Fusion of End-to-End Deep Learning Models for Sequence-to-Sequence Sleep Staging

Huy Phan*, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos

***Abstract*— Sleep staging, a process of identifying the sleep stages associated with polysomnography (PSG) epochs, plays an important role in sleep monitoring and diagnosing sleep disorders. We present in this work a model fusion approach to automate this task. The fusion model is composed of two base sleep-stage classifiers, SeqSleepNet and DeepSleepNet, both of which are state-of-the-art end-to-end deep learning models complying to the sequence-to-sequence sleep staging scheme. In addition, in the light of ensemble methods, we reason and demonstrate that these two networks form a good ensemble of models due to their high diversity. Experiments show that the fusion approach is able to preserve the strength of the base networks in the fusion model, leading to consistent performance gains over the two base networks. The fusion model obtain the best modelling results we have observed so far on the Montreal Archive of Sleep Studies (MASS) dataset with 200 subjects, achieving an overall accuracy of 88.0%, a macro F1-score of 84.3%, and a Cohen's kappa of 0.828.**

I. INTRODUCTION

Recent years have seen an explosive amount of sleep data. This offers additional resources and opens new doors to leverage deep learning algorithms to reduce the performance gap between automatic sleep staging systems and sleep experts' manual scoring. Significant performance improvements by deep learning algorithms in sleep stage classification have been recently reported on datasets obtained from hundreds [1] to thousands of subjects [2]. These results demonstrate the potential for automatic algorithms to replace, or at least assist, sleep experts in the sleep scoring task. As a result, the automated algorithms can help to ease the manual task, improve the accuracy of the diagnosis and assessment of sleep disorders [3], and scale sleep monitoring to benefit a lot of people in need [4], [5].

During the last few years, advances of deep learning have benefited the automatic sleep staging problem in various ways. First, deep neural network's capability in learning good features directly from raw signals has outdated hand-crafted features and liberated us from designing a handful of these features. Autoencoders [6], deep neural networks (DNNs) [7], convolutional neural networks (CNNs) [8], [9], [10], [11], and recurrent neural networks (RNNs) [12], [13], [14] are useful for this purpose. Second, they enable us to look for new classification schemes, which are impossible under more conventional machine learning paradigm, that take into account properties of the input signals. To this end, one-to-many [15] and many-to-many classification schemes [1] were recently introduced to enhance the efficiency of a deep learning model in encoding the sequential dependency of the sleep signals. Particularly, under the many-to-many scheme, the sleep staging was re-formulated as a sequence-to-sequence classification problem and deep learning models following this scheme were recently reported to significantly outperform existing methods on benchmarking datasets with hundreds to thousands of subjects [1], [2], [16].

While building more efficient network architectures and better classification schemes are active research topics in this field, developing ensemble methods for automatic sleep staging in the deep learning context has been left uncharted. However, ensembles of learned models [17], [18] is a well-established method in machine learning and in statistical science (in which it is termed *meta-analysis* [19]), allowing us to construct a fusion model which is better than its individual base models in general. They have found to work well for the automatic sleep staging task when more conventional methods are used as model bases [20], [21]. In this work, we study ensemble methods for the task under the deep learning prism. Specifically, we propose a fusion model composing of two deep network bases, namely *SeqSleepNet* [1] and *DeepSleepNet* [16]. These two networks comply to the sequence-to-sequence classification scheme and were trained end-to-end with the training strategies proposed in [1]. Given the ensemble methods' criteria, we also ask and answer the question why these two networks should construct a good cohort for fusion purpose. Empirical results on the MASS dataset reveal that the proposed fusion model is able to leverage the advantages of the deep network bases to achieve consistently better results than those of the individual models.

II. MONTREAL ARCHIVE OF SLEEP STUDIES (MASS) DATASET

We employed the public dataset Montreal Archive of Sleep Studies (MASS) [22] in this study. This dataset consists of whole-night recordings from 200 subjects aged between 18 and 76 years (97 males and 103 females). Manual annotation was done on each epoch of the recordings by sleep experts according to the AASM standard [23] (SS1 and SS3 subsets) or the R&K standard [24] (SS2, SS4, and SS5 subsets). As in [15], [1], we converted different annotations into five sleep stages {W, N1, N2, N3, and REM} and expanded 20-

HP is with the School of Computing, University of Kent, Chatham Maritime ME4 4AG, UK. OYC and MDV are with the Institute of Biomedical Engineering, University of Oxford, Oxford OX3 7DQ, UK. PK and AM are with the Institute of Signal Processing, University of Lübeck, Lübeck 23562, Germany.

*Corresponding author: h.phan@kent.ac.uk

second epochs into 30-second ones by including 5-second segments before and after each epoch. We adopted and studied combinations of an EEG channel (C4-A1), an EOG channel (ROC-LOC), and an EMG channel (CHIN1-CHIN2) in our experiments. The signals, originally sampled at 256 Hz, were downsampled to 100 Hz.

III. End-to-end deep learning models for sequence-to-sequence sleep staging

Sequence-to-sequence sleep staging scheme was recently proposed to improve encoding performance of long-term temporal dependencies of PSG epochs in a deep learning model [1]. Intuitively, given a sequence of consecutive PSG epochs, a sequence-to-sequence classification model aims to classify all the epochs at once. Formally, the sequence-to-sequence sleep staging problem [1] is formulated to maximize the conditional probability $p(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L \,|\, \mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_L)$, where $(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_L)$ denote the input sequence of L consecutive epochs and $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L)$ represent the sequence of corresponding L one-hot encoding vectors of the ground-truth output labels.

Fig. 1 illustrates a network architecture proposal for sequence-to-sequence sleep staging. The epoch processing block (EPB) is to extract a feature vector $\mathbf{x}_i$ to represent each epoch $\mathbf{S}_i$, $1 \leq i \leq L$. Furthermore, the EPB should be the same, i.e. being shared, for all epochs and preferably be a sub-network which can be trained jointly in an end-to-end fashion [1]. Afterwards, a bidirectional recurrent neural network (biRNN) reads the induced sequence of feature vectors forward and backward to encode their sequential interactions into a sequence of output vectors $(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_L)$. The sequence of output vectors is subsequently classified by a softmax layer to produce the output sequence of sleep stage probabilities $(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_L)$. Such an end-to-end network is trained to minimize the sequence classification loss [1] over N training sequences in the training data:

$$E(\boldsymbol{\theta}) = -\frac{1}{L}\sum_{n=1}^{N}\sum_{i=1}^{L}\mathbf{y}_i \log\left(\hat{\mathbf{y}}_i\left(\boldsymbol{\theta}\right)\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2, \qquad (1)$$

where $\boldsymbol{\theta}$ represents the network parameters and λ denotes the hyper-parameter that trades off the error terms and the ℓ_2-norm regularization term.

A. SeqSleepNet

SeqSleepNet was presented in [1] to deal with sequence-to-sequence sleep staging. The network makes use of time-frequency representations as input. The EEG, EOG, and EMG signals are transformed into log-power spectra via short-time Fourier transform (STFT) with a window size of two seconds and 50% overlap, followed by logarithmic scaling. Hamming window and 256-point Fast Fourier Transform (FFT) were used. This results in a 3-channel image $\mathbf{S} \in \mathbb{R}^{F \times T \times C}$ where $F = 129$ (the number of frequency bins), $T = 29$ (the number of spectral columns), and $C = 3$ (the number of channels).

In SeqSleepNet, the EPB is the combination of three *filterbank* layers [10], a two-layer biRNN realized by Gated Recurrent Unit (GRU) cell [25], and an attention layer [26] as illustrated in Fig. 2(a). Note that this epoch-level biRNN should not be confused with the sequence-level biRNN in Fig. 1. Each of the filterbank layers with 32 filters is firstly used to preprocess one input image channels. Afterwards, the resulting image channels are stacked in the frequency dimension to form a single image. The biRNN then treat the image as a sequence of T local feature vectors (image columns) $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)$ and encodes this sequence into a sequence of output vectors $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_T)$. These output vectors collectively form the epoch-wise feature vector $\mathbf{x}$ for an epoch $\mathbf{S}$:

$$\mathbf{x} = \sum_{t=1}^{T} \alpha_t \mathbf{a}_t, \qquad (2)$$

where α_t is the attention weight obtained via the attention layer [26] at the image column index t.

Concerning the sequence-level biRNN in Fig. 1, it is

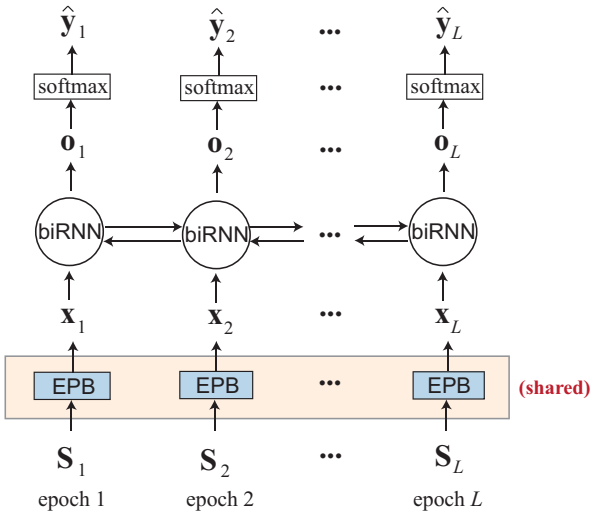

Fig. 1: Network architecture proposal for sequence-to-sequence sleep staging. The epoch processing block (EPB) plays the role of epoch-wise feature extractor and is the same, i.e. being shared, by all epochs.

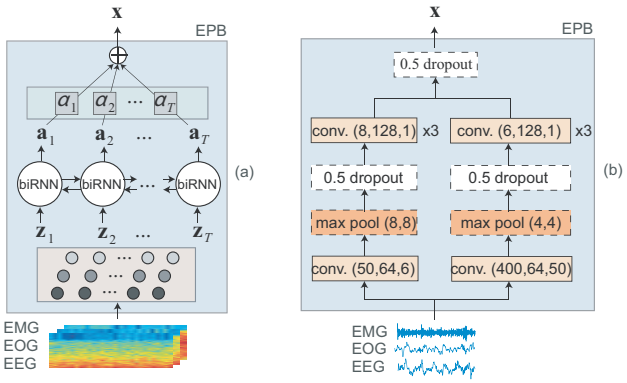

Fig. 2: Illustration of the EPBs of (a) SeqSleepNet and (b) DeepSleepNet. The former rely on an attentional biRNN coupled with preprocessing filterbank layers. The latter is a two-branch deep CNN.

implemented using GRU cell. Further details regarding the network's parameters can be found in [1].

B. DeepSleepNet

DeepSleepNet was proposed in [16] and its end-to-end variant was also presented in [1]. Here, at the network's input layer, the raw EEG, EOG, and EMG signals are stacked to form 3-channel input. The EPB is realized by a deep CNN sub-network as illustrated in Fig. 2(b). The CNN sub-network comprises two branches with 4 convolutional layers each. In Fig. 2(b), *conv. (n,w,s)* denotes a convolutional layer with *n* 1-D filters of size *w* and stride *s*. *max pool. (w,s)* denotes a 1-D max pooling layer with kernel size *w* and stride *s*. The convolutional kernels are designed to have different sizes in the two branches to be able to learn features at both fine and coarse temporal resolutions. Different from SeqSleepNet, the sequence-level biRNN is designed to have two layers with Long Short-Term Memory (LSTM) cells [27]. A residual connection is also exploited to combine the EPB's convolutional features $\boldsymbol{x}$ with the sequential output $\boldsymbol{o}$ of the sequence-level biRNN before classification takes place by the softmax layer [16].

C. Fusion of SeqSleepNet and DeepSleepNet

It was shown that SeqSleepNet outperformed DeepSleepNet on the MASS dataset [1]. However, the improvement was not evenly distributed over all sleep stages. Inspection on class-wise performance reveals that while SeqSleepNet worked better for N1 and REM, DeepSleepNet was favourable for N3. It is therefore natural to ask the question whether we can combine SeqSleepNet and DeepSleepNet in the way that they could compensate each other and collectively derive a better model.

Because of their respective advantages, we conjecture that SeqSleepNet and DeepSleepNet collectively form a good ensemble for model fusion. In the perspective of ensemble methods, model fusion allows us to achieve an accuracy which is often higher than those of single models. It was shown, theoretically and experimentally, that in order for a fusion model to be effective, the base classifiers should be (i) high-accuracy and (ii) diversified [28]. For the first criterion, adhering to the sequence-to-sequence classification scheme, both SeqSleepNet and DeepSleepNet have been recently shown to be highly accurate for the automatic sleep staging task, significantly outperforming those relying on other classification schemes [1]. For the second criterion, SeqSleepNet differ significantly from DeepSleepNet. First, at the input layer, they receive different signal types as inputs, i.e. time-frequency features with SeqSleepNet and raw signals with DeepSleepNet. Second, at the epoch processing block, SeqSleepNet employs an attentional RNN combined with filterbank layers as an epoch-wise feature learning engine whereas this component is operated by a deep CNN in DeepSleepNet. Third, at the sequence processing level, SeqSleepNet makes use of the GRU cell to implement the bidirectional RNN for sequence modelling while DeepSleepNet exploited the LSTM cell for this purpose.

Here, we employ a late fusion method in which the probabilistically multiplicative aggregation scheme is used to fuse decisions coming from two network bases. Moreover, since both SeqSleepNet and DeepSleepNet are multiple-output networks, one may shift the input sequence of length L by one epoch as in [1] when evaluating on a test recording to obtain L decisions at every epoch (except those at the recording's ends). The likelihood of a sleep stage $y_i \in \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ at an epoch index i after model fusion is given by

$$\mathcal{L}(y_i) = \frac{1}{L} \prod_{j=i-L+1}^{i} P_1(y_i \,|\, \mathcal{S}_j) P_2(y_i \,|\, \mathcal{S}_j). \tag{3}$$

where $\mathcal{S}_j = (\mathbf{S}_j, \mathbf{S}_{j+1}, \ldots, \mathbf{S}_{L-1})$ is the epoch sequence starting at index j. P_1 and P_2 represent the classification probabilities outputted by SeqSleepNet and DeepSleepNet, respectively. When the number of decisions involved in (3) is large, the aggregation should be conducted in the logarithmic domain to avoid possible numerical problems. In the logarithmic domain, the equation (3) can be re-written as

$$\log \mathcal{L}(y_i) = \frac{1}{L} \sum_{j=i-L+1}^{i} (\log P_1(y_i \,|\, \mathcal{S}_j) + \log P_2(y_i \,|\, \mathcal{S}_j)). \tag{4}$$

Subsequently, the output label $\hat{y}_i$ at epoch index i is determined by log-likelihood maximization:

$$\hat{y}_i = \underset{y_i}{\operatorname{argmax}} \log \mathcal{L}(y_i) \text{ for } y_i \in \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}. \tag{5}$$

IV. Experiments

A. Experimental setup

We conducted experiments on the MASS dataset via 20-fold cross-validation. At each iteration, 180, 10, and 10 subjects were employed for training, validation, and testing, respectively. During training, a network, i.e. SeqSleepNet and DeepSleepNet, was validated on the validation set after every 100 training steps and the one that yielded the best overall accuracy was retained for evaluation. The outputs of 20 cross-validation folds were pooled and considered as a whole for computing the sleep staging performance.

B. Network parameters

We experimented with different sequence length $L = \{10, 20, 30\}$ PSG epochs, equivalent to $\{5, 10, 15\}$ minutes. The sequences were sampled from the PSG recordings with a maximum overlapping (i.e. $L-1$ epochs). In this way, we generated all possible sequences from the training recordings for network training purpose.

Both SeqSleepNet and DeepSleepNet were implemented using *TensorFlow* framework [29]. The networks were parametrized similar to those in our previous work [1]. They were trained for 10 training epochs with a minibatch size of 32 sequences. The network training was performed using *Adam* optimizer [30] with a learning rate of 10^{-4}.

TABLE I: Performance obtained by SeqSleepNet, DeepSleepNet, and their fusion model on the MASS dataset.

System	Seq. length	Overall metrics					Class-wise sensitivity					Class-wise selectivity				
		Acc.	κ	MF1	Sens.	Spec.	W	N1	N2	N3	REM	W	N1	N2	N3	REM
SeqSleepNet	10	87.0	0.814	83.2	82.4	96.2	88.6	**59.9**	91.2	79.4	93.0	**91.3**	64.9	88.6	**85.1**	90.2
DeepSleepNet		86.3	0.804	82.0	81.6	96.1	88.4	55.6	90.3	**83.4**	90.6	88.8	62.0	89.0	82.3	90.2
Fusion		**87.9**	**0.827**	**84.2**	**83.5**	**96.5**	**90.3**	59.7	**91.9**	82.4	**93.4**	91.1	**68.4**	**89.5**	84.6	**91.8**
SeqSleepNet	20	87.0	0.815	83.3	82.8	96.3	89.4	**60.8**	90.7	80.3	92.9	90.0	65.1	89.1	84.0	90.8
DeepSleepNet		86.2	0.804	82.2	82.0	96.1	88.4	57.0	89.9	**84.1**	90.4	89.0	62.1	89.0	81.1	91.2
Fusion		**87.9**	**0.827**	**84.3**	**83.7**	**96.5**	**90.0**	**60.8**	**91.7**	82.9	**93.1**	**91.3**	**67.9**	**89.7**	**84.3**	**91.8**
SeqSleepNet	30	87.1	0.815	83.3	82.7	96.2	89.0	59.7	90.9	80.2	93.5	90.7	65.1	88.9	84.2	90.7
DeepSleepNet		86.4	0.805	82.2	81.8	96.1	*89.2*	55.8	90.5	**83.1**	90.3	88.8	62.6	88.8	82.0	91.1
Fusion		**88.0**	**0.828**	**84.3**	**83.8**	**96.5**	**89.9**	**59.9**	**92.1**	82.0	**93.5**	**91.5**	**68.6**	**89.5**	**85.2**	**91.2**

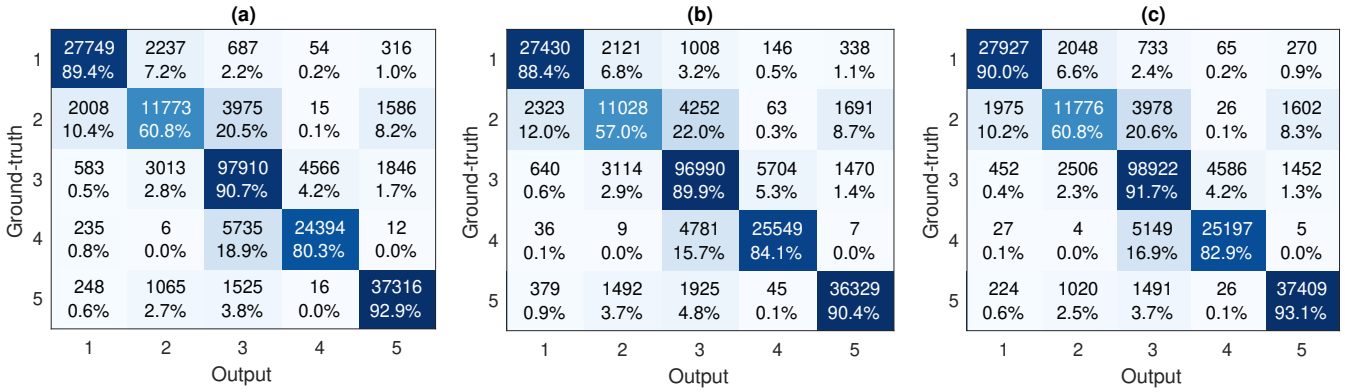

Fig. 3: Confusion matrix: (a) SeqSleepnet, (b) DeepSleepNet, and (c) the fusion model. the fusion model obtains better performances for all sleep stages, except for N3 whose performance seems to be the average of SeqSleepNet and DeepSleepNet.

C. Experimental results

We show in Table I the performances obtained by SeqSleepNet, DeepSleepNet, and their model fusion on the MASS dataset. The overall performance is reported using accuracy, macro F1-score (MF1), Cohen's kappa (κ), sensitivity, and specificity. In addition, class-specific performance is also assessed via sensitivity and selectivity as recommended in [31]. Note that in this work we only focus on comparing the proposed fusion model with its base networks (i.e. SeqSleepNet and DeepSleepNet). A comprehensive performance comparison between these networks with other methods can be found in [1].

As Table I show the fusion model yields good performance and consistently outperforms its model bases over all the sequence lengths. Taking $L = 20$ for example, the fusion model obtains an accuracy of 87.9%, an F1-score of 84.3%, and a κ value of 0.827. This performance improves that of SeqSleepNet by 0.9%, 1.0%, and 0.013 absolute in terms of overall accuracy, macro F1-score, and κ, respectively. The respective gains over the DeepSleepNet are even more noticeable, reaching 1.7%, 1.9%, and 0.023 absolute.

The effects of model fusion on individual sleep stages are also elucidated by class-wise results in Table I and further shown by the confusion matrices in Fig. 3. More specifically, model fusion leads to better performances for all sleep stages, except for N3 whose performance seems to be the average of the two network bases. Interestingly, the fusion model is able to preserve the strength of SeqSleepNet to compensate the weakness of DeepSleepNet in recognizing N1 and REM. This result is potentially meaningful as accurately recognizing these sleep stages plays an important role in diagnosis and assessment of many types of sleep disorders, such as narcolepsy [2] and REM Sleep Behavior Disorder (RBD) [32].

V. Conclusions

This work investigated ensemble methods with deep learning models for automatic sleep staging. A fusion model was composed of two high-quality and diversifying end-to-end deep networks, SeqSleepNet and DeepSleepNet, which were designed for sequence-to-sequence sleep staging. The experimental results showed that the fusion model consistently outperformed both of its high-end network bases, not only on overall performance but also on most of class-specific results, particularly those clinical-relevant sleep stages. These preliminary results suggest that ensemble methods with deep neural networks have the potential to further improve accuracy on the automatic sleep staging task which may not be achieved easily with a single network model.

REFERENCES

[1] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering (TNSRE)*, vol. 27, no. 3, pp. 400–410, 2019.

[2] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. D. Sorensen, P. Jennum, and E. Mignot, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications*, vol. 9, no. 1, pp. 5229, 2018.

[3] A. C. Krieger, Ed., *Social and Economic Dimensions of Sleep Disorders, An Issue of Sleep Medicine Clinics*, Elsevier, 2017.

[4] K. B. Mikkelsen, J. K. Ebajemito, M. A. Bonmati-Carrion, N. Santhi, V. L. Revell, G. Atzori, C. della Monica, S. Debener, D.-J. Dijk, A. Sterr, and M. de Vos, "Machine-learning-derived sleep-wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy," *Journal of Sleep Research*, p. e12786, 2018.

[5] D. Looney, V. Goverdovsky, I. Rosenzweig, M. J. Morrell, and D. P. Mandic, "Wearable in-ear encephalography sensor for monitoring sleep. preliminary observations from nap studies," *Annals of the American Thoracic Society*, vol. 13, no. 12, pp. 32–42, 2016.

[6] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Annals of Biomedical Engineering*, vol. 44, no. 5, pp. 1587–1597, 2016.

[7] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 324–333, 2018.

[8] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," *arXiv:1610.01683*, 2016.

[9] F. Andreotti, H. Phan, and M. De Vos, "Visualising convolutional neural network decisions in automatic sleep scoring," in *Proc. Joint Workshop on Artificial Intelligence in Health (AIH)*, 2018, pp. 70–81.

[10] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification," in *Proc. EMBC*, 2018, pp. 453–456.

[11] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[12] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks," in *Proc. EMBC*, 2018, pp. 1452–1455.

[13] P. Koch, H. Phan, M. Maass, F. Katzberg, and A. Mertins, "Recurrent neural network based early prediction of future hand movements," in *Proc. EMBC*, 2018, pp. 4710–4713.

[14] P. Koch, H. Phan, M. Maass, F. Katzberg, R. Mazur, and A. Mertins, "Recurrent neural networks with weighting loss for early prediction of hand movements," in *Proc. EUSIPCO*, 2018, pp. 1152–1156.

[15] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Trans. Biomedical Engineering (TBME)*, 2018.

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

[17] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015.

[18] T. G. Dietterich, *Multiple classifier systems*, chapter Ensemble methods in machine learning, pp. 1–15, Springer, 2000.

[19] L. V. Hedges and I. Olkin, *Statistical Methods for Meta-Analysis*, Academic Press, 1985.

[20] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1186–95, 2012.

[21] E. Alickovic and A. Subasi, "Ensemble SVM method for automatic sleep stage classification," *IEEE Trans. on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1258–1265, 2018.

[22] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: An open-access resource for instrument benchmarking & exploratory research," *Journal of Sleep Research*, pp. 628–635, 2014.

[23] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, "The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications," *American Academy of Sleep Medicine*, 2007.

[24] J. A. Hobson, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Electroencephalography and Clinical Neurophysiology*, vol. 26, no. 6, pp. 644, 1969.

[25] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.

[26] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. EMNLP*, 2015, pp. 1412–1421.

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

[28] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *Information Fusion*, vol. 6, no. 1, pp. 83–98, 2005.

[29] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, 2016.

[30] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *Proc. ICLR*, 2015, number 1-13.

[31] S. A. Imtiaz and E. Rodriguez-Villegas, "Recommendations for performance assessment of automatic sleep staging algorithms," in *Proc. EMBC*, 2014, pp. 5044–5047.

[32] N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. M. Hu, and M. De Vos, "Detection of REM sleep behaviour disorder by automated polysomnography analysis," *arXiv preprint arXiv:1811.04662*, 2018.