Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders

[Automatic sleep scoring with autoencoders]

Orestis Tsinalis[1], Paul M. Matthews[2], and Yike Guo[1*]

[1]Department of Computing, Imperial College London, London, UK

[2]Division of Brain Sciences, Imperial College London, London, UK

[*]email: y.guo@imperial.ac.uk, tel: +44 (0) 20 7594 8182, fax: +44 (0) 20 7581 8024

Abstract

We developed a machine learning methodology for automatic sleep stage scoring. Our time-frequency analysis-based feature extraction is fine-tuned to capture sleep stage-specific signal features as described in the American Academy of Sleep Medicine (AASM) manual that the human experts follow. We used ensemble learning with an ensemble of stacked sparse autoencoders for classifying the sleep stages. We used class-balanced random sampling across sleep stages for each model in the ensemble to avoid skewed performance in favour of the most represented sleep stages, and addressed the problem of misclassification errors due to class imbalance while significantly improving worst-stage classification. We used an openly available dataset from 20 healthy young adults for evaluation. We used a single channel of EEG from this dataset, which makes our method a suitable candidate for longitudinal monitoring using wearable EEG in real-world settings. Our method has both high overall accuracy (79%, range 76–81%), and high mean F_1-score (85%, range 83–87%) and mean accuracy across individual sleep stages (87%, range 85–88%) over all subjects. The performance of our method appears to be uncorrelated with the sleep efficiency and percentage of transitional epochs in each recording.

Key Terms— Electroencephalography, EEG, Deep learning, Ensemble learning

1 Introduction

Sleep is central to human health. The health consequences of reduced sleep, abnormal sleep patterns or desynchronized circadian rhythms can be emotional, cognitive, or somatic [26]. Associations between disruption of normal sleep patterns and neurodegenerative diseases are well recognised [26].

According to the American Academy of Sleep Medicine (AASM) manual [11], sleep is categorised into four stages. These are Rapid Eye Movement (stage R) sleep and 3 non-R

stages, stages N1, N2 and N3. Formerly, stage N3 (also called Slow Wave Sleep, or SWS) was divided into two distinct stages, N3 and N4 [20]. To these a Wake (W) stage is added. These stages are defined by electrical activity recorded from sensors placed at different parts of the body. The totality of the signals that are recorded through these sensors is called a polysomnogram (PSG). The PSG includes an electroencephalogram (EEG), an electrooculogram (EOG), an electromyogram (EMG), and an electrocardiogram (ECG). After the PSG is recorded, it is divided into 30-second intervals, called *epochs*. Then, one or more experts classify each epoch into one of the five stages (N1, N2, N3, R or W) by quantitatively and qualitatively examining the signals of the PSG in the time and frequency domains. Sleep scoring is performed according to the Rechtschaffen and Kales sleep staging criteria [20]. In Table 1 we reproduce the Rechtschaffen and Kales sleep staging criteria [22], merging the criteria for N3 and N4 into a single stage (N3).

Recent research suggests that detection of sleep/circadian disruption could be a valuable marker of vulnerability and risk in the early stages of neurodegenerative diseases, such as Alzheimer's disease, Parkinson's disease and multiple sclerosis, and that sleep stabilisation could improve the patients' quality of life [26]. There is therefore a pressing need for longitudinal sleep monitoring for both medical research and medical practice. In this case an affordable, portable and unobtrusive sleep monitoring system for unsupervised at-home use would be ideal. Wearable EEG is a strong candidate for such use. A core software component of such a system is a sleep scoring algorithm, which can reliably perform automatic sleep stage scoring given the patient's EEG signals.

In this study we present and evaluate a machine learning methodology for automatic sleep stage scoring using a single channel of EEG. Our methodology is based on time-frequency analysis [4] and stacked sparse autoencoders [1]. We compared the performance of our method with three existing studies. In [6] the data consisted of 16 subjects (aged 30–75 years) and the EEG channel used was C3-A1. The authors' method was time-frequency analysis using the Continuous Wavelet Transform (CWT) and Renyi's entropy for feature extraction, and the random forest classifier. In [16] the first dataset comprised 20 subjects (aged 20–22 years), using channel C3-A2. The second dataset [19] comprised

8 subjects (aged 21–35 years), and the authors chose channel Pz-Oz. The authors used multiscale entropy (MSE) for feature extraction from the EEG signal, and they also fitted an autoregressive (AR) model to the signal. They then trained a linear discriminant analysis (LDA) model using the MSE features and the fitted parameters of the AR model as features, employing a set of 11 *a priori* 'smoothing rules' on the hypnogram after the initial sleep scoring. In [3] the authors used a dataset comprising 15 subjects (aged 29.2 ± 8 years). The feature extraction methods and the machine learning algorithm are not described in detail in [3].

There are two main limitations in the existing literature. First, regarding the results of the proposed methods, in all three studies we observe imbalance in the scoring performance across sleep stages. For example, the F_1-score in the worst-classified sleep stage (N1) can be as low as 30% in [16]. Second, regarding the evaluation methodology, in all three studies the authors evaluated their methods using a single training-testing split of the data, and did not perform any type of cross-validation. Furthermore, in [6] the authors trained and tested their algorithm using epochs from all subjects, which means that the training and testing datasets were not independent. In this work we mitigated skewed sleep scoring performance in favour of the most represented sleep stages, and addressed the problem of misclassification errors due to class imbalance in the training data while significantly improving worst-stage classification. Our experimental design employs cross-validation across subjects, ensuring independence of training and testing data.

2 Materials and Methods

2.1 Data

The dataset that we used to evaluate our method is a publicly available sleep PSG dataset [14] from the PhysioNet repository [7] that can be downloaded from [18]. The data was collected from electrodes Fpz-Cz and Pz-Oz, instead of the standard C3-A2 and C4-A1. The sleep stages were scored according to the Rechtschaffen and Kales guidelines [20]. The epochs of each recording were scored by a single expert (6 experts in total). The sleep

stages that are scored in this dataset are Wake (W), REM (R), non-R stages 1–4 (N1, N2, N3, N4), Movement and Not Scored. For our study, we removed the very small number of Movement and Not Scored epochs (Not Scored epochs were at the start or end of each recording), and also merged the N3 and N4 stages into a single N3 stage, as it is currently the recommended by the American Academy of Sleep Medicine (AASM) [11, 22]. There were 61 movement epochs in our data in total, and only 17 of the 39 recordings had movement artifacts. The maximum number of movement epochs per recording was 12. The rationale behind the decision of removing the movement epochs was based on two facts. First, these epochs had not been scored by the human expert as belonging to any of the 5 sleep stages, as it is recommended in the current AASM manual [11, p. 31]. Second, their number was so small that they could not be used as a separate 'movement class' for learning. The public dataset includes 20 healthy subjects, 10 male and 10 female, aged 25–34 years. There are two approximately 20-hour recordings per subject, apart from a single subject for whom there is only a single recording. To evaluate our method we used the in-bed part of the recording. The sampling rate is 100 Hz and the epoch duration is 30 seconds.

2.2 Feature extraction

For feature extraction we performed time-frequency analysis using complex Morlet wavelets (see, for example, Chapters 12 and 13, pp. 141–174 in [4]). The reason for preferring a time-frequency-based feature extraction method over the Fourier transform was that we wanted to extract features that capture the mixture of frequencies and their interrelations at different points in time as features.

For time-frequency analysis using complex Morlet wavelets there are two sets of parameters that need to be chosen, the peak frequencies and the number of wavelet cycles per frequency. The number of wavelet cycles defines its width and controls the trade-off between temporal and frequency precision. Specifically, increasing the number of cycles increases the frequency precision but decreases the temporal precision, while decreasing the number of cycles increases the temporal precision but decreases the frequency preci-

sion. In this study we selected the peak frequencies and the number of cycles based on the sleep scoring criteria in Table 1, taking into account the transition rules in Table 2. In Table 3 we summarise the parameters chosen.

After extracting the frequency-band power for each peak frequency given in Table 3, the features that we computed for each epoch were the power of the frequency-band power signal, the power of the time-domain signal, the Pearson correlation coefficient between each pair of frequency-band power signals and the autocorrelation in the time-domain signal for 50 time lags (i.e. up to 0.5 seconds). Additionally, we used a sliding window to extract the power of the frequency-band power and the power of the time-domain signal at different intervals within each epoch. Specifically, we used a sliding window of duration of 5 seconds and step of 2.5 seconds, which resulted in 11 power of frequency-band power features per frequency band per epoch and 11 power of the time-domain signal features per epoch. All the extracted features are summarised in Table 4. We mapped all the features in the [0,1] interval, and centred their distribution using transformations (see Table 4), as this is beneficial for our learning algorithm. We then normalised the features from each trial of each subject.

The AASM manual [11] includes a number of rules that recommend taking into account neighbouring epochs for the scoring of each current epoch under certain circumstances. We identified 12 rules in total concerning the transition between certain sleep stage pairs that refer to 7 distinct transition patterns, as shown in Table 2. These rules apply to three sleep stage pairs, N1-N2, N1-R and N2-R. The transition patterns include up to two preceding or succeeding neighbouring epochs. Trying to capture the effect of these transition rules in an automatic sleep scoring algorithm by simply including transition probabilities between sleep stages is not a suitable approach. The reason is that the algorithm could overfit to hypnogram-level patterns from the subjects we used for training, especially when the training data do not include data from different sleep pathologies.

We incorporated transition information directly as features for our machine learning algorithm. Specifically, for the classification of each epoch, apart from the features

corresponding to itself, we included the features from the preceding two and succeeding two epochs. We addressed the possibility of overfitting which exists in this case in our experimental design (Subsection 2.4). In the literature, Liang *et al.* [16] used 11 *a priori* hypnogram 'smoothing rules' in order to capture transition information. These rules are applied on the scored epochs after automatic sleep scoring has taken place, effectively changing the classification of each epoch given the sleep stage of its neighbours. Unfortunately, the authors described only 2 of the rules in their paper, and, notably, did not discuss the order in which the rules are applied to the estimated hypnogram.

2.3 Machine learning methodology

Stacked sparse autoencoders [1] are a specific type of neural network model. The key difference between stacked autoencoders and standard neural networks is layer-wise pre-training using unlabelled data (i.e. without class labels) before fine-tuning the network as a whole [2]. Autoencoders are trained using iterative optimisation with the backpropagation algorithm. The optimisation method we used was L-BFGS, as recommended in [15]. The hyperparameters of a sparse autoencoder-based model are: (1) a regularisation weight λ which is used to decrease the magnitude of the parameters and prevent overfitting, (2) a sparsity weight β which controls the relative importance of the sparsity penalty term, (3) a sparsity parameter ρ which sets the desired level of sparsity, and (4) the number of units n in the hidden layer of the autoencoder. The only hyperparameter for the optimisation is the total number of iterations r.

The combinatorial space to explore all the possible combinations of hyperparameters is huge. Therefore, we decided to choose the same hyperparameters across all layers. Our final choice was $\lambda = 1 \times 10^{-5}$, $\beta = 2.0$, $\rho = 0.2$, $n = 20$, and $r = 60$. We used autoencoders with the sigmoid activation function, which is symmetric. This is the reason that our features were transformed so that their distribution be approximately centred around the mean.

The classes (sleep stages) in our dataset, as in any PSG dataset, were not balanced, i.e. there were a lot more epochs for some stages (particularly N2) than others (particularly W

and N1). In such a situation, if all the data is used as is, it is highly likely that a classifier will exhibit skewed performance favouring the most represented classes, unless the least represented classes are very distinct from the other classes. In order to resolve the issues stemming from imbalanced classes we decided to employ class-balanced random sampling with an ensemble of classifiers, each one being trained on a different sample of the data. Our final model consisted of an ensemble of 20 independent stacked sparse autoencoders (SSAEs) with the same hyperparameters. Each of the 20 SSAEs was trained using a sample of the data in which the number of epochs per stage per recording was equal to the number of epochs of the least represented stage (N1). The classification of the epochs in the testing recordings was done by taking the mean of the class probabilities that each of the 20 SSAEs outputs, and then selecting the class with the highest probability.

We used our own Matlab implementation for time-frequency analysis and stacked autoencoders, and the Matlab implementation by Mark Schmidt for L-BFGS (`http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`).

2.4 Evaluation

To evaluate the generalisability of our method, we obtained our results using 20-fold cross-validation. Specifically, in each fold we used the recordings of a single subject for testing and all other recordings for training. We used each subject's recordings only once for testing, thus obtaining a one-to-one correspondence of cross-validation folds and test subjects. We chose per-subject cross-validation as we also performed comparisons across individual recordings. With this experimental design, we were able to assess both the overall performance of our method and the performance across recordings with a single set of experimental results.

We report the evaluation metrics using their average across all recordings. Specifically, we report their mean value across all 5 sleep stages and their value for the most misclassified sleep stage, which gives information about the robustness of the method across sleep stages. We tested our method with both available EEG electrodes (Fpz-Cz and Pz-Oz). We report the scoring performance using the best electrode, which was Fpz-

Cz. Finally, we calculated 95% confidence intervals for each of the performance metrics by bootstrapping using 1000 bootstrap samples across the 39 recordings.

We also tested our algorithm using 5-fold cross-validation with non-independent training and testing sets by mixing the subjects' epochs as the authors in [6] did. This was done to show the improvement in the results that such a flawed practice can result into, and appropriately compare our method to [6]. We do not consider this performance indicative of the quality of our method, or any method targeted in EEG sleep scoring, as it is not practical in the real world. These results are separated from the others in Table 6.

To further evaluate the generalisability of our method, we performed two tests on our results to assess the correlation between scoring performance and (1) a measure of the sleep quality of each recording, and (2) the percentage of transitional epochs in each recording. Robust scoring performance across sleep quality and temporal sleep variability, can be seen as further indicators of the generalisability of an automatic sleep stage scoring algorithm. The reason is that low sleep quality and high sleep stage variability across the hypnogram are prevalent in sleep pathologies (see, for example, [17]).

We measured sleep quality with a widely-used index, called *sleep efficiency*. Sleep efficiency is defined as the percentage of the total time in bed that a subject was asleep [23, p. 226]. Our data contain a 'lights out' indicator, which signifies the start of the time in bed. We identified the sleep onset as the first non-W epoch that occurred after lights were out. We identified the end of sleep as the last non-W epoch after sleep onset, as our dataset does not contain a 'lights on' indicator. The number of epochs between the start of time in bed and the end of sleep was the total time in bed, within which we counted the non-W epochs; this was the total time asleep. We defined transitional epochs as those whose preceding or succeeding epochs were of a different sleep stage than them. We computed their percentage with respect to the total time in bed. In our experiments we computed the R^2 and its associated p-value between sleep efficiency and scoring performance, and between percentage of transitional epochs and scoring performance.

All scoring performance metrics are derived from the confusion matrix. Using a 'raw' confusion matrix in the presence of imbalanced classes implicitly assumes that the relative importance of correctly detecting a class is directly proportional to its frequency of occurrence. This is not desirable for sleep staging. What we need to mitigate the negative effects of imbalanced classes on classification performance measurement is effectively a normalised or 'class-balanced' confusion matrix that places equal weight into each class. Surprisingly, in the single-channel EEG sleep staging literature there are examples of such mistakenly reported performance results using the raw confusion matrix. For this reason, we compared our work only with the studies in the literature that provided the raw confusion matrix, from which we computed the performance metrics after class-balancing.

The metrics we computed were precision, sensitivity, F_1-score, per-stage accuracy, and overall accuracy. The F_1-score is the harmonic mean of precision and sensitivity and is a more comprehensive performance measure than precision and sensitivity by themselves. The reason is that precision and sensitivity can each be improved at the expense of the other. All the metrics apart from overall accuracy are binary. However, in our case we have 5 classes. Therefore, after we performed the classification and computed the normalised confusion matrix, we converted our problem into 5 binary classification problems each time considering a single class as the 'positive' class and all other classes combined as a single 'negative' class (*one-vs-all* classification).

Finally, we computed the scoring performance of our algorithm without and with features from neighbouring epochs. If we observed improvement in sleep stage pairs which are not included in the transition rules (i.e. any pair other than N1-N2, N1-R and N2-R, see Table 2), we would conclude that the algorithm learned spurious patterns that are an artifact of our training data. Additionally, we should observe at least some small improvement and certainly no decrease in the classification performance between pairs N1-N2, N1-R and N2-R. In this case, even without having data from different sleep pathologies we can evaluate whether the epoch-to-epoch or hypnogram-level patterns that our algorithm learned were akin to the generic guidelines or overfitting to the training data.

3 Results

As we show in the the normalised confusion matrix in Table 5, the most correctly classified sleep stage was N3, with around 90% of stage N3 epochs correctly classified. Stages N2, R and W follow, with around 80% of epochs correctly classified for each stage. The most misclassified stage was N1 with 60% of stage N1 epochs correctly classified. Most misclassifications occurred between the pairs N1-W and N1-R (about 15% and 13% respectively), followed by pairs N1-N2 and N2-N3 (about 8%), and N2-R and R-W (about 4%). The remaining pairs had either misclassification rates smaller than 4% (N2-W and N3-W) or almost no misclassifications at all (N1-N3 and N3-R). We also observe that the percentage of false negatives with respect to each stage (non-diagonal elements in each row) per pair of stages was approximately balanced between the stages in the pair (the only conspicuous exception is the pair N1-W, and, to a lesser extent, the pair N2-W). Effectively the upper and lower triangle of the confusion matrix are close to being mirror images of each other. This is a strong indication that the misclassification errors due to class imbalance have been mitigated.

As we show in Table 6, our method has both high overall accuracy (78%, range 75–80%), and high mean F1-score (84%, range 82–86%) and mean accuracy across individual sleep stages (86%, range 84–88%) over all subjects. From the scoring performance metrics results in Table 6 we observe that our method either outperformed or had approximately equal performance with the methods in the literature in all metrics apart from worst-stage precision (the non-independent testing results at the bottom row are not taken into account). In many cases, even the lower end of the 95% confidence interval (the top number in parentheses) was higher than the corresponding metric for the other methods. Table 6 also summarises the improvement of our method over the state of the art, i.e. the best of all the methods in the literature in that particular metric (negative numbers indicate worse performance than the state of the art). Overall, our method exhibits improved performance over the state of the art in automated sleep scoring using single-channel EEG across the five scoring performance metrics.

In Table 7 we show the results of the algorithm without and with information from

neighbouring epochs. We observe that there is no mutual improvement in any other stages apart from the targeted pairs N1-N2, N1-R and N2-R.

We also assessed the independence of the scoring performance (for F_1-score and overall accuracy) of our method across recordings relative to sleep efficiency and the percentage of transitional epochs per recording (Table 8). The p-values of the regression coefficients are all above 0.15, which means that we fail to reject the null hypothesis of zero R^2, which is already negligible (lower than 0.1) in all cases. For clarity we present the data for these tests graphically for the F_1-score results in Figures 1 and 2. Our dataset contained 10 recordings with sleep efficiency below 90% (in the range 60-89%), which is the threshold recommended in [23, p. 7] for young adults. The percentage of transitional epochs ranged from 10-30% across recordings.

Finally, in Figure 3 we present an original manually scored hypnogram and its corresponding estimated sleep hypnogram using our algorithm for a single PSG for which the overall F_1-score was approximately equal to the mean F_1-score across the entire dataset.

4 Discussion

Given the high disagreement across epochs between human experts [24] a 1–2% improvement in mean scoring performance may not be considered significant. We think that there are two characteristics that render our method better than the state of the art. First, we significantly decreased the gap between the mean performance over all sleep stages and the most misclassified stage performance (stage N1) compared to the state of the art with about 20% improvement in the F_1-score and 10% improvement in accuracy over the state of the art (with independent testing). Second, we mitigated the adverse effects of class imbalance to sleep stage scoring. This is an indication that our method could be generalised to data with varying proportions across sleep stages, and is not markedly affected by these proportions, as other methods in the literature seem to be by inspecting their normalised confusion matrices. In our future work we aim to replicate these results in independent datasets. After addressing class imbalance, the majority of the remain-

ing misclassification errors is likely due to either differences in EEG patterns that our feature extraction methodology cannot sufficiently capture, difficulty in capturing EOG and EMG-related that are important in distinguishing between certain sleep stage pairs features through the single channel of EEG, or inherent similarities between sleep stages in epochs that even experts would disagree with one another about.

The most misclassified pair of sleep stages using our method was N1-W; about 15% false negatives for each stage were accounted for by the other. We think that the root cause of the problem is the similarity in the characteristic EEG frequency patterns of sleep stages N1 and W, as described in the AASM sleep scoring manual [11]. Specifically, relatively low voltage mixed 2–7 Hz and alpha (8–13 Hz) activity are described as criteria for both stages. The second most misclassified pair of sleep stages was N1-R, for which the characteristic EEG frequency patterns are similar as well. There are 4 transition rules which pertain to the N1-R pair in the AASM manual, which have proven useful, as we showed in Table 7. However, some of these rules rely heavily on EOG and EMG, so it was difficult to exploit their full potential. The next most misclassified pairs of sleep stages were N1-N2 and N2-N3 (about 8%). The classification between stages N1 and N2 depends to a great extent on transition patterns (Table 2) that partly rely on the detection of arousals (and, in particular, on K-complexes associated or not with arousals), body movements and slow eye movements, which can be difficult to capture using a single channel of EEG. The misclassification between stages N2 and N3 could be partly attributed to the potential persistence of sleep spindles in stage N3 [11, p. 27].

Of the two electrodes in the dataset, we achieved better results using the signal from electrode Fpz-Cz. We hypothesised that this was due to fact that the Fpz-Cz position can better capture most of the frequency band activity that is important for sleep staging. Specifically, delta activity [5], K-complexes [9] and lower frequency sleep spindles [12] are predominantly frontal phenomena, and alpha activity, although it is predominantly an occipital phenomenon, can manifest itself in frontal derivations [5]. Theta activity [5] and higher frequency sleep spindles [12] are mostly parietal phenomena. However, theta activity is present in multiple sleep stages, so even if it were captured more effectively

from the Pz-Oz position it might not have been very beneficial by itself. In our future work, we aim to work with datasets with more electrodes so that we can rigorously test specific hypotheses about the suitability of different electrode positions.

Although we recognise that our dataset does not contain a very large number of recordings of bad sleep quality, we found no statistically significant correlation between sleep efficiency and mean scoring performance. Similarly, there was no statistically significant correlation between the percentage of transitional epochs (which are by definition more ambiguous) and mean sleep scoring performance. These statistical test results indicate that our method could be robust across a number of potentially adverse factors. In our future work we aim to perform the same tests in datasets containing a wider range of ages and sleep pathologies.

Mean interrater agreement between human sleep scorers across subjects and stages can vary significantly. For example, in [24] the consensus agreement among three experts was between 60-80%. It would therefore be desirable that the difference in the performance of an automated scoring algorithm across scorers is not significant (i.e. that the algorithm does not overfit to a specific expert's scoring style). Each recording in our dataset was scored by one of six different experts. In total there are 27 recordings scored by a single expert (expert C), and 12 recordings scored by all other five experts combined. The number of recordings per expert was not sufficiently large to perform a formal statistical test to assess the significance of differences in scoring performance across experts. Both the mean F_1-score for the recordings scored by expert C and the mean F_1-score for the recordings scored by any of the other experts were between 83–84%. Both values are close to each other and the overall F_1-score. In our future work we aim to work with datasets that either, preferably, are scored using consensus agreement or, alternatively, contain a larger number of recordings per expert.

For different pathologies that are related with sleep disorders, there are different sleep stages that are relatively more important for distinguishing them from normal sleep. For instance, to distinguish normal sleep from sleep in patients with depression stages R and N3 are relatively more important than other stages (see for example [21]). Common

measures of sleep quality, include sleep efficiency, wake after sleep onset and sleep latency [23, p. 226], for all of which detection of stage W is essential. Different drugs are associated with effects in all non-R sleep stages N1, N2 and N3 [23, p. 9]. Excessive daytime sleepiness and sudden-onset sleep (sudden W to N2 transition) are present in Parkinson's disease [10], and detection of stages N1 and N2 are particularly important for those. These examples indicate the broad range of sleep architecture aspects that need to be targeted across different pathologies. Therefore, the accurate scoring of the entire sleep architecture would be beneficial for a wide range of biomedical applications.

Our method can account for case-specific relative importance of sleep stages in a straightforward way. Our classification algorithm outputs class probabilities. Since in our paper we placed the same weight to each sleep stage, we classified each epoch to the stage that had the highest class probability. If we wanted to place different weight to each class, we could multiply each stage's probability with a stage-specific weight before choosing the stage with the highest class probability (of course, these weights should be the same for each classified epoch). This would incorporate the relative importance that a researcher places on each sleep stage given the specific sleep pathology that they are trying to identify.

To the best of our knowledge our method has the best performance in the literature when classification is done across all five sleep stages simultaneously using a single channel of EEG. This is different from doing fewer than five one-vs-all classification tasks, as in the latter case, if the eventual overall objective is simultaneous 5-class classification, the performance is likely overestimated. There are examples in the literature that achieve higher performance in a single or two one-vs-all classification tasks, especially for the most easily distinguishable stages N3 and W. However, this is not the same as achieving high performance in a 5-class classification problem, because the errors in the remaining classes are not taken into account. Therefore, since our method achieved very high performance for stages N3 and W, while *simultaneously* achieving good performance in the remaining stages, it is preferable to a method that achieves high performance in a stage W versus N3-only classification task.

5 Acknowledgements

The research leading to these results was funded by the UK Engineering and Physical Sciences Research Council (EPSRC) through grant EP/K503733/1. PMM acknowledges support from the Edmond J Safra Foundation and from Lily Safra and the Imperial College Healthcare Trust Biomedical Research Centre and is an NIHR Senior Investigator. We thank Chao Wu and the three anonymous reviewers for helpful comments and suggestions on the manuscript.

References

[1] Bengio, Y. Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2(1):1-127, 2009.

[2] Bengio, Y., P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In: NIPS. Vancouver, 2006.

[3] Berthomier, C., X. Drouot, M. Herman-Stoïca, P. Berthomier, J. Prado, D. Bokar-Thire, O. Benoit, J. Mattout, and M.-P. d'Ortho. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. *Sleep.* 30(11):1587–1595, 2007.

[4] Cohen, M. X. *Analyzing Neural Time Series Data: Theory and Practice.* Cambridge, MA:MIT Press, 2014.

[5] Finelli, L. A., A. A. Borbély, and P. Achermann. Functional topography of the human nonREM sleep electroencephalogram. *Eur. J. Neurosci.* 13(12):2282–2290, 2001.

[6] Fraiwan, L., K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus. Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. *Comput. Meth. Prog. Bio.* 108(1):10–19, 2012.

[7] Goldberger, A. L., L. A. N. 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.* 101(23):215–220, 2000.

[8] Goncharova, I. I., D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw. EMG contamination of EEG: spectral and topographical characteristics. *Clin. Neurophysiol.* 114(9):1580–1593, 2003.

[9] Happe, S., P. Anderer, G. Gruber, G. Klösch, B. Saletu, and J. Zeitlhofer. Scalp topography of the spontaneous K-complex and of delta-waves in human sleep. *Brain Topogr.* 15(1):43–49, 2002.

[10] Hobson, D. E., A. E. Lang, W. W. Martin, A. Razmy, J. Rivest, and J. Fleming. Excessive daytime sleepiness and sudden-onset sleep in Parkinson disease: A survey by the Canadian Movement Disorders Group. *J. Am. Med. Assoc.* 287(4):455–463, 2002.

[11] Iber, C., S. Ancoli-Israel, A. Chesson, and S. F. Quan. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* Westchester, IL:American Academy of Sleep Medicine, 2007.

[12] Jobert, M., E. Poiseau, P. Jähnig, H. Schulz, and S. Kubicki. Topographical analysis of sleep spindle activity. *Neuropsychobiology.* 26(4):210–217, 1992.

[13] Jones, B. E. From waking to sleeping: neuronal and chemical substrates. *Trends Pharmacol. Sci.* 26(11):578–586, 2005.

[14] Kemp, B., A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Trans. Biomed. Eng.* 47(9):1185–1194, 2000.

[15] Le, Q. V., J. Ngiam, A. Coates, A, Lahiri, B. Prochnow, and Andrew Y. Ng. On optimization methods for deep learning. In: ICML. Bellvue, WA, 2011.

[16] Liang, S.-F., C.-E. Kuo, Y.-H. Hu, Y.-H. Pan, and Y.-H. Wang. Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models. *IEEE Trans. Instrum. Meas.* 61(6):1649–1657, 2012.

[17] Norman, R. G., I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep.* 23(7):901–908, 2000.

[18] PhysioNet: The Sleep-EDF database [Expanded]: `http://www.physionet.org/physiobank/database/sleep-edfx/` (Accessed January 2015).

[19] PhysioNet: The Sleep-EDF database: `http://www.physionet.org/physiobank/database/sleep-edf/` (Accessed January 2015)

[20] Rechtschaffen, A., and A. Kales. (eds.). *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects.* Washington, DC: Public Health Service, U.S. Government Printing Office, 1968.

[21] Riemann, D., M. Berger, and U. Voderholzer. Sleep and depression–results from psychobiological studies: An overview. *Biol. Psychol.* 57(1):67–103, 2001.

[22] Silber, M. H., S. Ancoli-Israel, M. H. Bonnet, S. Chokroverty, M. M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. A. Keenan, M. H. Kryger, T. Penzel, M. Pressman, and C. Iber. The visual scoring of sleep in adults. *J. Clin. Sleep Med.* 3(2):121–131, 2007.

[23] Spriggs, W. H. *Essentials of Polysomnography: A Training Guide and Reference for Sleep Techicians.* Burlington, MA:Jones & Bartlett Learning, 2014.

[24] Stepnowsky, C., D. Levendowski, D. Popovic, I. Ayappa, and D. M. Rapoport. Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters. *Sleep Med.* 14(11):1199–1207, 2013.

[25] Whitham, E. M., T. Lewis, K. J. Pope, S. P. Fitzgibbon, C. R. Clark, S. Loveless, D. DeLosAngeles, A. K. Wallace, M. Broberg, and J. O. Willoughby. Thinking activates EMG in scalp electrical recordings. *Clin. Neurophysiol.* 119(5):1166–1175, 2008.

[26] Wulff, K., S. Gatti, J. G. Wettstein, and R. G. Foster. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nat. Rev. Neurosci.* 11(8):589–599, 2010.

[27] Yuval-Greenberg, S., O. Tomer, A. S. Keren, I. Nelken, and L. Y. Deouell. Transient induced gamma-band response in EEG as a manifestation of miniature saccades. *Neuron.* 58(3):429–441, 2008.

Figure captions

Figure 1: F_1-score as a function of sleep efficiency.

Figure 2: F_1-score as a function of transitional epochs.

Figure 3: The original manually scored hypnogram (top) and the estimated hypnogram using our algorithm (bottom) for the second night of subject number 2.

Table 1: The Rechtschaffen and Kales sleep staging criteria [20], adapted from [22].

Sleep Stage	**Scoring Criteria**
Non-REM 1 (N1)	50% of the epoch consists of relatively low voltage mixed (2-7 Hz) activity, and $< 50\%$ of the epoch contains alpha (8-13 Hz) activity. Slow rolling eye movements lasting several seconds often seen in early N1.
Non-REM 2 (N2)	Appearance of sleep spindles and/or K complexes and $< 20\%$ of the epoch may contain high voltage (> 75 μV, < 2 Hz) activity. Sleep spindles and K complexes each must last > 0.5 seconds.
Non-REM 3 (N3)	$20\% - 50\%$ (formerly N3) or $> 50\%$ (formerly N4) of the epoch consists of high voltage (> 75 μV), low frequency (< 2 Hz) activity.
REM (R)	Relatively low voltage mixed (2-7 Hz) frequency EEG with episodic rapid eye movements and absent or reduced chin EMG activity.
Wake (W)	$> 50\%$ of the epoch consists of alpha (8-13 Hz) activity or low voltage, mixed (2-7 Hz) frequency activity.

Table 2: The transition rules summarised from the AASM sleep scoring manual [11, Chapter IV: Visual Rules for Adults, pp. 23–31].

Stage Pair	Transition Pattern	Rule	Differentiating Features
N1-N2	N1-{N1,N2}	5.A.Note.1	Arousal, K-complexes, sleep spindles
	(N2-)N2-{N1,N2}(-N2)	5.B.1	K-complexes, sleep spindles
		5.C.1.b	Arousal, K-complexes, sleep spindles
	N2-{N1-N1,N2-N2}-N2	5.C.1.c	Alpha, body movement, slow eye movement
N1-R	R-R-{N1,R}-N2	7.B	Chin EMG tone
		7.C.1.b	Chin EMG tone
		7.C.1.c	Chin EMG tone, arousal, slow eye movement
	R-{N1-N1-N1,R-R-R}	7.C.1.d	Alpha, body movement, slow eye movement
N2-R	R-R-{N2,R}-N2	7.C.1.e	Sleep spindles
	(N2-)N2-{N2,R}-R(-R)	7.D.1	Chin EMG tone
		7.D.2	Chin EMG tone, K-complexes, sleep spindles
		7.D.3	K-complexes, sleep spindles

Curly braces indicate choice between the stages or stage progressions in the set based on the distinctive features, and parentheses indicate optional epochs.

Table 3: Peak frequencies and number of wavelet cycles per frequency for time-frequency analysis using complex Morlet wavelets.

Target Frequency Band	**Target Sleep Stages**	**Frequency or Time Precision**	**Peak Frequency (Hz)**	**Number of Wavelet Cycles**
slow (0.5-2 Hz)	N3	Time	0.7	3
slow (0.5-2 Hz)	N3	Time	1	3
slow (0.5-2 Hz)	N3	Time	1.5	3
slow (0.5-2 Hz)	N3	Time	2	3
K-complex (1.6-4 Hz) [9]	N2	Time	2	3
K-complex (1.6-4 Hz) [9]	N2	Time	3.2	3
delta/theta (2-7 Hz)	N1,R,W	Intermediate	3	5
delta/theta (2-7 Hz)	N1,R,W	Intermediate	4	5
delta/theta (2-7 Hz)	N1,R,W	Intermediate	5	5
delta/theta (2-7 Hz)	N1,R,W	Intermediate	6	5
alpha (8-13 Hz)	N1,W	Frequency	8	10
alpha (8-13 Hz)	N1,W	Frequency	10	10
alpha (8-13 Hz)	N1,W	Frequency	12	10
spindle (12-15 Hz)	N2,N3	Time	12	3
spindle (12-15 Hz)	N2, N3	Time	13	3
spindle (12-15 Hz)	N2,N3	Time	14	3
spindle (12-15 Hz)	N2,N3	Time	15	3
beta (15-30 Hz)	N1 (arousal)	Time	16	3
beta (15-30 Hz)	N1 (arousal)	Time	18	3
beta (15-30 Hz)	W	Intermediate	20	5
gamma (30-100 Hz) *	N1,N2,N3,R,W	Intermediate	40	5

* There is evidence in the literature that features from modalities other than EEG, such as eye movements [27], stage R sleep [13] and EMG activity [8, 25], can manifest themselves in the gamma activity of EEG.

Table 4: Features extracted from the single-channel EEG signal.

Feature	Number	Purpose	Transform
Power of frequency-band power over the entire epoch	22	Capture the overall presence of the particular frequency band in the signal	$\log(x)$
Power of frequency-band power using a sliding window	231	Capture the presence of the particular frequency band in the signal across time	$\log(x)$
Time-domain signal power over the entire epoch	1	Capture the overall amplitude characteristics of the signal	$\log(x)$
Time-domain signal power using a sliding window	11	Capture the amplitude characteristics of the signal over time	$\log(x)$
Frequency-band power-power correlation	242	Capture the relationships between the different frequency bands over time	None
Time-domain signal autocorrelation	50	Capture long-term dependencies in the signal	x^2
ALL	557		

Table 5: Confusion matrix from cross-validation using the Fpz-Cz electrode.

	N1 (algorithm)	**N2 (algorithm)**	**N3 (algorithm)**	**R (algorithm)**	**W (algorithm)**
N1 (expert)	**1654** (60%)	**262** (10%)	**8** (0%)	**366** (13%)	**472** (17%)
N2 (expert)	**1270** (7%)	**13696** (78%)	**1231** (7%)	**760** (4%)	**621** (4%)
N3 (expert)	**7** (0%)	**469** (8%)	**4966** (89%)	**6** (0%)	**143** (3%)
R (expert)	**899** (12%)	**340** (4%)	**0** (0%)	**6164** (80%)	**308** (4%)
W (expert)	**441** (13%)	**34** (1%)	**23** (1%)	**138** (4%)	**2744** (81%)

This confusion matrix is the sum of the confusion matrices from each fold. The numbers in bold are numbers of epochs. The numbers in parentheses are the percentage of epochs that belong to the class classified by the expert (rows) that were classified by our algorithm as belonging to the class indicated by the columns.

Table 6: Comparison between our method and the literature across the five scoring performance metrics (precision, sensitivity, F_1-score, per-stage accuracy, and overall accuracy).

	Scoring performance metrics								
	Precision		**Sensitivity**		**F_1-score**		**Accuracy**		
Study	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Overall
Independent training and testing									
[16]	93	**89**	77	29	82	43	86	63	77
[16]	90	82	73	19	77	31	83	57	73
[3]	92	88	74	36	81	51	84	66	74
	(92)	(86)	(75)	(55)	(82)	(68)	(84)	(74)	(75)
current	**93**	88	**78**	**60**	**84**	**71**	**86**	**76**	**78**
	(94)	(90)	(80)	(65)	(86)	(75)	(88)	(78)	(80)
	0	*-1*	*+1*	*+24*	*+2*	*+20*	*0*	*+10*	*+1*
Non-independent training and testing									
[6]	93	88	77	53	84	68	86	75	77
current	**95**	**91**	**82**	**65**	**88**	**76**	**89**	**79**	**82**
	+2	*+3*	*+5*	*+8*	*+4*	*+8*	*+3*	*+4*	*+5*

For the binary metrics, we report the mean performance (over all five sleep stages) as well as the worst performance (in the most misclassified sleep stage, always stage N1). We present the results for our method using the Fpz-Cz electrode with cross-validation using both independent and non-independent training and testing. The numbers in parentheses are the bootstrap 95% confidence interval bounds for the mean performance across subjects. The signed numbers in italics indicate the improvement (positive) or deterioration (negative) in performance over the second best (improvement) or best (deterioration) method in the literature.

Table 7: Normalised confusion matrices from 20-fold cross-validation using the Fpz-Cz electrode without and with neighbouring epochs. All values are percentages. Pairs of stages with mutual improvement are in bold (N1-N2, N1-R and N2-R).

	Algorithm									
	Without neighbouring epochs					**With neighbouring epochs**				
	N1	**N2**	**N3**	**R**	**W**	**N1**	**N2**	**N3**	**R**	**W**
N1 (expert)	53	**11**	0	**17**	18	60	**9**	0	**13**	17
N2 (expert)	**8**	77	7	**5**	4	**7**	78	7	**4**	4
N3 (expert)	0	8	89	0	3	0	8	89	0	3
R (expert)	**18**	**5**	0	73	5	**12**	**4**	0	80	4
W (expert)	13	1	1	4	82	13	1	1	4	81

Table 8: Correlation between sleep efficiency and percentage of transitional epochs, and scoring performance (F_1-score and overall accuracy).

	Recording parameters			
	Sleep efficiency		**Percentage of transitional epochs**	
Metric	R^2	p-value	R^2	p-value
F_1-score	0.02	0.42	0.04	0.20
Overall accuracy	0.02	0.46	0.05	0.17