varied

- the number of convolutional blocks between 3 and 7,
- the number of filters in each layer between 10 and 96,
- the filter sizes between 4 and 32,
- the number of dense layers between 1 and 3,
- the number of neurons in dense layers between 20 and 400, and
- the dropout rate between 0 and 0.5.

For each cross validation run, these hyperparameters were the same in all layers they applied to. The general training setup was Adam optimizer with learning rate 10^{-4}, categorical cross entropy as loss function and early stopping.

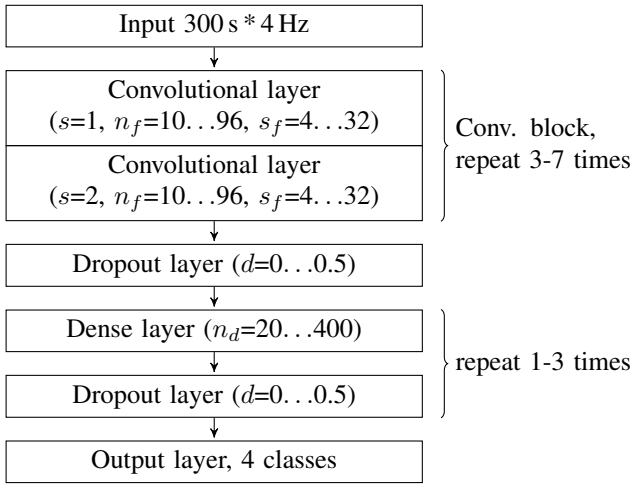

Figure 1. Schematic architecture of CNNs in the grid search. With stride s, number of filters n_f, filter size s_f, number of dense neurons n_d, dropout rate d.

2.4. Specializing Classifiers

We used two strategies to specialize classifiers: (a) only use a subset of the training data to train the model and (b) train the model with all training data, and then use transfer learning to retrain with a subset of the training data.

To explore these strategies, we used the best architecture from the previous grid search and a 5fcv. We therefore calculated five versions (i. e. by each fold) of

- a general model M^g from S^{train}_{all},
- four narrowed models M^n_{0-5}, M^n_{5-15}, M^n_{15-30}, $M^n_{>30}$ from S^{train}_{0-5}, S^{train}_{5-15}, S^{train}_{15-30}, $S^{train}_{>30}$, respectively, and
- four transfer learning models M^t_{0-5}, M^t_{5-15}, M^t_{15-30}, $M^t_{>30}$ from M^g and S^{train}_{0-5}, S^{train}_{5-15}, S^{train}_{15-30}, $S^{train}_{>30}$, respectively.

3. Results

The overall performance (i. e. κ) of the general model M^g of 0.53 is equivalent to comparable current research (Table 2). This best model from our grid search has 4 convolutional blocks with 32 filters of size 8, 1 dense layer with 400 neurons, and a dropout rate of 0.3. It yielded a mean κ of 0.47 in the 3fcv with S^{train}_{heart}.

Comparing the general model to specialized models, it stands out that for each test set, M^g performs at least as good as any specialized model (Table 3).

The narrowed models M^n with their massively reduced training set show a strong performance for their respective test set. However they only come close to M^g for AHIs <15. Note that for S^{test}_{15-30} and $S^{test}_{>30}$, any model M^n trained on patients with lower AHI will still perform at least as good as M^n_{15-30} and $M^n_{>30}$, respectively.

In contrast, the transfer learning models M^t perform as well as M^g on their respective test sets. However, they show a slight drop in general performance on the other test sets.

Table 2. Model performance (κ) on test data, comparison to the literature for different combinations of sleep stages into classes. All models use input data from only the ECG or PPG. Note that Sun et al. improved κ to 0.59 for 5 classes when including an additional respiratory signal.

Number of classes	2	3	4	5
M^g	0.65	0.65	0.53	0.51
Malik et al.	0.38	-	-	-
Korkalainen et al.	-	0.65	0.54	0.51
Sun et al.	-	0.65	-	0.49

Table 3. Model performances (κ) on different test sets.

Test set	S^{test}_{all}	S^{test}_{0-5}	S^{test}_{5-15}	S^{test}_{15-30}	$S^{test}_{>30}$
M^g	**0.53**	**0.53**	**0.54**	**0.51**	**0.51**
M^n_{0-5}	**0.50**	**0.51**	0.51	0.47	0.45
M^n_{5-15}	**0.51**	0.50	**0.52**	0.49	0.49
M^n_{15-30}	**0.46**	0.44	0.47	**0.47**	0.47
$M^n_{>30}$	**0.40**	0.37	0.40	0.41	**0.45**
M^t_{0-5}	**0.53**	**0.53**	0.54	0.50	0.50
M^t_{5-15}	**0.53**	0.53	**0.54**	0.51	0.51
M^t_{15-30}	**0.52**	0.52	0.53	**0.51**	0.51
$M^t_{>30}$	**0.51**	0.50	0.52	0.50	**0.51**

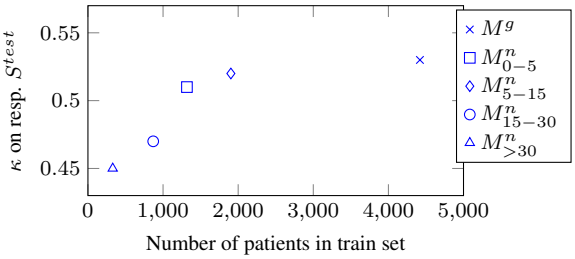

Figure 2. Mean κ of models M^g and M^n on the respective S^{test} over number of patients in the respective S^{train} (Table 1), i. e. the first boldly printed diagonal in Table 3.

As almost all models show the best performance on S^{test}_{5-15}, we cannot assume a direct correlation between AHI and performance. Rather, there seems to exists a correlation between size of train set and performance (Fig. 2).

Further examination by randomly downsampling all S^{train} to the same number of patients (i. e. 330) and training new models confirmed this hypothesis, as validation performance only varied between κ of 0.44 and 0.46 for all AHI subgroups in that experiment. The grid search results on S^{train}_{heart} also support this.

4. Discussion

As we aimed to classify sleep stages based on tachograms using CNNs, we found a CNN architecture that performs similar to state of the art CNN-LSTM architectures.

We confirm that the tachogramm contains essential information about sleep stages by using inputs that only consists of the location of R-peaks in the ECG, just as Malik et al. and Sun et al., whereas Korkalainen et al. account for the waveform of the PPG. Our results rest on a large, independent test set of 1000 patients. However, we have not tested our model on data from studies other than SHHS1. Our main improvement to Malik et al. with its comparable architecture is most likely the amount of data and some optimization of the architecture. And even though Korkalainen et al. also use significantly more data than Malik et al., there are only ca. 800 PSGs for training and 90 for testing. Only Sun et al. include significantly more PSGs and slightly more segments than our work.

In contrast to our work, both Korkalainen et al. and Sun et al. use an additional bi-directional LSTM layer, therefore applying a more complex model to the data. We assume that adding an LSTM layer to our own architecture will further improve the results, just as reversely, we assume the results of Korkalainen et al. and Sun et al. would improve by further optimization of the CNN. Widening the view to general cardiorespiratory sleep staging performance in the literature, we find that that adding respiratory

features and using an LSTM seem to be essential parts to further improve the classification. This was shown for raw signals ([5] with κ 0.59 for 5 classes) and feature engineering ([2] with κ 0.61 for 4 classes).

Regarding our second aim, the pursued strategy for specializing classifiers did not yield any advantages over general classifiers. Neither specializing by AHI-narrowed models nor transfer learning improved the classification performance compared to the general classifier. Similarly, the specialized models did not have any advantages over the general classifier (e. g. computing time, amount of data needed) that did not come with a decrease in performance. On the contrary, we found a significant disadvantage in using narrowed models M^n for subgroups with AHI $>$30, as they seem to profit most from additional training data. We assume, the reason for this are critically small subgroups in our calculations, because of (a) the rather good results of M^n_{5-15} on S^{test}_{15-30} and $S^{test}_{>30}$ and (b) our appended experiment that showed a strong decrease in performance when randomly downsampling the train sets. Therefore, a secondary finding of our experiments with narrowed classifiers is a strong dependence of the classification performance on the train set size. This goes well with the machine learning truism "More data is always better".

5. Conclusion

Even though literature generally shows that Neural Networks are capable of classifying sleep stages from the tachogram and other time series, our research amplifies that mere CNNs have a higher potential in this task than previously assumed. Our next steps will be towards adding an LSTM layer and using different cardio-respiratory time series parallelly. Just as the literature suggests, our preliminary research in these courses shows promising results for both strategies.

Despite our negative results for specializing models by AHI, we will still look further into other patient features that might influence the automatic classification (e. g. body mass index, age, medications), in accordance with our general research in strategies for individualized classification.

However, from the dependency on train set sizes and the undefeated performance of the general model with its diverse training data, we assume that combining data from various sources and studies will be essential for further improvements in sleep staging.

Acknowledgements

We thank the Center for Information Services and High Performance Computing (ZIH HPC) at TU Dresden for generous allocations of computer time.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute.

References

[1] Fonseca P, Long X, Radha M, Haakma R, Aarts RM, Rolink J. Sleep stage classification with ECG and respiratory effort. Physiological Measurement apr 2015;36(10):2027.

[2] Radha M, Fonseca P, Moreau A, Ross M, Cerny A, Anderer P, Long X, Aarts RM. Sleep stage classification from heart-rate variability using long short-term memory neural networks. Scientific Reports apr 2019;9(1):14149.

[3] Malik J, Lo YL, Wu HT. Sleep-wake classification via quantifying heart rate variability by convolutional neural network. Physiological Measurement aug 2018; 39(8):085004.

[4] Korkalainen H, Aakko J, Duce B, Kainulainen S, Leino A, Nikkonen S, Afara IO, Myllymaa S, Töyräs J, Leppänen T. Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. Sleep may 2020;.

[5] Sun H, Ganglberger W, Panneerselvam E, Leone MJ, Quadri SA, Goparaju B, Tesh RA, Akeju O, Thomas RJ, Westover MB. Sleep staging from electrocardiography and respiration with deep learning. Sleep jul 2020;43(7).

[6] Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW. The Sleep Heart Health Study: design, rationale, and methods. Sleep dec 1997;20(12):1077.

[7] Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S. The National Sleep Research Resource: towards a sleep data commons. Journal of the American Medical Informatics Association oct 2018; 25(10):1351.

[8] Afonso VX, Tompkins WJ, Nguyen TQ, Luo S. ECG beat detection using filter banks. IEEE Trans Biomed Eng feb 1999;46(2):192.

[9] Vidaurre C, Sander TH, Schlögl A. BioSig: The Free and Open Source Software Library for Biomedical Signal Processing. Comput Intell Neurosci 2011;2011:935364.

[10] Wichterle D, Simek J, La Rovere MT, Schwartz PJ, Camm AJ, Malik M. Prevalent low-frequency oscillation of heart rate: novel predictor of mortality after myocardial infarction. Circulation sep 2004;110(10):1183.

[11] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement apr 1960; 20(1):37.

Address for correspondence:

Miriam Goldammer
miriam.goldammer@tu-dresden.de