SLEEP, 2023, 46, 1–2
https://doi.org/10.1093/sleep/zsad034
Advance access publication 16 February 2023
Editorial

Editorial

Exploiting labels from multiple experts in automated sleep scoring

Samaneh Nasiri[1,2,3,4,*], Wolfgang Ganglberger[2,3,4], Haoqi Sun[2,3,4,5], Robert J. Thomas[2,4] and M. Brandon Westover[1,2,3,4,5]

[1]Department of Neurology, Massachusetts General Hospital (MGH), Boston, MA, USA,
[2]Harvard Medical School, Boston, MA, USA,
[3]Clinical Data Animation Center (CDAC), Boston, MA, USA,
[4]McCance Center for Brain Health, MGH, Boston, MA, USA and
[5]Department of Medicine, Division of Pulmonary, Critical Care & Sleep, Beth Israel Deaconess Medical Center, Boston, MA, USA

*Corresponding author. Samaneh Nasiri, 399 Revolution Drive, Suite 1160, Somerville, MA 02145, USA. Email: snasiri@mgh.harvard.edu.

The current "ground truth" for sleep staging is manual scoring of the electroencephalogram following American Academy of Sleep Medicine (AASM) rules [1]. These rules specify how to label each 30-s epoch into one of five stages: Wake (W), Rapid Eye Movement (REM), and Non-REM 1–3 (N1, N2, and N3). However, AASM rules are not precise enough to be directly programmed into a computer. Moreover, NREM sleep from a biological standpoint exists along a continuous spectrum rather than in discrete stages [2]. This imprecision and artificial discretization lead to variable and imperfect inter-rater scoring agreement, ranging from 60% internationally to ~80% within the same institute [3]. Recently, several papers have developed deep neural networks for automated sleep staging [4–6]. This "AI-enabled sleep staging," although proposed as a way to achieve objective and repeatable sleep staging, is ultimately limited by imprecision in the "gold standard" training labels. This is particularly true for AI methods which consider datasets annotated by a single scorer [7, 8]. One way to overcome the problem of noisy labels is to utilize datasets scored by multiple experts.

Most prior efforts to train sleep staging models using labels from multiple experts have combined labels using a simple majority vote scheme, which does not make optimal use of information about disagreement in voting among experts [9, 10]. In the current issue of *SLEEP*, Fiorillo et al. propose a framework for training deep learning algorithms that leverages labels from multiple experts more effectively than majority voting.

The authors adopt "label smoothing" to leverage multiple labels from different scorers efficiently [11]. Label smoothing assigns a non-zero probability to multiple classes, treating them as "soft," as opposed to a baseline approach that uses "hard" labels, in which one class is treated as correct with 100% confidence [12]. In the baseline approach, hard labels are assigned by majority vote. In case of ties, the correct answer is taken to be the vote of the "most reliable" rater (the rater whose answers most frequently agrees with the majority). The paper then compares two label smoothing approaches to the baseline approach.

(1) *Label smoothing by a uniform distribution*: In this approach, if the majority label for a given epoch is wake (W), then the "hard label" would be $L = [p_w, p_{N1}, p_{N2}, p_{N3}, p_{REM}] = [1, 0, 0, 0, 0]$, where the 5 positions represent the probability that we assign to each of the 5 possible sleep stages. The smoothed label based on the uniform distribution would then be a mixture of the original hard label L and the uniform distribution vector $U = [\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}]$ where each stage is assigned a 1/5 probability, $L_{su} = \alpha L + (1 - \alpha)U$. Here, the "mixing" parameter α is a number between 0 and 1 that determines how much weight is given to the hard label vector vs. the uniform distribution. This number is determined empirically. For illustration suppose this number $\alpha = 0.9$, meaning that 90% of the weight is given to the hard label (majority vote) L, and 10% to the uniform distribution U. In this case, the smoothed label would be $L_{su} = [0.920.020.020.020.02]$. Note that most of the weight is still given to the label that received the majority vote, but the smoothed label allows for some uncertainty and thus might be expected to prevent the model from becoming overconfidence about the correct label for this example. Specifically, this smoothed label gives 90% of the total probability to the majority label and distributes the remaining 10% equally among the other possibilities. Note that the total weight (probability) of the smoothed label still adds up to one.

(2) *Label smoothing by soft consensus*: The second label smoothing strategy is based on Soft Consensus (SC). The SC for a given epoch is the vector containing the proportions of experts who voted for each stage. For example, if 6 experts labeled a given epoch, where 4 labeled the epoch as wake (W), and 2 as N1, then the SC vector would be $SC = [\frac{4}{6}, \frac{2}{6}, 0, 0, 0]$. The smoothed label in this case would then be $L_{sc} = \alpha L + (1 - \alpha)SC$. If we again assume as the value for the smoothing parameter $\alpha = 0.9$, then the smoothed label is $L_{sc} = [0.970.03000]$ This smoothed label again gives 90% of the probability to the majority label, but

then distributes the remaining 10% unequally, according to the "soft consensus."

For comparing the two label smoothing methods with the baseline majority voting approach, three open-access datasets scored by multiple physicians have been used; ISRC (Inter-scorer Reliability Cohort; N = 70 PSGs, n = 6 scorers) [13]; DOD-H (Dreem Open Dataset—Healthy; N = 25 PSGs, n = 5 scorers), and DOD-O (Dreem Open Dataset—Obstructive; N = 55 PSGs, n = 5 scorers) [10]. The authors used two deep learning-based sleep staging algorithms, DeepSleepNet-Lite (DSN-L) [14] and SimpleSleepNet (SSN) [10] to classify sleep stages into the five AASM sleep stages (Wake, REM, N1, N2, and N3). The authors used K-fold cross-validation for training each model (for ISRC, DOD-H, and DOD-O, K = 10, 25, and 10, respectively). During K-fold cross-validation, each dataset is split into K number of folds, onefold is considered as a test set, and the model is trained and validated on the remaining subjects' data in K – 1 folds. This process is repeated until each fold takes a turn being the test set [15].

The authors use an averaged cosine similarity metric (ACS) to quantify the similarity between the hypnodensity graph generated by the models using label smoothing with SC and the hypnodensity graph generated by the scorer consensus (majority vote). The hypnodensity graph provides a probability distribution over sleep stages per epochs (i.e. each 30-s window). The authors used ACS to quantitatively evaluate the ability of the model to adapting to the consensus of the group of scorers, where a higher ACS value means a higher similarity between these two hypnodensity graphs. Based on ACS, the label smoothing by SC enabled both deep learning models to learn to perform substantially better than when label smoothing was not utilized, and better than label smoothing based on the uniform distribution.

A key limitation of this study is that the datasets used for training and evaluating the proposed method are small (N = 70, 25, 55 for three different datasets). To train a staging model that generalizes across clinically relevant parameters (e.g. age, gender, ethnicity, medical and neurological disorders) would require large datasets scored by multiple experts. However, this is challenging because no currently available datasets are large both in terms of number of patients and number of scorers. In this direction, crowdsourcing could be a viable solution to create larger multiply scored datasets [16]. Another limitation is that the number of experts needed to overcome the noise inherent in the human sleep staging process is not known. Finally, it is not clear how best to select a group of experts, although some guidance is available from other fields where crowd sourcing has proven effective; for example, the "crowd" should be large and diverse, and the judgments must be independent (e.g. from different institutions).

Despite these limitations, the proposed method is a welcome addition to the literature. Label smoothing provides a principled approach to leveraging the variability among multiple scorers to improve the performance of automated sleep scoring algorithms.

Authors' Contribution

All authors have been involved in the following aspects, but can be mostly categorized as: drafting the work: S.N., W.G., H.S., and M.B.W.; revising critically: S.N., W.G., H.S., M.B.W., and R.J.T.

Funding

Dr. Westover's laboratory is supported by grants from the NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598) and NSF (2014431). Dr. Westover is a co-founder of Beacon Biosignals. Dr. Thomas received grant support from an American Academy of Sleep Medicine Foundation Strategic Research Award, 1RF1AG064312, and reports consulting (Jazz Pharmaceuticals, GLG Councils, Guidepoint Global); grant support, license, and intellectual property (DeVilbiss Healthcare) for auto-CPAP algorithm; and license, intellectual property, and royalties (MyCardio, LLC) for ECG/PPG spectrogram. Dr. Nasiri is a consultant at LifeBell AI, LLC. No other disclosures were reported.

Disclosure Statement

None declared.

References

1. Berry RB, *et al*. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Vol. **176**. Darien, IL: American Academy of Sleep Medicine; 2012:2012.
2. Cesari M, *et al*. Sleep modelled as a continuous and dynamic process predicts healthy ageing better than traditional sleep scoring. *Sleep Med*. 2021;**77**:136–146.
3. Basner M, *et al*. Inter-rater agreement in sleep stage classification between centers with different backgrounds. *Somnologie*. 2008;**12**(1):75–84.
4. Sun H, *et al*. Sleep staging from electrocardiography and respiration with deep learning. *Sleep*. 2020;**43**(7). doi:10.1093/sleep/zsz306
5. Biswal S, *et al*. Expert-level sleep scoring with deep neural networks. *J Am Med Inform Assoc*. 2018;**25**(12):1643–1650.
6. Nasiri S, *et al*. Attentive adversarial network for large-scale sleep staging. In: Machine Learning for Healthcare Conference, PMLR; 2020:457–478.
7. Nasiri S, *et al*. Boosting automated sleep staging performance in big datasets using population subgrouping. *Sleep*. 2021;**44**(7). doi:10.1093/sleep/zsab027
8. Danker-Hopfe H, *et al*. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res*. 2004;**13**(1):63–69.
9. Stephansen JB, *et al*. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun*. 2018;**9**(1):1–15.
10. Guillot A, *et al*. Dreem open datasets: multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Trans Neural Syst Rehabil Eng*. 2020;**28**(9):1955–1965.
11. Yuan L, *et al*. Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020:3903–3911.
12. Lienen J, *et al*. From label smoothing to label relaxation. *Proc AAAI Conference Artif Intell*. 2021;**35**(10):8583–8591.
13. Kuna ST, *et al*. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*. 2013;**36**(4):583–589. doi:10.5665/sleep.2550
14. Fiorillo L, *et al*. DeepSleepNet-Lite: a simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE Trans Neural Syst Rehabil Eng*. 2021;**29**:2076–2085.
15. Memar P, *et al*. A novel multi-class EEG-based sleep stage classification system. *IEEE Trans Neural Syst Rehabil Eng*. 2017;**26**(1):84–95.
16. Warby SC, *et al*. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nat Methods*. 2014;**11**(4):385–392.