The Journal of Open Source Software

Automated Sleep Stage Scoring Using *k*-Nearest Neighbors Classifier

Tamás Kiss[1, 2], **Stephen Morairty**[3], **Michael Schwartz**[3], **Thomas S. Kilduff**[3], **Derek L. Buhl**[1, 4], **and Dmitri Volfson**[1, 4]

1 Global Research and Development, Pfizer Inc, Groton, CT, USA **2** Department of Computational Sciences, Wigner Research Centre for Physics, Budapest, Hungary **3** Center for Neuroscience, SRI International, Menlo Park, CA, USA **4** Current affiliation – Takeda Pharmaceuticals, Inc., Cambridge, MA, USA

DOI: 10.21105/joss.02377

Software

- Review
- Repository
- Archive

Editor: Olivia Guest

Reviewers:

- @Emma-k-ward
- @sbuergers
- @Trisbek
- @samikakumar

Submitted: 20 May 2020
Published: 24 July 2020

Polysomnographic Sleep Stage Scoring

Many features of sleep, such as the existence of rapid eye movement (REM) sleep or non-REM sleep stages, as well as some of the underlying physiological mechanisms controlling sleep, are conserved across different mammalian species. Sleep research is important to understanding the impact of disease on circadian biology and optimal waking performance, and to advance treatments for sleep disorders, such as narcolepsy, shift work disorder, non-24 sleep-wake disorder, and neurodegenerative disease. Given the evolutionary relatedness of mammalian species, sleep architecture and changes therein may provide reliable translational biomarkers for pharmacological engagement in proof-of-mechanism clinical studies.

Key physiological indicators in sleep include electroencephalography (EEG) or electrocorticography, electrooculography (EOG), and electromyography (EMG). Polysomnography (PSG) is the simultaneous collection of some or all of these measurements and is typically performed in a specialized sleep laboratory. Determination of the wake or sleep stage someone is in (i.e., wake, REM sleep, or non-REM sleep, which is broken down into stages 1, 2, or 3), relies on the judgment of a trained professional who scores the data based on the standardized criteria for the recording and staging of human PSG set forth by Berry et al. (2017). Disagreement between individual recordings might arise due to differences in instrumentation or to the subjective opinion of the individual scoring the stages. Animal sleep studies show even greater variability (Robert, Guilpin, & Limoge, 1999), as each laboratory uses methods that best suit their individual needs (e.g., electrode/reference positions, muscle choice for EMG implantation, use of EOG, etc.). While these technical differences make it difficult to compare studies, the variability in scoring of sleep stages makes it even more challenging. Although numerous scoring algorithms exist, most are unreliable, especially following drug treatment. After nearly half a century of PSG studies, the gold standard of scoring sleep architecture remains a complete and thorough examination of the PSG signals, which are scored in 4-, 10-, or 12-second epochs in animal studies and 30-second epochs in human studies, making it very difficult to screen through drugs in animal studies and cumbersome to implement large clinical trials.

Applications and Advantage

To expedite the tedious process of visually analyzing PSG signals and to further objectivity in the scoring procedure, a number of sleep staging algorithms have been developed both for animals (Barger, Frye, Liu, Dan, & Bouchard, 2019; Bastianini et al., 2014; Stephenson, Caron, Cassel, & Kostela, 2009; Vladimir, Ting-Chuan, Yuting, Bryan, & Steven, 2020) and human subjects (Gunnarsdottir et al., 2020; Penzel & Conradt, 2000; Zhang et al., 2020) as reviewed most recently by Fiorillo et al. (2019) and Faust, Razaghi, Barika, Ciaccio, & Acharya (2019). However, computer-based methods are typically tested on data obtained from healthy subjects or control animals, and performance is assessed only in a few cases in subjects with sleep disorders or following drug treatment (Allocca et al., 2019; Boostani, Karimzadeh, & Nami, 2017). Furthermore, scoring sleep for hundreds of animals in a typical preclinical drug discovery effort often becomes a bottleneck and a potential source of subjectivity affecting research outcomes.

In this paper, we present an automated approach intended to eliminate these potential issues. The initial application of our approach is for basic and discovery research in which experiments are conducted in large cohorts of rodents, with the expectation that results can be translated to higher-order mammals or even humans. Building on features classically extracted from EEG and EMG data and machine learning-based classification of PSG, this approach is capable of staging sleep in multiple species under control and drug-treated conditions, facilitating the detection of treatment-induced changes or other manipulations (e.g., genetic). Using human interpretable features calculated from EEG and EMG will be important to understand drug mechanisms, for prediction of treatment outcomes, and as biomarkers or even translational biomarkers. For example, one of the features used by the algorithm is the power in the theta frequency band (called `eeg_theta` in the code), which is the 4 Hz to 12 Hz range and it is known that an increase of theta activity together with low EMG activity (our relevant features are called `emg_high` and `emg_RMS`) are the hallmark of REM sleep (see the figure in Wikipedia contributors (2020)). However, theta power is also associated with other phenomena, like anxiety (John, Kiss, Lever, & Érdi, 2014), thus our `eeg_theta` feature, besides being used for sleep scoring can also be used as a biomarker of drug effect.

Multiple software applications have been developed to address the problem of automated sleep stage scoring. In their comparative review, Boostani et al. (2017) found that the best results could be achieved when entropy of wavelet coefficients along with a random forest classifier were chosen as feature and classifier, respectively. Another recent method (Miladinović et al., 2019) used cutting-edge machine learning methods combining a convolutional neural network-based architecture to produce domain invariant predictions integrated with a hidden Markov model to constrain state dynamics based upon known sleep physiology. While our method also builds on machine learning techniques, it is based on interpretable features and uses a simpler algorithm for classification – which should make it an ideal choice for the broader community as well as for sleep experts who might not be too familiar with complex machine learning approaches. Furthermore, we chose not to constrain the number of identifiable sleep/wake states or the probability of transition from one state to another, as we and others have found that drug interventions (Harvey et al., 2013) and disease processes (de Mooij et al., 2020) tend to change not only the amount of time spent in different sleep stages but their transition probabilities as well. Finally, our method is a supervised method that requires a training set. While this might seem to be a disadvantage over non-supervised methods, we have found that drug treatment or pathological conditions can result in sleep stages not observed in healthy controls. Thus, the algorithm must be trained to these new stages.

Brief Software Description

Our software package, implemented in Matlab, is available for download on GitHub (Kiss et al., 2020). Automatic sleep staging consists of the classical consecutive steps of machine learning-based sleep scoring algorithms Figure 1. First, offline stored EEG and EMG data are loaded into memory to allow for the uniform processing of time-series data and segmented into consecutive 10-second, non-overlapping epochs that correspond to manually scored epochs. Second, features are extracted from the raw signal for all epochs. Features consist of the power contained in physiologically-relevant frequency bands, as well as Hjorth parameters for both EEG and EMG data. Third, features undergo a pre-processing step including the following operations: unusable epochs that contain too much noise or contain no signal are removed. Features are then transformed using the logarithm function making feature distributions more Gaussian-like, thereby facilitating subsequent machine classification. Finally, each feature is normalized to its median wake value within an animal to enable usability of the algorithm across laboratories. Wake periods can be identified before running the algorithm using the manually-scored training set or an experiment can be performed such that a given period is expected to be comprised of an extended period of wakefulness. Following feature extraction, a combined filter and wrapper method-based feature selection step is applied. This step ensures that features with the most predictive value are chosen and also helps to prevent over-fitting. For classification, the *k*-nearest neighbors classifier is used on data pre-processed following the procedure described above.

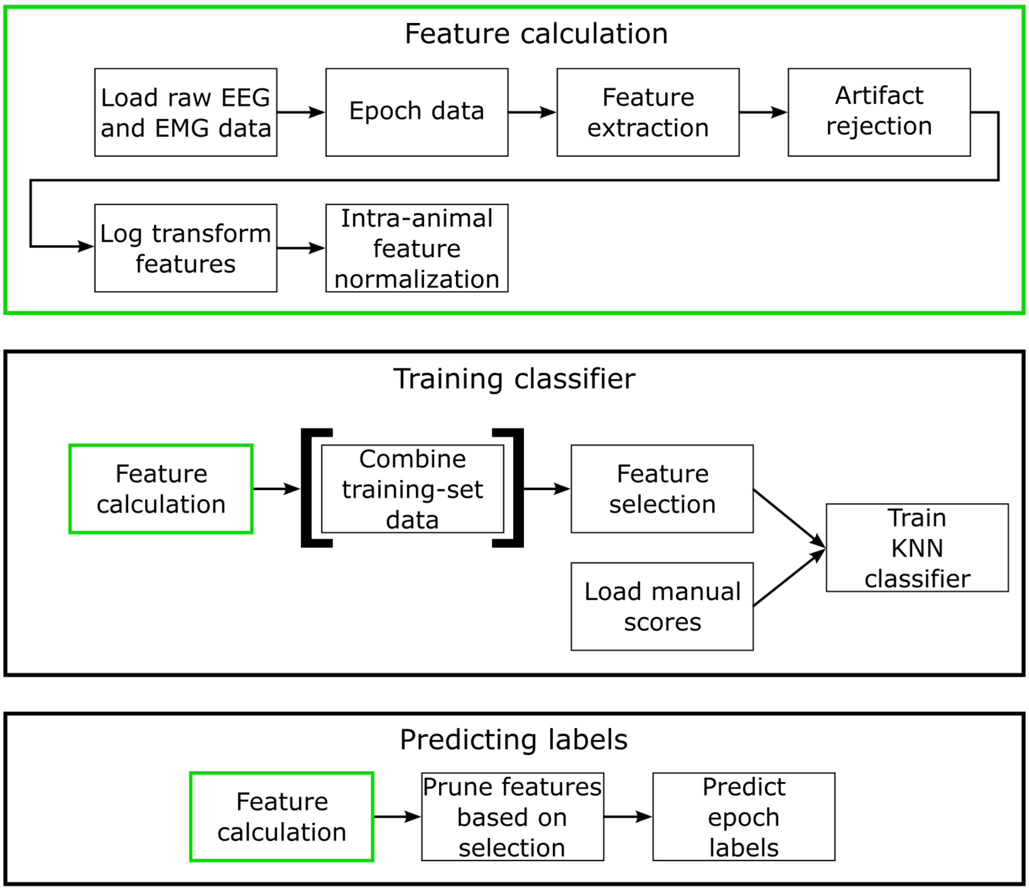

Figure 1: Summary of training and using the *k*-nearest neighbors algorithm for predicting sleep stage labels.

The algorithm was used to predict sleep stages in mice (Figure 2), rats (Figure 3) and non-human primates (data not shown). Prediction accuracy was found to depend on a number of parameters of the input data, including consistency of manual scores and physiological signals, as well as the amount of artifacts. Furthermore, relative frequency of predicted labels can influence efficacy, with rare labels being harder to predict. The code on GitHub (Kiss et al., 2020) accompanying this paper contains the abridged version of two datasets, one from male Trace Amine-Associated Receptor 1 (TAAR1) knockout mice described in detail in Schwartz, Palmerston, Lee, Hoener, & Kilduff (2018) (Figure 2) and the other from male Sprague-Dawley rats collected in the Sleep Neurobiology Laboratory at SRI International (Figure 3).

The rodents in both datasets received an oral dosing of a water-based vehicle solution.

Three labels were predicted: wake (W), non-REM sleep (NR), and REM sleep (R), and prediction efficacy was calculated. (However, note that any number of stages can be trained depending on how elaborate the manual scoring is.) The model was first used to train a single classifier merging training data from all animals (Figure 2 A, Figure 3 A), then individual models were trained, one for each animal (Figure 2 B, Figure 3 B). The GitHub repository includes additional information on prediction accuracy, including detailed values of true and false positive rates, as well as a method to deal with imbalanced data.

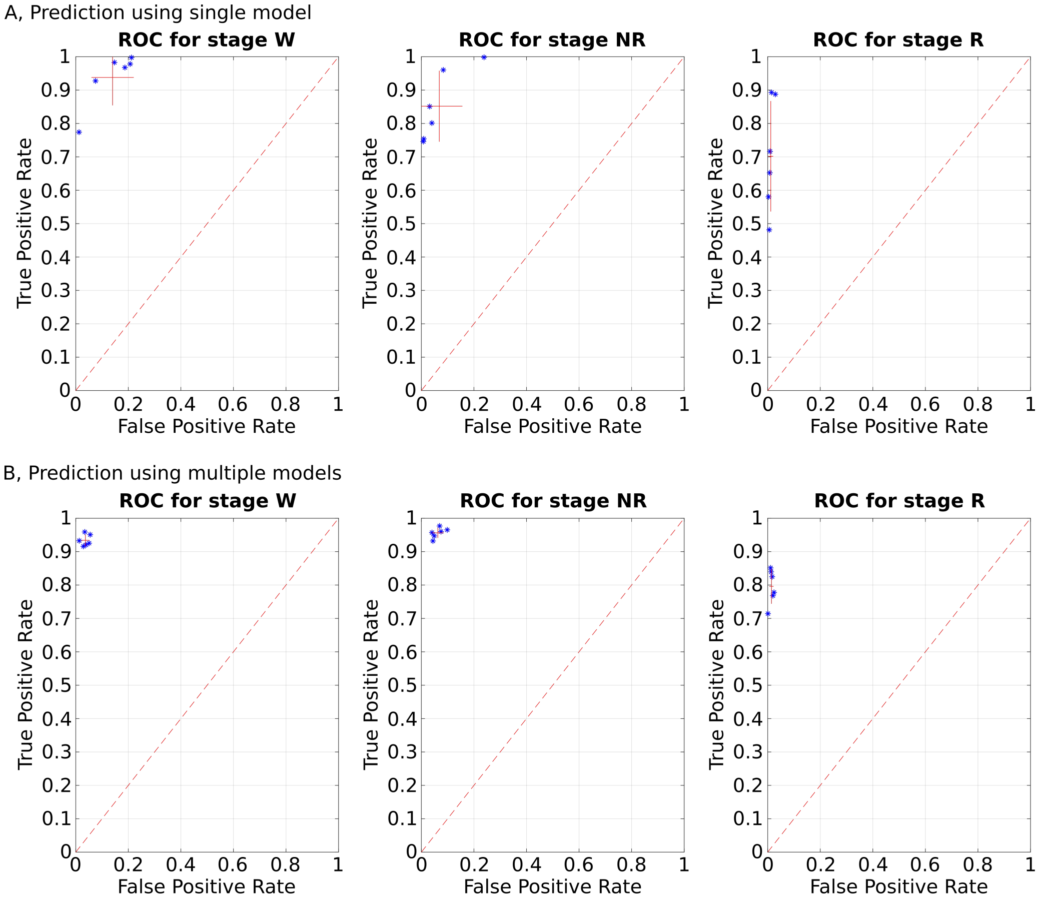

Figure 2: Estimation of prediction accuracy for the transgenic mouse data. For each state (wake – W, non-REM – NR, REM – R) and animal (points on plots) true and false positive rates are calculated. Red crosses denote mean and SEM. In A, training data was merged and one single classifier was trained to predict sleep stages of all animals. In B, an individual classifier was trained for each animal separately.

State labels were predicted the same way for the rat data (the same set of GitHub scripts were run) and prediction accuracy represented on Figure 3 shows very similar results.

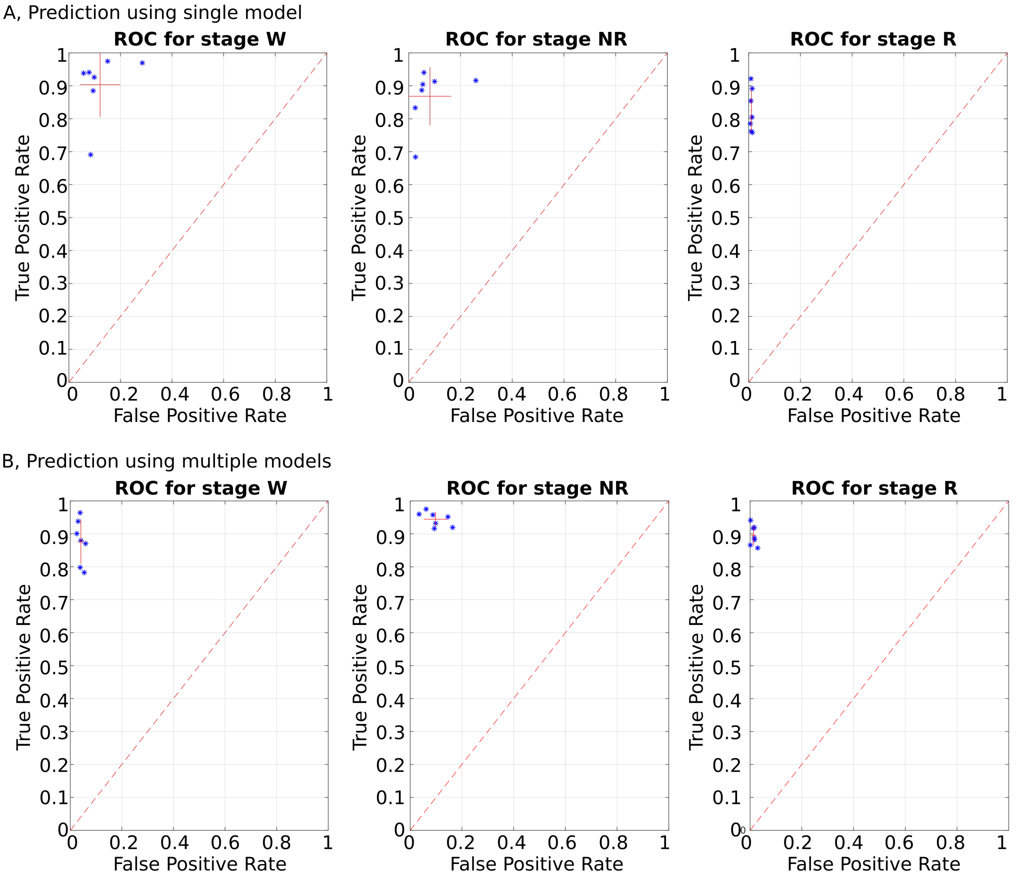

Figure 3: Estimation of prediction accuracy for the rat data. Prediction and figure set up as in Figure 2.

Acknowledgments

TK, DV, and DLB were full time employees and shareholders of Pfizer Inc. during development of this software package. This work was supported by Pfizer Inc. and SRI International.

Author contributions

DV and TK developed the software, SM, MS, TSK, and DLB contributed data for development and testing, all authors took part in debugging and testing the software, and all authors wrote or contributed to writing the manuscript.

References

Allocca, G., Ma, S., Martelli, D., Cerri, M., Del Vecchio, F., Bastianini, S., Zoccoli, G., et al. (2019). Validation of 'Somnivore', a Machine Learning Algorithm for Automated Scoring and Analysis of Polysomnography Data. *Front Neurosci*, *13*, 207. doi:https://doi.org/10.3389/fnins.2019.00207

Barger, Z., Frye, C. G., Liu, D., Dan, Y., & Bouchard, K. E. (2019). Robust, automated sleep scoring by a compact neural network with distributional shift correction. *PLoS ONE*, *14*(12), e0224642. doi:https://doi.org/10.1371/journal.pone.0224642

Bastianini, S., Berteotti, C., Gabrielli, A., Del Vecchio, F., Amici, R., Alexandre, C., Scammell, T. E., et al. (2014). SCOPRISM: A new algorithm for automatic sleep scoring in mice. *Journal of Neuroscience Methods*, *235*, 277–284. doi:https://doi.org/10.1016/j.jneumeth.2014.07.018

Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M. T., et al. (2017). AASM scoring manual updates for 2017 (version 2.4). *Journal of Clinical Sleep Medicine*, *13*(05), 665–666. doi:10.5664/jcsm.6576

Boostani, R., Karimzadeh, F., & Nami, M. (2017). A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer Methods and Programs in Biomedicine*, *140*, 77–91. doi:https://doi.org/10.1016/j.cmpb.2016.12.004

de Mooij, S. M. M., Blanken, T. F., Grasman, R. P. P. P., Ramautar, J. R., Van Someren, E. J. W., & Maas, H. L. J. van der. (2020). Dynamics of sleep: Exploring critical transitions and early warning signals. *Comput Methods Programs Biomed*, *193*, 105448. doi:https://doi.org/10.1016/j.cmpb.2020.105448

Faust, O., Razaghi, H., Barika, R., Ciaccio, E. J., & Acharya, U. R. (2019). A review of automated sleep stage scoring based on physiological signals for the new millennia. *Comput Methods Programs Biomed*, *176*, 81–91. doi:https://doi.org/10.1016/j.cmpb.2019.04.032

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P. L., Favaro, P., Roth, C., Bargiotas, P., et al. (2019). Automated sleep scoring: A review of the latest approaches. *Sleep Med Rev*, *48*, 101204. doi:https://doi.org/10.1016/j.smrv.2019.07.007

Gunnarsdottir, K. M., Gamaldo, C., Salas, R. M., Ewen, J. B., Allen, R. P., Hu, K., & Sarma, S. V. (2020). A novel sleep stage scoring system: Combining expert-based features with the generalized linear model. *J Sleep Res*, e12991. doi:https://doi.org/10.1111/jsr.12991

Harvey, B. D., Siok, C. J., Kiss, T., Volfson, D., Grimwood, S., Shaffer, C. L., & Hajós, M. (2013). Neurophysiological signals as potential translatable biomarkers for modulation of metabotropic glutamate 5 receptors. *Neuropharmacology*, *75*, 19–30. doi:https://doi.org/10.1016/j.neuropharm.2013.06.020

John, T., Kiss, T., Lever, C., & Érdi, P. (2014). Anxiolytic drugs and altered hippocampal theta rhythms: the quantitative systems pharmacological approach. *Network*, *25*(1-2), 20–37. doi:https://doi.org/10.3109/0954898x.2013.880003

Kiss, T., Morairty, S., Schwartz, M., Kilduff, T. S., Buhl, D. L., & Volfson, D. (2020). *k*NNSS: A Matlab package for automated sleep scoring using the k-nearest neighbors algorithm. *GitHub repository*. GitHub. Retrieved from https://github.com/teamPSG/kNN_Sleep_Scorer_kNNSS

Miladinović, D., Muheim, C., Bauer, S., Spinnler, A., Noain, D., Bandarabadi, M., Gallusser, B., et al. (2019). SPINDLE: End-to-end learning from eeg/emg to extrapolate animal sleep scoring across experimental settings, labs and species. *PLOS Computational Biology*, *15*(4), 1–30. doi:10.1371/journal.pcbi.1006968

Penzel, T., & Conradt, R. (2000). Computer based sleep recording and analysis. *Sleep Medicine Reviews*, *4*(2), 131–148. doi:https://doi.org/10.1053/smrv.1999.0087

Robert, C., Guilpin, C., & Limoge, A. (1999). Automated sleep staging systems in rats. *Journal of Neuroscience Methods*, *88*(2), 111–122. doi:https://doi.org/10.1016/S0165-0270(99)00027-8

Schwartz, M. D., Palmerston, J. B., Lee, D. L., Hoener, M. C., & Kilduff, T. S. (2018). Deletion of Trace Amine-Associated Receptor 1 Attenuates Behavioral Responses to Caffeine. *Front Pharmacol*, *9*, 35. doi:https://doi.org/10.3389/fphar.2018.00035

Stephenson, R., Caron, A. M., Cassel, D. B., & Kostela, J. C. (2009). Automated analysis of sleep–wake state in rats. *Journal of Neuroscience Methods*, *184*(2), 263–274. doi:https://doi.org/10.1016/j.jneumeth.2009.08.014

Vladimir, S., Ting-Chuan, W., Yuting, X., Bryan, J. H., & Steven, V. F. (2020). A Deep Learning Approach for Automated Sleep-Wake Scoring in Pre-Clinical Animal Models. *J. Neurosci. Methods*, *337*, 108668. doi:https://doi.org/10.1016/j.jneumeth.2020.108668

Wikipedia contributors. (2020). Rapid eye movement sleep — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Rapid_eye_movement_sleep&oldid=966961239.

Zhang, X., Xu, M., Li, Y., Su, M., Xu, Z., Wang, C., Kang, D., et al. (2020). Automated multi-model deep neural network for sleep stage scoring with unfiltered clinical data. *Sleep Breath*. doi:https://doi.org/10.1007/s11325-019-02008-w