Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks

Fernando Andreotti[1], Huy Phan[1], Navin Cooray[1], Christine Lo[2], Michele T.M. Hu[3] and Maarten De Vos[1]

***Abstract*— Current sleep medicine relies on the analysis of polysomnographic measurements, comprising amongst others electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG) signals. This analysis currently requires supervision of a trained expert. Convolutional neural networks (CNN) provide an interesting framework to automated classification of sleep epochs based on raw EEG, EOG and EMG waveforms. In this study, we apply CNN approaches from the literature to four databases from pathological and physiological subjects. The best performing model resulted in Cohen's Kappa of $\kappa = 0.75$ on healthy subjects and $\kappa = 0.64$ on patients suffering from a variety of sleep disorder. Further, we show the advantages of using additional sensor data such as EOG and EMG. Last, to cope with smaller datasets of less prevalent diseases, we propose a transfer learning procedure using large freely available databases for pre-training. This procedure is demonstrated using a private REM Behaviour Disorder database, improving sleep classification by 24.4%.**

I. INTRODUCTION

Sleep is a fundamental biological process, widely present in the animal kingdom, that plays a critical role in the maintenance of human mental and physical health [1], [2]. Sleep medicine relies on the analysis of polysomnographic (PSG) recordings, which include EEG, EMG, EOG amongst other physiological signals [3]. In order to understand these signals, guidelines that divide sleep into a handful of stages (e.g. R&K [4] and the AASM [5] norms) have been proposed. These subjective definitions have been the focus of criticism over the last 50 years [6], nonetheless manual scoring following these rules remains the gold-standard in clinical practice. In addition to being subjective, visual analysis of recordings is time consuming, tedious, and prone to subject variability. These drawbacks lead to a mounting number of papers published on computerised classification of PSGs [7], [3]. While automated scoring provides a more objective approach, methods usually make use of hundreds of hand-engineered features obtained from physiological signals. Based on these features, traditional classification methods such as support vector machines, decision trees or hidden Markov models are typically applied (for a review the reader is referred to [8]).

[1] Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, UK.

[2] Department of Neurology, Sheffield Teaching Hospitals, Sheffield Institute of Translational Neuroscience, Sheffield, UK.

[3] Nuffield Department of Clinical Neurosciences, Oxford Parkinsons Disease Centre (OPDC), University of Oxford, UK.

This research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and the Engineering and Physical Sciences Research Council (EPSRC – grant EP/N024966/1).

Correspondence author: F. Andreotti (fernando.andreotti@eng.ox.ac.uk)

Deep learning methods are increasingly popular due to their ability to automatically generate features at multiple levels of abstraction (i.e. layers). This allows the system to learn complex functions by mapping the input to the output directly from data, not relying on hand-engineered features [9]. These methods have been successfully applied in the field of computer vision and speech analysis, however applications to 1-dimensional biomedical signals (e.g. EEG) are only emerging in literature. Some of the early works [10], [11], [12] proposed the usage of deep learning techniques for sleep staging. However, these studies are limited with respect to sensors used and the fact that none of these investigate sleep staging in patients suffering from sleep disorders.

In this study, we evaluate different deep learning models proposed in the literature for automatic sleep scoring. Three freely available databases, containing physiological subjects and patients with a variety of pathologies were used to assess these methods' performance. The advantages of using additional sensors such as EOG and EMG is evaluated on multiple databases. At last, a transfer learning procedure is suggested for improving classification when data is scarce. For this purpose, a private modest-sized dataset of REM Behaviour Disorder (RBD) patients is used. As RBD disease has low prevalence and patients sleep is plagued by arousals [13], sleep classification is a challenging task. Therefore, this dataset provides an ideal case for demonstrating the benefits of transfer learning.

II. MATERIALS

In this study, to avoid a saturation on the number of needed channels in a recording setup, a single central EEG lead (C4-A1 or C3-A2, where available), an average EOG (ROC-LOC), and/or an average EMG (CHIN1-CHIN2) channels are used. Although EOG and EMG signals are used by human experts for sleep staging, they are rarely present in automated systems [12]. The choice for these derivations is due to their common usage in literature. Several public sleep databases exist including different groups of healthy/diseased subjects, ages and genders. This work uses 3 openly accessible databases and a private clinical database. All recordings were resampled at 100 Hz and divided in 30 s epochs. Preprocessing was done by using a zero-phase 100^{th} order FIR filter with 0.1 Hz high-pass cutoff frequency for EEG/EOG signals and 10 Hz for EMG. Annotations using R&K were converted into AASM guidelines by assigning $S3$ and $S4$ stages to $N3$, while $\{S0, S1, S2\}$ were relabelled as $\{W, N1, N2\}$, respectively. The datasets here investigated are described in the following sub-sections.

A. Physionet Sleep-EDF Database (SLPEDF-DB)

The SLPEDF-DB [14], [15] comprises 38 two-night recordings from 19 healthy subjects (recording $SC4131E0$ was excluded due to a missing second night). Recordings comprise 9 young males aged 28.3 ± 2.3 years and 10 young females (29.1 ± 3.4 years). Unlike most studies, EEGs were annotated using Fpz-Cz (or Pz-Oz) derivations and EOG using a horizontal derivation. Signals were originally sampled at 100 Hz except EMG, which was sampled at 1 Hz. A total of 37,147 epochs were produced, being 11.8% W, 20.3% REM, 7.3% N1, 46.0% N2, 14.6% N3.

B. Montreal Archive of Sleep Studies (MASS-DB)

The MASS-DB [16] is a large database comprising 200 healthy participants with ages ranging between 18 and 76 years, including 98 males aged 42.7 ± 19.4 years and 102 females aged 38.1 ± 18.9 years. The database contains single nights and is divided into 5 cohorts all of which were used in this study. Recordings with 20 s epochs were converted into 30 s by including 5 s before and after each segment. A total of 228,870 epochs were produced, being 13.6% W, 17.6% REM, 8.5% N1, 47.2% N2, 13.3% N3.

C. CAP Sleep Database (CAPSLP-DB)

The CAPSLP-DB [17], [15] consists of 108 single night PSG recordings of 16 healthy and 92 pathological subjects. The dataset includes 66 male (aged 48.4 ± 19.2 years) and 42 female (aged 40.0 ± 19.4 years). Between the pathologies are periodic leg movements, insomnia, sleep-apnea as well as 22 RBD subjects. The record $brux1$ was excluded from our analysis due to inconsistent sampling frequency, $n04$, $n08$, and $n16$ were excluded due to absence of either signal modality. A total of 154,094 epochs were produced, being 12.2% W, 11.8% REM, 4.5% N1, 42.5% N2, 28.9% N3.

D. RBD Database (RBD-DB)

The RBD-DB consists of 21 double-night recordings of 20 male (aged 61.5 ± 7.0 years) and a female patient aged 69 years all suffering from RBD. Data was acquired by our local partners from the John Radcliffe hospital, Nuffield Department of Clinical Neurosciences at the University of Oxford. This study complies with the requirements of the Department of Health Research Governance Framework for Health and Social Care 2005 and was approved by the Oxford University hospitals NHS Trust (HH/RA/PID 11957). A total of 45,410 epochs were produced, being 24.1% W, 11.1% REM, 12.2% N1, 36.4% N2, 16.1% N3.

III. METHODS

Convolutional and Recurrent Neural Networks (i.e. CNN and RNNs, respectively) are the most used techniques for deep supervised learning. Due to its computationally efficient algorithm and properties such as translation invariance, parameter sharing and sparse connectivity, CNNs are often the method of choice for operating over grid-like structures (e.g. images or fixed segment windows) [18]. In its hidden layers, CNNs produce feature maps with a high degree of abstraction. In this work, we apply CNN architectures from literature to classify individual epochs of sleep data. These methods are described in the following sub-sections. For reproducibility, fully-connected (FC) layers were removed and the CNNs are followed by a single softmax layer. Removing FC layers forces the network to learn good representations in the convolutional layers, potentially leading to better generalisation [19].

A. Two-layer approach [10]

The approach by [10] proposed a two-layer CNN model specifically for sleep scoring and evaluated on the SLPEDF-DB. The resulting filters of the first 1-dimensional convolution were stacked and further processed by a 2D convolution, which results in 496k parameters. To evaluate the necessity of such a stacking procedure, we also evaluate a simpler network comprising two 1D-CNNs with the same temporal dimensions as in [10] resulting in 97k parameters.

B. DeepSleepNet [11]

In [11] two branches of 4 convolutional layers each were proposed. Each branch operates with different kernel sizes, aiming to generate feature maps with low and high frequency content. The CNN was then followed by a bidirectional Long-short Term Memory (LSTM - a type of RNN). The authors used a single lead which mixes EEG and EOG information and tested their models on a subset of the MASS-DB. In this work we make use of the proposed CNN network, which has 844k parameters.

C. Residual Network (ResNet) [20]

In this work we apply the pre-activation ResNet (also called v2), max-pooling the first two layers to reduce dimensionality as in [19]. Similarly to [19], a receptive field of 30 samples was used in the first layer to allow the model a higher level of abstraction with a lower number of layers. We evaluated ResNet models with 12, 22 and 34 layers totalling from 608k to 4.7M parameters.

IV. EXPERIMENTS

A. Experiment 1: Input Channels and Performance

In this experiment, we aim to assess if using multiple sensors improves the classification accuracy. For this purpose, we chose to apply the DeepSleepNet model [11] due to its reported accuracy. The model is applied to various channel combinations of each dataset, using both nights from the SLPEDF-DB and RBD-DB merged into a single subject. Raw EEG, EOG and EMG waveforms are added to the input as additional features. The results of a 5-fold cross-validation are shown in Table I. Results are reported using Cohen's κ coefficient for nominal multi-class agreement, which takes chance into consideration. As a rule of thumb, $\kappa \in [0.6, 1]$ is considered good whereas $\kappa \in [0, 0.4]$ is fair to poor. The model was trained in 100 epochs, with batch size 256.

From Table I we observe that including both EEG and EOG sensors significantly improves sleep stage classification. EMG improves the performance on the healthy subjects

TABLE I

EXPERIMENT 1: MEAN AND STANDARD DEVIATION OF COHEN'S KAPPA COEFFICIENTS FOR 5-FOLD CROSS-VALIDATION ON DATABASES AND DIFFERENT INPUT CHANNELS USING [11].

Input Signals	Dataset			
	SLPEDF-DB	MASS-DB	CAPS-DB	RBD-DB
EEG	0.65±0.04	0.67±0.02	0.58±0.02	0.46±0.06
EOG	0.58±0.04	0.66±0.01	0.58±0.01	0.43±0.05
EMG	0.07±0.01	0.34±0.02	0.18±0.02	0.13±0.04
EEG+EOG	0.68±0.04	0.72±0.01	0.62±0.02	0.49±0.06
EEG+EOG+EMG	0.67±0.05	0.74±0.01	0.61±0.01	0.48±0.07

of the MASS-DB, while on pathological cases it slightly worsens results. The SLPEDF-DB is an exception since EMG is sampled at 1 Hz, therefore much of the information is lost. The MASS-DB results agree with [12], where multiple channels of each modality were used. Different from the method presented in [12], in this work all three signals undergo the same pipeline of neural networks, i.e. using non-linear transformations rather than linear on the input and allowing the network to deal with different statistical properties of those signals. The original work by [10] made use of a single EEG channel on the SLPEDF-DB, while [11] proposed combining EEG and EOG leads into one channel of the MASS-DB. Results should be taken with care since the architecture used may have an influence. Despite its slightly worse results, EMG was kept on following experiments since it contains crucial information for pathologies like RBD [13].

B. Experiment 2: Models on Different Databases

To evaluate each model's performance on the individual databases, we performed a 5-fold cross-validation using all three signals as input. Hyper-parameters were kept the same as in the previous experiment. In Fig. 1, the methods' performance are depicted in terms of macro-averaged F_1-measure, sensitivity (SE), and specificity (SP) using the largest healthy/disease databases available (i.e. MASS-DB and CAPS-DB). From this figure it is noticeable that the stacking procedure proposed by [10] considerably worsens the performance on the MASS-DB, however performs better on the CAPSLP-DB. As expected, increasing the number of layers on the ResNet increases performance, which differs from [19], who attributed a worsen in accuracy to ResNets overfitting the training set. From both Fig. 1 and 2 we note that the DeepSleepNet [11] consistently outperforms all other models with a modest number of required parameters.

In Fig. 2, classification performance on each database are shown using the best performing variants of each method (based on Fig. 1). It is visible that the model performs well on different datasets, except the RBD-DB which is more challenging especially for states $N1$ and REM. REM detection is crucial for RBD as the pathology is defined based on muscular atonia occurring during this state [13].

C. Experiment 3: Fine-tuning Model to Subject

Transfer learning is an important strategy when dealing with deep neural networks. Particularly when data is scarce, such as in the cases of less prevalent disorders (e.g. RBD). This strategy allows higher classification accuracy by 1) pre-training a model in a more readily available data of a similar task; and 2) fine-tuning the network to the specific task at hand. In this study, the MASS-DB and CAPSLP-DB are used to pre-train the best performing methods from previous experiments (i.e. DeepSleepNet [11] and ResNet 34-layers [20]). The pre-trained model is then fine-tuned to each patient's first night of the RBD-DB and evaluated on the second night. This personalisation procedure is similar to the one performed in [21]. As a baseline method we perform a leave-one-subject-out procedure on the second night of the RBD-DB using the DeepSleepNet, which resulted in $\kappa = 0.45 \pm 0.15$.

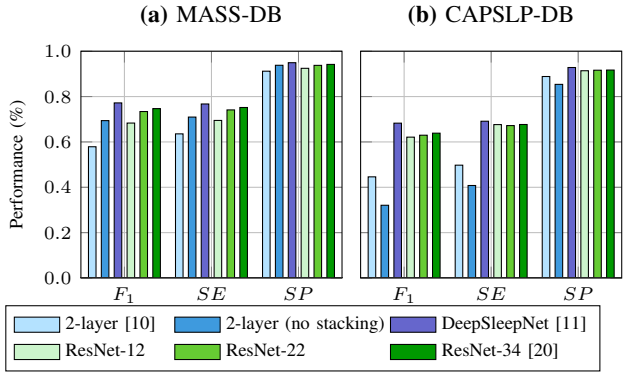

Fig. 1. Performance of all methods evaluated in this study on largest databases MASS-DB and CAPSLP-DB. Metrics used are F_1-measure, sensitivity (SE) and specificity (SP).

During fine-tuning, the DeepSleepNet performed best when all layers were adapted, with an average $\kappa = 0.43 \pm 0.21$. For the ResNet-34, parameters were only changed from the 5$^{\text{th}}$ residual block onward, producing $\kappa = 0.56 \pm 0.17$. From these results it is evident that pre-training the ResNet using large databases considerably improves classification performance on the RBD-DB. On the other hand, the DeepSleepNet [11] does not seem to produce transferable feature representations.

V. DISCUSSION

In acoustic signal processing very deep networks are uncommon, instead raw waveforms are converted into spectrograms and a few convolutional layers produce similar results [19]. Future work should compare both performance and transferability of these models. Moreover, the temporal dependency between epochs was not explored in this study. In order to treat sequences of sleep epochs, i.e. transitions between stages, [11] suggested bi-directional LSTM layers on top of the CNN network. Another limitation of this study is the amount of data on the RBD-DB. Additional subjects may be used to confirm the results here shown. Last, sleep staging is a common method in medicine, however, it only provides a partial understanding of the process itself. End-to-end learning of how to diagnose sleep disorders, on the other hand, may be more clinically relevant than improving sleep staging.

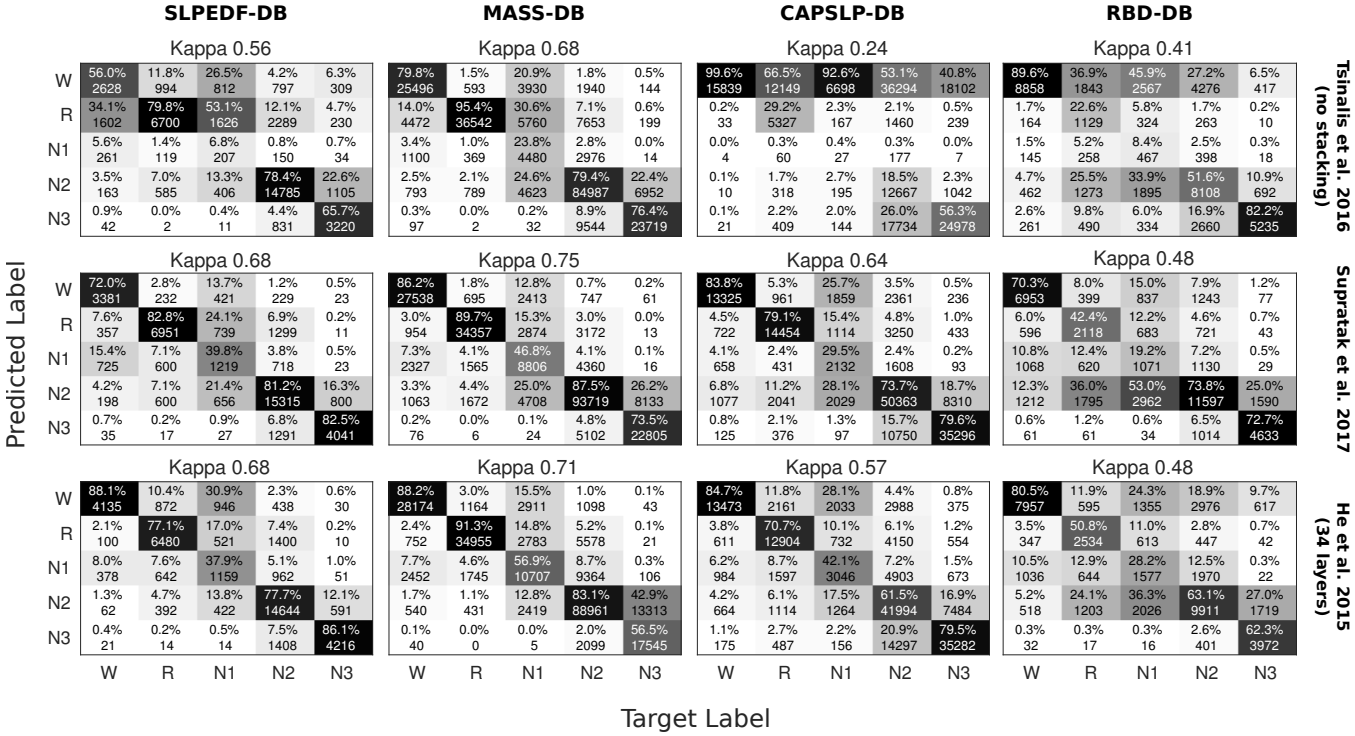

Fig. 2. Resulting confusion matrices for experiment 2 produced during 5-fold cross-validation over each database and method.

VI. CONCLUSION

In this work, we apply several deep convolutional neural networks to the task of automated sleep staging. Approaches are compared in terms of input sensors (EEG, EOG and/or EMG) with the help of four different databases from pathological and physiological subjects. At last, the generalisation power of these methods is demonstrated by pre-training a network on a combined large database consisting of both healthy and diseased subjects and fine-tuning this network's weights so that it improves classification performance on a small, more challenging database (i.e. RBD-DB).

REFERENCES

[1] K. Wulff, S. Gatti, J. G. Wettstein, and R. G. Foster, "Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease," *Nat. Rev. Neurosci.*, vol. 11, no. 8, pp. 589–599, 2010.

[2] F. Weber and Y. Dan, "Circuit-based interrogation of sleep control," *Nature*, vol. 538, no. 7623, pp. 51–59, oct 2016.

[3] R. Agarwal and J. Gotman, "Computer-assisted sleep staging," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 12, pp. 1412–1423, 2001.

[4] A. Rechtschaffen and A. Kales, *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*, 1968.

[5] R. Berry, R. Brooks, C. Gamaldo, S. Harding, R. Lloyd, C. Marcus, and B. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, v2.2 ed. American Academy of Sleep Medicine, 2015.

[6] S.-L. Himanen and J. Hasan, "Limitations of Rechtschaffen and Kales," *Sleep Med. Rev.*, vol. 4, no. 2, pp. 149–167, apr 2000.

[7] T. Penzel and R. Conradt, "Computer based sleep recording and analysis," *Sleep Med. Rev.*, vol. 4, no. 2, pp. 131–148, 2000.

[8] S. Motamedi-Fakhr, M. Moshrefi-Torbati, M. Hill, C. M. Hill, and P. R. White, "Signal processing techniques applied to human sleep EEG signalsA review," *Biomed. Signal Process. Control*, vol. 10, no. 1, pp. 21–33, mar 2014.

[9] Y. Bengio, "Learning Deep Architectures for AI," *Found. Trends Machine Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[10] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks," *arXiv preprint arXiv:1610.1683*, p. 12, oct 2016.

[11] 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. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, nov 2017.

[12] S. Chambon, M. Galtier, P. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," pp. 1–14, 2017.

[13] B. F. Boeve, M. H. Silber, and et al., "Pathophysiology of REM sleep behaviour disorder and relevance to neurodegenerative disease," *Brain*, vol. 130, no. 11, pp. 2770–2788, 2007.

[14] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, 2000.

[15] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." *Circulation*, vol. 101, no. 23, pp. E215—-E220, jun 2000.

[16] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research," *J. Sleep Res.*, vol. 23, no. 6, pp. 628–635, dec 2014.

[17] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep." *Sleep Med.*, vol. 2, no. 6, pp. 537–53, nov 2001.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 2016.

[19] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 421–425, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," *arXiv 1603.05027v3*, no. 1, pp. 1–15, mar 2016.

[21] K. Mikkelsen and M. De Vos, "Personalizing deep learning models for automatic sleep staging," *arXiv:1801.02645 [q-bio.NC]*, pp. 1–9.