2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Enhancing Current Cardiorespiratory-based Approaches of Sleep Stage Classification by Temporal Feature Stacking

Lucas Weber[1], Maksym Gaiduk[2], Ralf Seepold[3],
Natividad Martínez Madrid[4], Martin Glos[5], Thomas Penzel[6]

***Abstract*— This paper presents a generic method to enhance performance and incorporate temporal information for cardiorespiratory-based sleep stage classification with a limited feature set and limited data. The classification algorithm relies on random forests and a feature set extracted from long-time home monitoring for sleep analysis. Employing temporal feature stacking, the system could be significantly improved in terms of Cohen's κ and accuracy. The detection performance could be improved for three classes of sleep stages (Wake, REM, Non-REM sleep), four classes (Wake, Non-REM-Light sleep, Non-REM Deep sleep, REM sleep), and five classes (Wake, N1, N2, N3/4, REM sleep) from a κ of 0.44 to 0.58, 0.33 to 0.51, and 0.28 to 0.44 respectively by stacking features before and after the epoch to be classified. Further analysis was done for the optimal length and combination method for this stacking approach. Overall, three methods and a variable duration between 30 s and 30 min have been analyzed. Overnight recordings of 36 healthy subjects from the Interdisciplinary Center for Sleep Medicine at Charité-Universitätsmedizin Berlin and Leave-One-Out-Cross-Validation on a patient-level have been used to validate the method.**

***Clinical relevance*— The method can be employed generically to feature sets for (small scale) datasets to improve classification performance for classification problems with temporal relations with random forest classifiers.**

I. INTRODUCTION

Sleep plays an essential role in memory consolidation, muscle and tissue rejuvenation, and productivity during the day [12]. With the increasing prevalence of certain sleep disorders, long-term and mobile/remote monitoring of sleep structure becomes more critical, with sleep stage classification being one of the main parts [8]. Due to the limitations of a measurement environment at home, research in this field recently pivots in the direction of cardiorespiratory-based sleep stage classification [11]. The restriction to heart rate and respiratory activity enables easily applicable, mobile, and long-time measurements that can be carried out at home. Recent publications have shown that the classification of sleep stages from these features is feasible and can be enhanced to great performance [6, 11]. Also, in recent years there have been various systems with different sensor setups and measurements [2, 15, 14, 3]. Due to various sensor setups and the nature of these different measurements systems, it is often not possible to extract annotated large scale databases of thousands of patients like in [6, 11] which enable the use of intense neural networks with millions of trainable parameters that learn highly specialized features from millions of different sleep phases. In order to deal with restricted features and data amounts, researchers often employ classical machine learning algorithms along with handcrafted feature sets [2, 3]. Even with classical machine learning algorithms, there is a vast homogeneity in the number of handcrafted features. The number of features varies between 10 (e.g., in [3]) up to over a hundred different features (e.g., in [9]). The objective of this paper was to improve classification performance for a restricted feature set that can be extracted from unobtrusive sensors in a home environment. The feature set has been taken from [3]. The goal was to improve classification performance by constructing a generic method to incorporate temporal information in sleep stage classification, which has been proven to improve performance [3, 4]. The method is based on random forests, a limited feature set from [3] and a private dataset obtained from the Interdisciplinary Center for Sleep Medicine at Charité-Universitätsmedizin Berlin containing

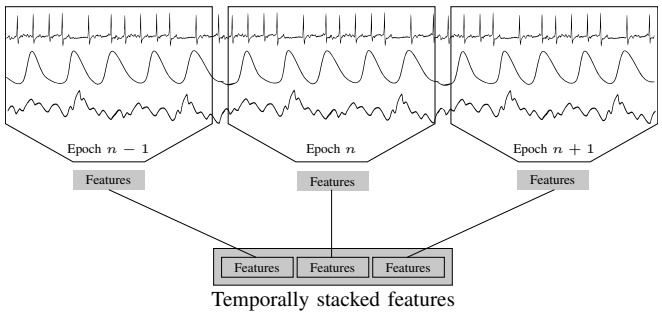

Fig. 1. Visualization for the feature stacking of $k = 1$.

[1]Lucas Weber is with the Ubiquitous Computing Lab at HTWG Konstanz, Konstanz, Germany, (lucas.weber@htwg-konstanz.de)

[2]Maksym Gaiduk is with the Ubiquitous Computing Lab at HTWG Konstanz, Konstanz, Germany and the Universidad de Sevilla, Seville, Spain, (maksym.gaiduk@htwg-konstanz.de)

[3]Ralf Seepold is with the Ubiquitous Computing Lab at HTWG Konstanz, Konstanz, Germany and the Department of Information and Internet Technology at I.M. Sechenov First Moscow State Medical University, Moscow, Russia (email: ralf@ieee.org)

[4]Natividad Martínez Madrid is with the IoT Lab at Reutlingen University, Reutlingen, Germany and the Department of Information and Internet Technology at I.M. Sechenov First Moscow State Medical University, Moscow, Russia (e-mail: nati@ieee.org)

[5]Martin Glos is with the Interdisciplinary Center for Sleep Medicine, Charité-Universitätsmedizin Berlin, Berlin, Germany (martin.glos@charite.de)

[6]Thomas Penzel is with the Interdisciplinary Center for Sleep Medicine, Charité-Universitätsmedizin Berlin, Berlin, Germany, (thomas.penzel@charite.de) and supported by the Russian Federation government mega-grant Nº 075-15-2019-1885

whole night records of polysomnography (PSG) from 36 patients. Incorporating temporal feature information from previous and following epochs has improved performance by around 50 % in terms of Cohen's κ for multiple sets of sleep stage grouping. Capturing the temporal information essentially is based on concatenating features from surrounding epochs, which is visualized in Figure 1. The basic reasoning is that information from previous and following epochs might enhance the classification by capturing temporal correlations between consecutive sleep stages.

II. Materials and Methods

A. Dataset

The Charité dataset contains 277 h of recording out of PSG from n=37 who were healthy and were nearly equal distributed with males and females (Age: 38.5 $\pm$ 14.5 years, BMI: 24.4 $\pm$ 4.9 $\frac{kg}{m^2}$). The initial study was carried out in Charité-Universitätsmedizin Berlin, Center of Sleep Medicine. Due to missing recordings, one patient has been excluded from the dataset, which leaves 276 h of recordings. After the exclusion of heavy movement phases and not identifiable sleep stages and the exclusion of initial sleep stages (for purposes of calculating longer duration features), but before the exclusion of epochs due to the temporal stacking, the dataset consists of 33230 epochs with a length of 30 s that have been annotated with their respective sleep stage according to the rules of the American Academy of Sleep Medicine (AASM) by a trained scorer. The class distribution can be seen in Table I. For this paper, there are three different groups of sleep stages. For these three classifications tasks, the stages are grouped as follows:

- **Three classes**. The sleep stages are divided in Wake, Non-REM (N1, N2, N3, N4), REM.
- **Four classes**. The sleep stages are divided in Wake, Non-REM light Sleep (N1+N2), Non-REM deep sleep (N3, N4), REM.
- **Five classes**. The sleep stages are divided in Wake, N1, N2, N3/4, REM.

From the overnight PSG-recordings the ECG (Lead II), thoracical effort signal measured by respiratory inductive plethysmography and a movement signal (calculated as the absolute value from a three-dimensional accelerometer sensor). ECG, respiration, and acceleration have sampling rates of 256 Hz, 32 Hz, and 32 Hz, respectively.

B. Data preprocessing and feature extraction

The features are mainly extracted from [3]. They are strongly oriented towards non-invasive measurements and do not rely on high-quality signals. Most importantly, breathing, heart rate, and the amplitude of movement are used to extract a small feature set of 10 different features. Signal preprocessing mostly is filtering the ECG-Signal to improve QRS-Detection. A short overview of the features, including a short description, is given in Table II. For in-depth calculation details, please refer to [3]. The feature values have been normalized by calculating the mean for the whole night for every patient [3].

TABLE I
Class distribution

Class Number	Class Names				
Five Classes	Wake	N1	N2	N3/4	REM
	6621	5296	11460	5921	3922
Four Classes	Wake	Light		Deep	REM
	6621	16756		5921	3922
Three Classes	Wake	NREM			REM
	6621	22677			3922

C. Temporal Feature Stacking

Sleep stages are by default assigned every 30 s, but recent papers have shown that accessing information before and after the epoch improves classification [3, 5, 11]. Goldammer et al. [5], as well as Sun et al. [11] incorporate 270 s and 300 s centered around the current epoch to enable their convolutional neural networks to extract features from a longer duration before and after the epoch that should be classified. On top of that Sun et al. [11] also employ a recurrent neural network to represent even longer temporal contexts in their classification. This approach reaches very high performances in terms of Cohens Kappa and Accuracy. Others used post-processing steps like transition probabilities between different stages [3, 10]. In these approaches, we decided to employ a method that we call temporal feature stacking to incorporate information of surrounding epochs to the classifier. The idea of this paper and the distinction to

TABLE II
Table of features.

Abbreviation	Short description
HR	mean heart rate [1/min] of epoch
HBI	Mean temporal RR-Interval [msec] epoch
HRV	Heart Rate Variability, calculated as mean difference of successive R-Peaks intervalls of epoch
RA	R(k)-Algorithm by [7]
BM	The sample mean of body movement during the epoch
TSDM	Mean respiratory depth of exhalation during the epoch
PSDM	Mean respiratory depth of inhalation during the epoch
VBR	Median respiratory volume during breathing cycles
VIN	Median respiratory volume during inhalation
DA	D(k)-algorithm by [7]

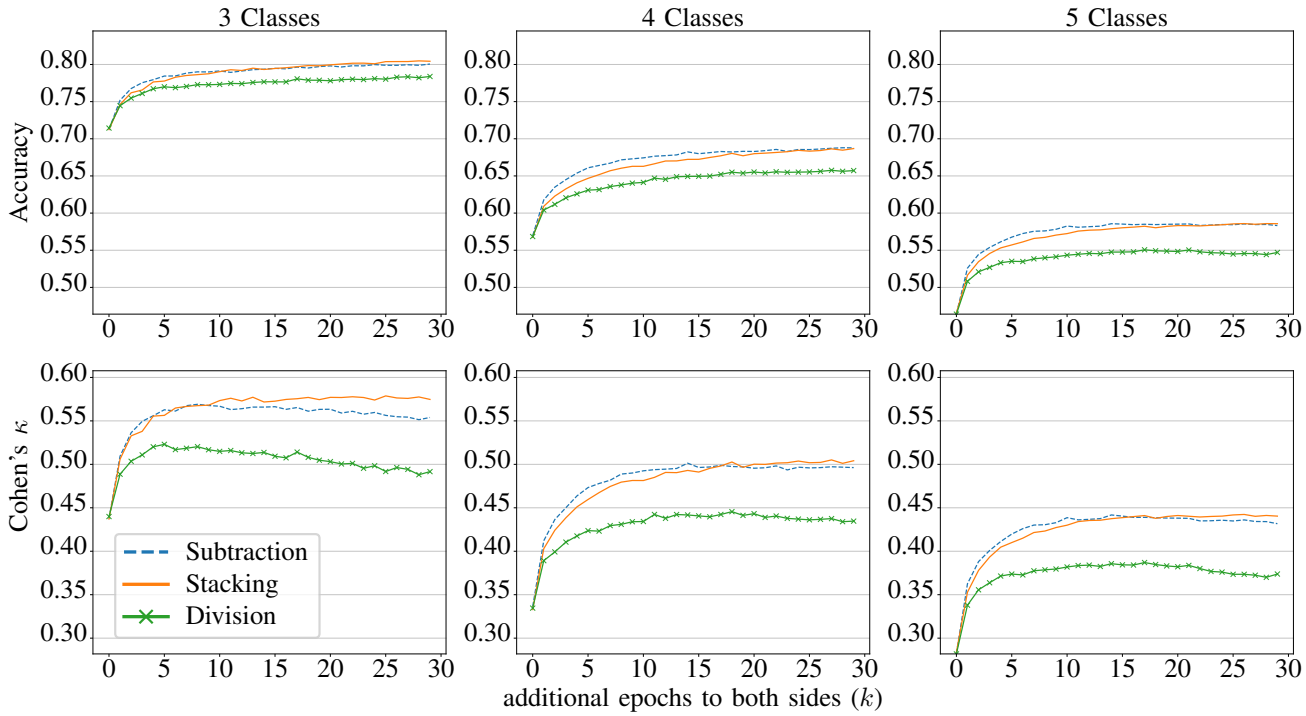

Fig. 2. The results over multiple methods and the number of surrounding epochs to stack.

[3] is to concatenate features from surrounding epochs to the current epoch's features. This approach, in conjunction with the chosen machine learning method, sets us apart from our previous publication.

The concatenation results in a feature vector with $(2k+1) * N$ features, where k is the number of epochs prior and after the current one and N is the number of features one calculates per epoch (N equals 10 in the case of this paper). Unfortunately, for large k, this might result in the curse of dimensions, where the feature space has such a high dimensionality that the classifier can separate each sample perfectly and, as a consequence, overfits heavily. To counteract this, we employ random forests which are known to be resistant to high dimensionalities [1]. This enables us to stack the features without additional feature selection. Unfortunately, due to our stacking method, we lose $2k$ samples for every patient since the epochs at the beginning of a night, and the end does not have previous or following epochs. In the process, we, therefore, lose a maximum of $36 * 2k$ samples. For our analysis, we change k from zero up to 30. Therefore we lose a maximum of 2160 samples or 6.5 %. In our analysis, we also employ three methods to concatenate features from surrounding epochs.

- **Stacking**. Concatenate features from surrounding epochs to both sides.
- **Subtraction**. Here we subtract the features from the current epoch from the features of surrounding epochs and concatenate this subtraction.
- **Division**. Here we divide features of surrounding epochs by the features of the current epoch and concatenate the result. In order to avoid zero division, we set all zeros in the features of the current epoch to the machine ϵ of 8 Byte floating-point numbers.

D. Classifier design and training method

The classifier is a Random Forest Classifier that consists of 500 Trees. It is trained using the bagging method and only learning from a maximum of 12000 samples per tree. As mentioned in the previous chapter, we employ an ensemble of binary trees to counteract the high dimensionality of up to 610 features (for $k = 30$).

During training, we incorporate the heavy class imbalance already visible in table I by setting an upper bound of 10000 samples per class. This, together with the 12000 samples per tree, guarantees that each tree sees at least samples of more than one class. For evaluation, we implement Leave-One-Out-Cross-Validation on a patient level, which guarantees the maximum amount of training data and patient-level separation.

III. Results

The results in this research paper will be twofold. In the first step, we are empirically trying to find the optimal stacking method and stacking length by training the ensemble for three different class separations, three methods, and 31 different durations. For each of these combinations, we employ Leave-One-Out-Cross-Validation on patient-level to obtain a gross statistic for Cohen's κ and accuracy (calculated with the common/merged confusion matrix of all results). We then analyze the best combination by class accuracy, kappa, and confusion matrix to give in-depth information about the classification results. This also helps to make the algorithm comparable to other results, as accuracy and

kappa are heavily influenced by the class distribution of the test data. The first stage of evaluation results can be seen in Figure 2. The confusion matrix for the best performing classifier and four classes can be found in Table III.

IV. Discussion

A. Optimal length of interest and method comparison

The results show an increase for almost all stacking methods within a k of zero to five (up to ten additional epochs). Overall the accuracy and κ could be increased significantly within a range of up to 15 % κ. These results show that temporal feature stacking can improve classification results for the temporally related classes of sleep staging. Comparing the three different class divisions, performance is inversely proportional with the amount of classes from three to five classes and therefore the difficulty of the classification tasks. In terms of the different methods stacking and subtraction show the best performances. Comparing stacking and division, one can notice a small section at the beginning of the graphs ($0 < k < 12$) where subtraction shows a sharper increase in classification performance, but after this point, stacking is equal or better for most cases. For all cases, there is no improvement after $k = 30$. For the stacking method, we achieve the optimal κ for all class divisions around $k = 25$. Here the performance reaches a κ of 0.44, 0.50, and 0.58 for five, four, and three classes.

B. Analysis of the final classifier

From the previous section, it became clear that a k of 25 with a simple stacking of features improves classification performance best. We will do further analysis on four stages of sleep since it is the most common class division. We now employ the best parameter $k = 25$ and allow the training algorithm to search through all features for every split and every tree. To give an in-depth view of the classification performance, the confusion matrix for four classes is shown in Table III. The best performing classifier reaches a κ of 0.51 and an accuracy of 69 % for four classes. Also, it achieves class accuracies over 50 % for every class, which shows the balanced classification performance. The overall performance is comparable or better to other state-of-the-art classifiers found in Table II in [13]. Moreover, the results are comparable with [13], when looking at one step classification. We achieve these results using just ten features and the proposed temporal feature stacking in comparison to the 74 features and about 6 % of the training data used in [13] (36 to 625 patients). Radha et al. [9] use ten times more features and eight times the amount of patients to achieve their extraordinary results of $\kappa = 0.61$. In the most recent, most comprehensive, and best-performing research work of [11], the deep neural networks achieve a κ of 0.58 using breathing and heart rate signal. In comparison to our approach, we achieve a lower performance, but the used network in [11] has around 15 million trainable parameters and is trained using 8682 PSG recordings compared to the 36 recordings in this paper.

TABLE III
Confusion matrix for stacking, $k = 25$ and four classes

		Classifier			
		Wake	Light	Deep	REM
Expert	Wake	3492	1793	82	70
	Light	1461	12491	1611	780
	Deep	187	1920	3743	46
	REM	354	1400	79	1911
Accuracies		0.64	0.76	0.63	0.51

V. Conclusion

Looking at these results, we want to emphasize that we do not aim for maximum performance in comparison to the most recent State-Of-the-Art-Classifiers, but we want to show a generic approach that can be used to enhance classification performance to temporally related classes while dealing with a restricted feature set and small-scale datasets. The presented approach with temporal feature stacking can be employed to other feature sets and temporal classifications problems. It isn't restricted to sleep stage classification and the features employed in this study. In our experiment, we could enhance kappa and accuracy significantly (around 0.15). Comparing these metrics to other State-of-the-Art classifiers, we achieve comparable performance with much fewer features and training data available. For the presented problem of sleep stage classification 51 timesteps ($k = 25$) and features stacking resulted in the best performance for all tested cases. In the future, we will conduct further analyses combining State-of-the-Art feature sets, for example from [13], and our approach to improve the classification performance. We hypothesize that results from some of the comparison studies mentioned in the discussion could also be improved by temporal feature stacking applied to their respective feature sets.

ACKNOWLEDGMENT

This research was partially funded by the EU Interreg V-Program "Alpenrhein-Bodensee-Hochrhein": Project "IBH Living Lab Active and Assisted Living", grants ABH040, ABH04, ABH066 and ABH068; and SRP grant of HTWG Konstanz.

References

[1] Gérard Biau. "Analysis of a Random Forests Model". In: *Journal of Machine Learning Research* 13.5 (2010), pp. 1063–1095.

[2] Reza Boostani, Foroozan Karimzadeh, and Mohammad Nami. "A comparative review on sleep stage classification methods in patients and healthy individuals". In: *Computer Methods and Programs in Biomedicine* 140 (2017), pp. 77–91.

[3] Maksym Gaiduk et al. “Automatic sleep stages classification using respiratory, heart rate and movement signals”. In: *Physiological Measurement* 39.12 (Dec. 2018), p. 124008.

[4] Miriam Goldammer et al. “Individualized Sleep Stage Classification from Cardiorespiratory Features”. In: *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*. Vol. 2019-May. IEEE, May 2019, pp. 1–6.

[5] Miriam Goldammer et al. “Specializing CNN Models for Sleep Staging Based on Heart Rate”. In: *2020 Computing in Cardiology Conference (CinC)*. Vol. 47. Dec. 2020, pp. 1–4.

[6] Henri Korkalainen et al. “Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea”. In: *Sleep* 43.11 (Nov. 2020), pp. 1–10.

[7] Yosuke Kurihara and Kajiro Watanabe. “Sleep-Stage Decision Algorithm by Using Heartbeat and Body-Movement Signals”. In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 42.6 (Nov. 2012), pp. 1450–1459.

[8] Thomas Penzel, Christoph Schöbel, and Ingo Fietze. “New technology to assess sleep apnea: wearables, smartphones, and accessories”. In: *F1000Research* 7.0 (Mar. 2018), p. 413.

[9] Mustafa Radha et al. “Sleep stage classification from heart-rate variability using long short-term memory neural networks”. In: *Scientific Reports* 9.1 (Dec. 2019), p. 14149.

[10] Haoqi Sun et al. “Large-Scale Automated Sleep Staging”. In: *Sleep* 40.10 (Oct. 2017).

[11] Haoqi Sun et al. “Sleep staging from electrocardiography and respiration with deep learning”. In: *Sleep* 43.7 (2020), pp. 1–12.

[12] Nico Surantha, Tri Fennia Lesmana, and Sani Muhamad Isa. “Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data”. In: *Journal of Big Data* 8.1 (Dec. 2021), p. 14.

[13] Alexander Tataraidze et al. “Sleep architecture measurement based on cardiorespiratory parameters”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Vol. 2016-Octob. IEEE, Aug. 2016, pp. 3478–3481.

[14] T. Willemen et al. “An Evaluation of Cardiorespiratory and Movement Features With Respect to Sleep-Stage Classification”. In: *IEEE Journal of Biomedical and Health Informatics* 18.2 (Mar. 2014), pp. 661–669.

[15] Yuezhou Zhang et al. “Sleep Stage Classification Using Bidirectional LSTM in Wearable Multi-sensor Systems”. In: *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, Apr. 2019, pp. 443–448.