Automatic sleep staging from pulse oximeter using RNN

Ramiro Casal
Lab. of Signals and Nonlinear Dynamics
IBB - CONICET
Oro Verde, Argentina
rcasal@conicet.gov.ar

Leandro E. Di Persia
sinc(i) Research Institute
UNL - CONICET
Santa Fe, Argentina
ldipersia@sinc.unl.edu.ar

Gastón Schlotthauer
Lab. of Signals and Nonlinear Dynamics
IBB - CONICET
Oro Verde, Argentina
gschlotthauer@conicet.gov.ar

***Abstract*—The sleep stage scoring allows the analysis and characterization of several sleep disorders. Since manual labeling is a tedious task subject to human errors, many proposals to perform this classification automatically have been made. Methods based on electroencephalogram (EEG) are the gold-standard, achieving the best results. However, they have complex instrumentation, which is a disadvantage for screening methods. For this reason, we propose an automatic sleep staging method using heart rate (HR) and peripheral oxygen saturation (SpO_2) signals obtained from pulse oximeter, an ideal device for screening due to its low cost and simplicity. This method consists of two stacked layers of bidirectional gated recurrent units and a softmax layer to classify the output according to the American Academy of Sleep Medicine. To evaluate the performance, we use the Sleep Heart Health Study dataset, using 2500 HR and SpO_2 signals corresponding to different patients for training, 1250 for validation, and 1250 for testing the models. The obtained results in the testing subset were 73.2% for accuracy and 0.63 for the Cohen's Kappa coefficient. This performance shows that our model is able to outperform alternative methods that use cardiac signals from both pulse oximeter and electrocardiogram, but there is still an important gap to achieve the performances obtained using EEG.**

***Index Terms*—pulse oximeter, heart rate, recurrent neural networks, automatic sleep staging**

I. Introduction

The gold standard to assess sleep disorders is polysomnography (PSG), which consists of the recording of several biological signals including electroencephalography (EEG), electrooculography (EOG), oxygen saturation, chin and leg electromyography (EMG), electrocardiography (ECG), breathing effort, among others [1], [2]. The PSG is expensive and its availability is scarce. The visual data interpretation of the PSG signals is the most common approach to diagnose, but the scoring is a time-consuming process and depends on the expert's experience. Further, it has a lot of variability among different professionals [3].

There are two available standards that represent a guideline for diagnosing sleep pathologies, the traditional Rechtschaffen and Kales (R&K) [4] and, since 2007, the later standard published by the American Academy of Sleep Medicine (AASM) [2]. According to the R&K standard, the PSG recordings are split into consecutive 30-seconds-long segments and each segment is classified in wakefulness (W), two stages of light sleep (N1 and N2), two of deep sleep (N3 and N4), and rapid eye movement sleep (REM), which are differentiated based on characteristic waveforms that can be found in EEG, EOG and EMG [4], [5]. The AASM modifies the R&K rules with the aim of increasing the inter-rater reliability of sleep staging, unifying N3 and N4 in a single stage, called simply N3 or slow-wave sleep.

To overcome the disadvantages of visual inspection, numerous approaches for automatic sleep staging have been developed. Many studies have shown that the EEG signal is almost sufficient for obtaining a reliable scoring [6]. In addition to this, due to the complexity of the PSG studies, there is a generalized underdiagnosis of several sleep pathologies [7], [8]. Hence, the development of portable screening methods to assess sleep studies at home is becoming particularly relevant in the research community.

Pulse oximeter is an ideal choice for screening due to its low cost, accessibility, and simplicity [9]. This device has proven to be useful for the screening of obstructive sleep apnea/hypopnea syndrome [10], one of the most prevalent sleep disorders [11]. In addition, this technology can easily be adapted to wearable devices and be used for personal health monitoring. There are many commercial devices that provide sleep measures, but not many studies have validated these measures [12].

The main hypothesis of this work is that information about sleep stages can be inferred by means of the adequate processing of cardiac-related signals. The heart rate (HR) signal is affected by the regulation of the autonomic nervous (sympathetic and parasympathetic balance), decreasing during sleep to adapt to reduced metabolism. The average HR falls steadily from wake to deep sleep stages, increasing slightly

This work was partially supported by the National Agency for Scientific and Technological Promotion (ANPCyT) under projects PICT 2014-2627 and 2015-0977, Universidad Nacional de Entre Ríos and the National Council on Scientific and Technical Research (CONICET) under projects PID-UNER 6171, PIO-UNER-CONICET 146-201401-00014-CO, and Universidad Nacional del Litoral under projects CAI+D 50020150100059LI and 50020150100082LI.

This work used computational resources from High Performance Computing Center (CCAD) - Universidad Nacional de Córdoba (http://ccad.unc.edu.ar/), in particular the Nabucodonosor Cluster, which is part of National High Performance Computing System (SNCAD) - MinCyT, República Argentina.

Further, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the two Titan Xp GPU used for this research.

during REM [5]. Further, HR presents greater variability during wakefulness and REM. We expect that the proposed algorithm will be able to exploit these changes to perform the automatic sleep stage classification.

This paper is a preliminary study to classify sleep stages using signals provided by the pulse oximeter, namely HR estimated from photoplethysmography and peripheral oxygen saturation (SpO_2). These developments are derived from the modification of a previous work in which the sleep stages were classified as awake and sleep [13]. The classification is performed by applying recurrent neural networks (RNNs) to the raw signals. The RNNs are able to store information about the entire sequence in the state vectors to learn the temporal dependencies of the internal structure of the sleep [14]. We suppose that the network will be able to discriminate the sleep stages based on the changes in the HR dynamic caused by the different regulation of the autonomic nervous system during sleep [5].

II. Materials and Methods

The automatic sleep staging is performed using a specific type of RNN called gated recurrent networks (GRU) [15], which is a simplified variant of the well-known long short-term memories (LSTM) [16]. The proposed approach consists of a simple preprocessing of the raw data with the aim of removing invalid data and standardize the inputs to the network. Then, two stacked layers of bidirectional GRUs receive the input preprocessed data to learn the transition rules of the sleep stages exploiting both past and future information [17]. Finally, a 5-softmax layer is used to classify each segment into five classes according to the AASM rules. An overview of these network architecture is shown in Fig. 1.

In this section we will explain briefly the used dataset, the different parts of the designed architecture, and the performed post-processing.

A. Dataset

In this work, we used 5000 recordings of HR and SpO_2 corresponding to the Sleep Heart Health Study (SHHS) dataset [18]. The SHHS dataset contains PSG studies acquired automatically at the patient's home. Full details can be found in [19].

The HR signal has a sampling rate of 1 Hz and precision of 3 beats per minute. Further, the oximeter provides a quality-related signal with information about the status of the sensor connection, the value of which is 0 for a good connection and 1 for a defective connection.

Additionally, we used the sleep staging included in the SHHS as the target of the system. These sleep stages are labeled in consecutive 30-seconds-long segments and they were processed to be in accordance with the AASM rules.

1) Input data preprocessing : We split the dataset into three subsets: 2500 subjects were used for training the network, while two subsets of 1250 subjects were used for validating and testing the trained models.

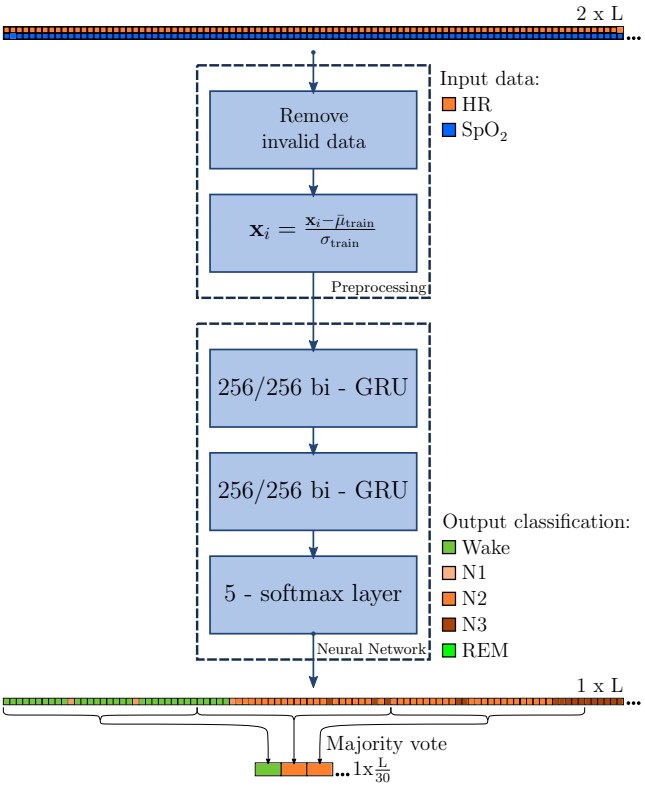

Figure 1. An scheme of the best architecture consisting of two stacked layers of bidirectional GRU, a softmax layer to classify the outputs of the GRUs, and finally a majority vote performed to obtain the final classification.

To remove invalid data due to defective sensor contact, we use the quality-related signal to mask the signals. Once identified the invalid data, we interpolate linearly between the previous and posterior valid data.

In order to reduce the inter-subject variability, we standardize the input data using the global mean and standard deviation of the training dataset. Then, these values were used to standardize the training, validation, and testing datasets.

2) Recurrent neural networks: RNNs are a family of neural networks that have proven to be very useful for processing sequential data. They can store in an "internal state" the information of the history of the signals. In this way, the outputs of the network do not depend only on the current input, but they also depend on the previous inputs and outputs. This is accomplished using recurrent connections that feedback the outputs into the inputs [14]. In classical RNNs, this theoretical persistence of the information is not easy to achieve in practice because the backpropagated gradients used to train the network tend to vanish rapidly [16].

To overcome this limitation, called vanishing gradient, the LSTMs networks were designed. In these networks, the information flow is controlled by structures called gates that allow learning the long-term dependencies introducing a persistent internal state present in each LSTM unit.

Many variations of the LSTM have been proposed since their emergence, but the GRUs [15] are a simplified version

of the LSTM that have become very popular. For time step t and GRU-cell i, the follow equations are used:

$$\begin{aligned}
\mathbf{u}_i^{(t)} &= \sigma\left(\mathbf{W}_{u,i}\left[\mathbf{h}_i^{(t-1)}, \mathbf{x}^{(t)}\right] + \mathbf{b}_{u,i}\right), \\
\mathbf{r}_i^{(t)} &= \sigma\left(\mathbf{W}_{r,i}\left[\mathbf{h}_i^{(t-1)}, \mathbf{x}^{(t)}\right] + \mathbf{b}_{r,i}\right), \\
\tilde{\mathbf{h}}_i^{(t)} &= \tanh\left(\mathbf{W}\left[\mathbf{r}_i^{(t)}\mathbf{h}_i^{(t-1)}, \mathbf{x}^{(t)}\right] + \mathbf{b}_{s,i}\right), \\
\mathbf{h}_i^{(t)} &= \left(1 - \mathbf{u}_i^{(t)}\right) \circ \mathbf{h}_i^{(t-1)} + \mathbf{u}_i^{(t)} \circ \tilde{\mathbf{h}}_i^{(t)}
\end{aligned} \tag{1}$$

where $\mathbf{u}$ is the "update" gate, $\mathbf{r}$ is the "reset" gate, and $\mathbf{h}_i^{(t)}$ is the state vector of the i-th GRU cell. $\mathbf{W}_{(\cdot),i}$ and $\mathbf{b}_{(\cdot),i}$ are the weights and bias, $\mathbf{x}$ represents the input to the network, and σ represents a sigmoid function [20]. Finally, the operator $\circ$ represents an element-wise product.

The update gate controls by means of the sigmoid function how much information from the last state vector $\mathbf{h}_i^{(t-1)}$ and how much information from the new candidate of state vector $\tilde{\mathbf{h}}_i^{(t)}$ are used for the new state vector $\mathbf{h}_i^{(t)}$. The reset gate controls which parts of the current state are used to compute the next state [20].

As can be seen from (1), the state vector stores only information from past and present inputs, i.e. it has a causal behavior. When the processing is off-line, we prefer to be able to extract information not only from the past but also from the future, which allows a better understanding of the context and can eliminate ambiguities. Schuster and Paliwal [17] created a bidirectional RNN combining two RNNs, one that moves forward through time and the other that moves backward.

In our work, we have used GRU-based architectures instead of LSTM since they make less use of memory. Moreover, since the processing in our case is done off-line, bidirectional GRUs were considered.

B. Softmax layer

With the aim of classifying the outputs of the GRU, a 5-*softmax* layer is applied by:

$$\mathbf{y} = \text{relu}(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{2}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the weights and bias, respectively, and $\mathbf{x}$ is the input to this layer, that is the output of the second bidirectional GRU. We use a rectified linear unit activation $\text{relu}(x) = \max(0, x)$. This output vector y is mapped to a class probability with a *softmax* function. The used loss function is cross-entropy and the optimization algorithm is Adam [21].

The approach presented in this paper performs a sleep stage classification sample to sample, that is, with a resolution of 1 second. According to AASM, which recommends labeling the sleep stages every 30 seconds, we conducted a majority vote for non-overlapping segments of 30 seconds. The reported results correspond to this vote.

III. Results and discussion

We evaluated many variants of network architectures, changing the number of GRUs stacked and the number of hidden layer sizes in the bidirectional-GRUs. From these experiments, the architecture with the best performance was two stacked bidirectional-GRUs with 256 hidden units.

As we previously stated, we have used the Adam optimizer. The parameters of which are the learning rate α and the exponential decay rates for the first moment and second moment, β_1 and β_2 respectively. These parameters were set to 10^{-4}, 0.9, and 0.99 respectively.

We trained the model using the training dataset during 120 epochs. After each epoch, the model was evaluated using the validation dataset. To avoid over-fitting, we used early stopping, selecting the model with the best accuracy in the validation dataset.

A. Sleep staging performance

The performance of the designed network was evaluated in the test dataset, composed of 1250 signals corresponding to unseen patients.

Sleep stages can be classified with different degrees of discrimination. Thus, methods designed for screening have a tendency to group several sleep stages in order to perform a simpler classification. With the aim of favor the methods comparison, we present several tables with the obtained results considering different sleep stage groupings and the main algorithms of the state-of-the-art which performed the same grouping. It should be noted that, as different signals and datasets were used, direct comparisons between results cannot be performed.

Table I presents the performance classifying sleep stages according to the AASM, namely awake, N1, N2, N3, and REM. As far as we know, there are no screening algorithms that classify sleep stages according to the AASM. For this reason, it was necessary to compare against methods that use EEG. Supratak et al. [22] used an architecture based on convolutional neural networks (CNN, to automatically learn features from EEG signals) and RNN (to learn the sleep stage transition rules), the performance of which was evaluated in two different datasets. We report both results in order to show the variability of the results for the same algorithm. Mousavi et al. [23] developed a method based on CNN and RNN using a single-EEG channel. Unlike Supratak, they used an encoder-decoder architecture and attention mechanisms. To the best of our knowledge, that work obtains the best performance in automatic sleep staging using a single-EEG channel. While we cannot yet compete with these results, we must say that EEG signals have complex instrumentation. These results give us an idea of the upper bound that we could reach, and allow us to compare against the gold-standard.

Table II summarized the state-of-the-art methods grouping N1 and N2. We compared our obtained results with Beattie et al. [24], which used photoplethysmography (PPG) and accelerometer signals. In that work, the authors considered 4 classes. The used database was composed of 60 participants that were self-reported normal sleepers. We can not make a direct comparison because they have additional information. They used the PPG signal (not only the HR calculated from

Table I
COMPARISON WITH THE LITERATURE (AASM RULES)

Method	Input	N	Acc	Acc_W	Acc_{N1}	Acc_{N2}	Acc_{N3}	Acc_{REM}	κ
256-biGRU	HR+SpO_2	5000	73.2	85.6	0	75.7	60.8	75.4	0.63
CNN+RNN [22]	EEG	62	86.2	87.3	59.8	90.3	81.5	89.3	0.80
CNN+RNN [22]	EEG	20	82.0	84.7	46.6	85.9	84.8	82.4	0.76
CNN + RNN [23]	EEG	61	84.3	89.2	52.2	86.8	85.1	85.0	0.79

it), in addition to the accelerometer signals. In spite of this, it can be seen that our algorithm obtains a better performance than that work, even when we are discriminating between N1 and N2. The best accuracy obtained in [24] was 68.7%.

Finally, table III shows the performance classifying sleep stages as W, REM, and non-REM (grouping N1, N2, and N3). Here we can compare with Yücelbaş et al. [25] and Xiao et al. [26], both using ECG. Yücelbaş used two different datasets, which contain 10 and 18 signals corresponding to different patients, respectively. The results were reported for healthy subjects and patients (people suffering from sleep disease). In the second dataset, the reported data do not allow to compare all performance measures. The authors performed 10-fold cross-validation, but segments corresponding to the same patient were used for training and testing. Ideally, signals from patients that are used for training should not be used for testing. The work by Xiao et al. [26] extracted 41 features from ECG and used random forests to classify. The authors only analyzed data labeled with "stationary", that is they classified 5-minute windows corresponding to a single class. The accuracy obtained was 75.6%. The authors performed two different schemes, subject-specific classifier (training and testing set with the same record) and subject independent classifier (training and testing with independent records). To be fair, we only compared our result with the best result in the independent scheme obtained by Xiao et al [26].

Based on the obtained results, we can see that acceptable performance is achieved. The most noticeable is the inability of the algorithm to classify N1, a very minor stage in relation to the rest (approximately 3% of the total-register-time).

Fig. 2 shows three hypnograms obtained using the designed network: the first hypnogram represents an average error, and the second and the third hypnograms obtained a performance higher and lower than the average, respectively.

IV. CONCLUSION

In this paper, we proposed an RNN-based model for classifying sleep stages using raw HR and SpO_2 obtained from a pulse oximeter. Our model uses bidirectional-GRUs to learn the transition rules among the sleep stages. It was shown that the HR and SpO_2 dynamic is useful to perform sleep staging. As far as we know, with exception of EEG, this research shows better results than the others research in the field that used signals that were harder to register in the same field. Further, this approach can be easily adapted to screening devices of sleep pathologies, wearable devices for personal health monitoring, among others. The size of the dataset used is bigger than in other related works. As future work, we will improve this network and try other architectures with the aim to obtain results comparable with the EEG.

REFERENCES

[1] R. K. Malhotra and A. Y. Avidan, "Chapter 3 - sleep stages and scoring technique," in *Atlas of Sleep Medicine (Second Edition)*, second edition ed., S. Chokroverty and R. J. Thomas, Eds. St. Louis: W.B. Saunders, 2014, pp. 77 – 99. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9781455712670000035

[2] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, and B. Vaughn, "The AASM manual for the scoring of sleep and associated events," *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 2012.

[3] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset." *Sleep*, vol. 23, no. 7, pp. 901–908, 2000.

[4] A. Rechtschaffen, "A manual of standardized terminology, technique and scoring system for sleep stages of human subjects," *Public Health Service*, 1968.

[5] T. Penzel, J. W. Kantelhardt, L. Chung-Chang, K. Voigt, and C. Vogelmeier, "Dynamics of heart rate and sleep stages in normals and patients with sleep apnea," *Neuropsychopharmacology*, vol. 28, no. S1, p. S48, 2003.

[6] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals," *Computer methods and programs in biomedicine*, vol. 140, pp. 77–91, 2017.

[7] S. Ram, H. Seirawan, S. K. Kumar, and G. T. Clark, "Prevalence and impact of sleep disorders and sleep habits in the united states," *Sleep and Breathing*, vol. 14, no. 1, pp. 63–70, 2010.

[8] C. Fuhrman, B. Fleury, X.-L. Nguyên, and M.-C. Delmas, "Symptoms of sleep apnea syndrome: high prevalence and underdiagnosis in the french population," *Sleep medicine*, vol. 13, no. 7, pp. 852–858, 2012.

[9] K. P. Pang and D. J. Terris, "Screening for obstructive sleep apnea: an evidence-based analysis," *American Journal of Otolaryngology*, vol. 27, no. 2, pp. 112–118, 2006.

[10] G. Schlotthauer, L. E. Di Persia, L. D. Larrateguy, and D. H. Milone, "Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry," *Medical Engineering & Physics*, vol. 36, no. 8, pp. 1074–1080, 2014.

[11] M. J. Sateia, "International classification of sleep disorders: highlights and modifications," *Chest Journal*, vol. 146, no. 5, pp. 1387–1394, 2014.

[12] J. Mantua, N. Gravel, and R. Spencer, "Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography," *Sensors*, vol. 16, no. 5, p. 646, 2016.

[13] R. Casal, L. E. Di Persia, and G. Schlotthauer, "Classifying sleep–wake stages through recurrent neural networks using pulse oximetry signals," *Biomedical Signal Processing and Control*, vol. 63, p. 102195, 2021.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[15] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

Table II
COMPARISON WITH THE LITERATURE (W, LIGHT, DEEP, AND REM)

Method	Inputs	N	Acc	Acc_W	Acc_{Light}	Acc_{Deep}	Acc_{REM}	κ
256-biGRU	HR+SpO_2	5000	75.7	85.6	73.7	60.1	75.4	0.65
Feat.+LDA [24]	PPG + acc	60	68.7	69.3	69.2	62.5	71.6	0.52

Table III
COMPARISON WITH THE LITERATURE (W, NREM, AND REM). P1 AND H1 ARE THE PATIENT AND HEALTHY SUBSET OF THE FIRST DATASET. ANALOGOUSLY FOR P2 AND H2.

Method	Inputs	N	Acc	Acc_W	Acc_{NREM}	Acc_{REM}	κ
256-biGRU	HR+SpO_2	5000	85.2	85.6	87.5	75.4	0.74
Feat.+RF [25]	ECG (P1)	5	78.1	74.3	82.8	40.9	0.57
	ECG (H1)	5	87.1	83.5	91.1	52.1	0.74
	ECG (P2)	16	77.0	–	–	–	0.57
	ECG (H2)	2	76.8	–	–	–	0.43
Feat.+RF [26]	ECG	45	72.6	56.4	81.3	59.8	0.46

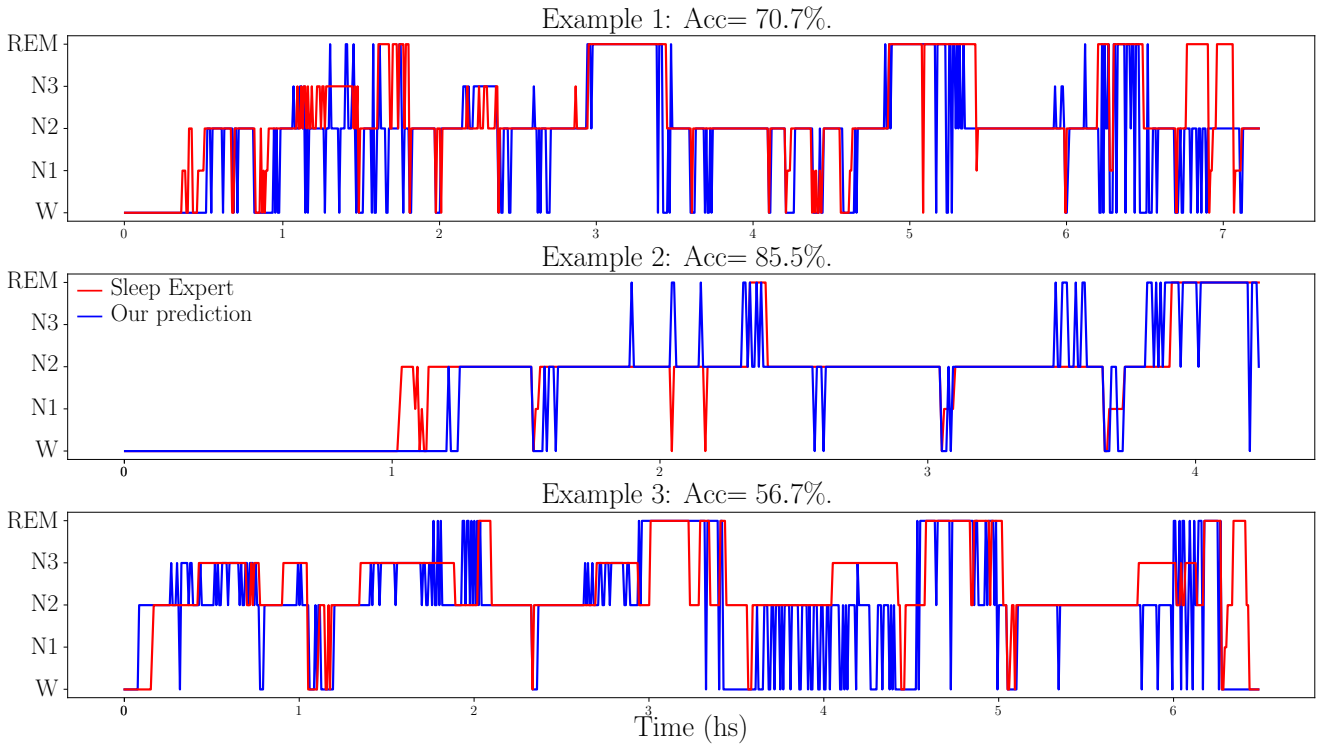

Figure 2. Obtained hypnograms according to AASM rules.

[18] S. Redline, M. H. Sanders, B. K. Lind, S. F. Quan, C. Iber, D. J. Gottlieb, W. H. Bonekat, D. M. Rapoport, P. L. Smith, J. P. Kiley *et al.*, "Methods for obtaining and analyzing unattended polysomnography data for a multicenter study," *Sleep*, vol. 21, no. 7, pp. 759–768, 1998.

[19] E. J. Nieto, G. T. O'Connor, D. M. Rapoport, and S. Redline, "The sleep heart health study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[20] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[23] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PloS one*, vol. 14, no. 5, p. e0216456, 2019.

[24] Z. Beattie, Y. Oyang, A. Statan, A. Ghoreyshi, A. Pantelopoulos, A. Russell, and C. Heneghan, "Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals," *Physiological Measurement*, vol. 38, no. 11, p. 1968, 2017.

[25] Ş. Yücelbaş, C. Yücelbaş, G. Tezel, S. Özşen, and Ş. Yosunkaya, "Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal," *Expert Systems with Applications*, vol. 102, pp. 193–206, 2018.

[26] M. Xiao, H. Yan, J. Song, Y. Yang, and X. Yang, "Sleep stages classification based on heart rate variability and random forest," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 624–633, 2013.