Transfer Learning for Subject-Independent Sleep Deprivation Detection from Resting-State EEG

Daya Kumar[1, 2], **Uday Devulapalli**[1, 2], **Saptharishi Lalgudi Ganesan**[3, 4, 5, 6], **Apurva Narayan**[1]
[1]Western University, London, ON, Canada
[2]International Center for Applied Systems Science for Sustainable Development (ICASSSD), Cambridge, ON, Canada
[3]Department of Paediatrics & Clinical Neurological Sciences, Schulich School of Medicine & Dentistry, Western University, London, ON, Canada
[4]Paediatric Critical Care Medicine, Children's Hospital - London Health Sciences Centre, London, ON, Canada
[5]Western Institute for Neuroscience, Western University, London, ON, Canada
[6]Children's Health Research Institute, London Health Sciences Center Research Institute (LHSC-RI), London, ON, Canada
dkumar55@uwo.ca, udevulap@uwo.ca, rishi.ganesan@lhsc.on.ca, apurva.narayan@uwo.ca

Abstract

Sleep deprivation (SD) impairs cognition and heightens safety risks, yet reliable electroencephalography (EEG)-based detection remains challenging in low-data settings. We evaluated transfer learning with a compact Convolutional Neural Network (CNN) (EEGNetv4) to classify SD versus well-rested wakefulness using an open-source EEG dataset containing eyes-open resting-state data from 71 healthy young adults. EEGNetv4 was initialized with publicly available weights pretrained on an Event-Related Potential (ERP) dataset. Shape-compatible layers were transferred and frozen, with the remaining layers trained on the target data. Baselines comprised EEGNetv4, a bidirectional Long Short-Term Memory (LSTM), and a Transformer model, each trained without pretraining. Five-fold subject-independent cross-validation was used to evaluate model performance. EEGNetv4 with transfer learning achieved the highest mean accuracy (70.79% ± 4.17), outperforming EEGNetv4 trained from scratch (65.75% ± 5.48), the Transformer (63.35% ± 2.78), and the LSTM (61.70% ± 3.20). These findings suggest that leveraging pretrained EEG representations can enhance subject-generalizable SD classification in small-sample contexts, supporting transfer learning as a pragmatic strategy for neurophysiological applications.

Introduction

Sleep is a fundamental biological process essential for cognitive, emotional, and physiological health (Krause et al. 2017). Insufficient sleep is consistently associated with impairments in executive function, decision-making, and psychomotor vigilance (Lim and Dinges 2010; Van Dongen et al. 2003). In safety-critical domains such as healthcare and transportation, sleep-related fatigue has been linked to higher rates of serious errors and accidents, underscoring the need for reliable methods to detect and classify sleep deprivation (SD) (Landrigan et al. 2004; Williamson et al. 2011).

Electroencephalography (EEG) has emerged as one of the most informative neurophysiological modalities for characterizing sleep and its deprivation. EEG signals capture oscillatory brain activity across multiple frequency bands, providing direct insights into the neural mechanisms underlying fatigue, attention, and cognitive decline (Makeig et al. 2004; Klimesch 1999). Traditional approaches for analyzing EEG in SD research have focused on handcrafted features derived from spectral power, connectivity measures, or statistical descriptors, which are then classified using conventional machine learning models such as support vector machines or random forests (Wan Masri et al. 2024; Baygin 2025; Ren et al. 2021; Dong, Lin, and Chang 2022). While these approaches have yielded promising results, their reliance on manual feature engineering limits generalizability across datasets and often requires large amounts of labeled data for robust performance (Roy et al. 2019).

Recent advances in deep learning (DL) have shifted attention toward data-driven methods capable of automatically extracting hierarchical representations from raw EEG signals. Convolutional neural networks (CNNs) and recurrent architectures such as long short-term memory (LSTM) networks have demonstrated strong performance in various EEG-based classification tasks, such as sleep staging, by learning spatio-temporal features directly from the data (Supratak et al. 2017; Chambon et al. 2018; Schirrmeister et al. 2017). However, a central challenge in applying deep learning to SD classification is the scarcity of large, labeled datasets (Lashgari, Liang, and Maoz 2020; Craik, He, and Contreras-Vidal 2019). EEG data are costly to acquire, and public datasets are typically small and heterogeneous, which increases the risk of overfitting and hampers model generalization (He et al. 2023).

Transfer learning (TL) offers a compelling solution to this problem by leveraging knowledge learned from a source task or dataset to improve performance on a target domain with limited data (Zhuang et al. 2021; Azab et al. 2018). In computer vision and natural language processing, transfer learning has enabled substantial improvements in accuracy and efficiency (Azunre 2021; Brodzicki et al. 2020), yet its application to EEG-based SD classification remains relatively underexplored. By adapting pretrained models trained on large-scale EEG or related datasets, it is possible to capture generalizable neural representations and fine-tune them for SD detection, thereby reducing dependence on extensive

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

labeled training data.

In this study, we investigate the feasibility of using a transfer learning approach to classify SD using an open-source EEG dataset. Our aim is to assess whether pretrained neural representations can enhance SD classification accuracy compared to models trained from scratch, while maintaining generalizability across subjects. This work contributes to the development of reliable, scalable, and practical tools for monitoring cognitive impairment associated with sleep loss.

Related Work

Deep learning has been widely applied to EEG for tasks such as automatic sleep staging and drowsiness detection, leveraging convolutional and recurrent architectures to capture spatio-temporal dynamics in neural signals (Phan et al. 2021; Mousavi, Afghah, and Acharya 2019). These approaches have achieved strong performance in large-scale datasets, but their reliance on extensive labeled data and substantial computational resources limits their applicability in domains with smaller cohorts, such as SD classification. Although open-source SD datasets have begun to standardize evaluation protocols, they remain modest in size and subject to high inter-individual variability, restricting the generalizability of purely DL-based approaches (Xiang et al. 2024).

To address these limitations, TL has emerged as a promising strategy. By pretraining on large EEG or sleep corpora and subsequently fine-tuning on smaller target datasets, TL can improve performance in low-resource contexts while mitigating issues of overfitting. Several studies have demonstrated the effectiveness of TL in the sleep domain, showing that pretrained representations can be successfully adapted for tasks such as sleep staging, sleep quality, and fatigue detection (Wang et al. 2022; He et al. 2023; Zhang, Zheng, and Lu 2017; Olesen et al. 2020). These studies highlight TL as a viable solutions to overcome the data scarcity and heterogeneity that characterizes SD research.

Recent machine learning studies have also explored direct EEG-based classification of SD. Masri et al. examined the influence of channel selection using Random Forest, k-Nearest Neighbors, Support Vector Machine, and Artificial Neural Networks on data from 10 participants, reporting accuracies as high as 99.7% with 19-channel recordings and 94% with only frontal channels (Wan Masri et al. 2024). Similarly, Baygin introduced the Melatonin Pattern (MelPat) feature extraction method, which combined wavelet decomposition, feature selection, and SVM classification to achieve 97.7% accuracy on an open-access dataset of 71 subjects (Baygin 2025). While these studies demonstrate the potential of EEG for SD detection, both relied on random train–test splits rather than subject-level separation, raising concerns about data leakage and limiting conclusions regarding cross-subject generalizability.

Methodology

Dataset

We used an open-access eyes-open resting-state EEG dataset comprising 71 healthy young adults (34 females, 37 males; age range 17-23 years, mean 20 $\pm$ 1.44 years) (Xiang et al. 2024). Participants reported no history of psychiatric or sleep disorders, and compliance with normal sleep patterns was confirmed via screening and actigraphy. Each subject completed two sessions in a counterbalanced within-subject design: (i) well-rested wakefulness and (ii) SD (24-30 h of sustained wakefulness), separated by 7-30 days. Sessions were scheduled within matched morning or afternoon timeframes to minimize circadian effects, and both followed identical protocols including cognitive testing (e.g., Psychomotor Vigilance Task) prior to EEG recording. Resting-state EEG was acquired using a 61-channel system (Brain Products GmbH, Germany) at 500 Hz (impedance <5 kΩ) during five minutes of eyes-open fixation. Data from one participant were excluded due to incompleteness.

For transfer learning, we employed publicly available pretrained weights of EEGNetv4 (Lawhern et al. 2018), which were originally trained on an ERP dataset (Lee et al. 2019) included in the MOABB benchmark (Aristimunha et al. 2025). This dataset comprises recordings from 54 healthy subjects, using a 62-channel Ag/AgCl electrode configuration sampled at 1,000 Hz (nasion reference, AFz ground; electrode impedance <10 kΩ). The experimental paradigm used a visual row-column "speller" design augmented with face stimuli to evoke P300 event-related potentials. Each trial had a duration of 1 second, and the dataset contained approximately 6,900 non-target and 1,380 target trials per subject across two sessions.

Preprocessing

EEG data were pre-processed in MATLAB (The MathWorks, Inc. 2024) using the EEGLAB toolbox (Delorme and Makeig 2004). Signals from all 61 electrodes were retained, downsampled to 256Hz, and bandpass filtered between 0.1-45Hz to remove slow drifts and high-frequency noise while preserving neurophysiologically relevant activity. Artifacts were corrected using Independent Component Analysis (ICA), with ocular and muscular components automatically identified through the ICLabel classifier. Components exceeding a 70% probability threshold for artifact classification were removed. The cleaned data were subsequently re-referenced to the average reference. Continuous recordings were segmented into non-overlapping 20-second epochs. Epochs containing excessive noise, defined as a standard deviation (STD) greater than 50 μV in any channel, were discarded, resulting in the exclusion of 18.6% of epochs. The remaining data were retained for downstream analysis.

Models

We employed EEGNetv4, a compact convolutional neural network specifically designed for EEG decoding tasks. EEGNetv4 integrates depthwise and separable convolutions to efficiently learn both temporal and spatial filters, thereby reducing the number of trainable parameters while preserving discriminative capacity. The architecture begins with a temporal convolution layer, which acts as a frequency filter to capture oscillatory patterns in EEG signals. This is

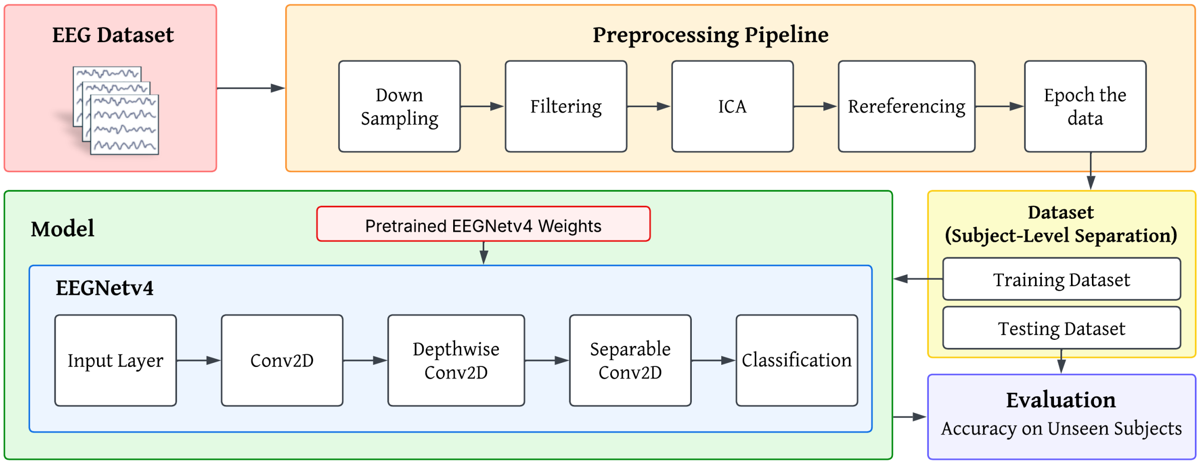

Figure 1: Overview of the training pipeline.

followed by a depthwise convolution layer that learns spatial filters across multiple channels, mimicking traditional spatial filtering techniques (e.g., common spatial patterns). A separable convolution is subsequently applied to combine spatial and temporal representations, further enhancing computational efficiency. Batch normalization, dropout, and average pooling layers are interleaved throughout the network to improve generalization and mitigate overfitting. The network concludes with a fully connected classification layer producing softmax outputs for class probabilities. The architecture of the model is shown in Figure 1.

For comparative evaluation, we implemented two additional deep learning architectures. The LSTM model processed multichannel EEG segments as sequences, with each timestep represented by a 61-dimensional feature vector. A single bidirectional LSTM layer with 128 hidden units was employed to capture forward and backward temporal dependencies, followed by dropout regularization (0.3) and a fully connected output layer for binary classification. The Transformer model projected channel-wise inputs into a 64-dimensional embedding space with sinusoidal positional encodings and a prepended learnable [CLS] token for sequence-level representation. The architecture consisted of stacked multi-head self-attention encoder layers with feed-forward sublayers, residual connections, and normalization, and the [CLS] embedding was passed through a classification head to produce the final prediction.

Transfer Learning

For transfer learning, we initialized EEGNetv4 with weights pretrained on the ERP dataset. The pretrained weights were obtained from the Hugging Face model repository guido151/EEGNetv4 (guido151 2024). Given architectural differences between the source and target datasets, the input dimensionality, number of EEG channels, and length of the input window differed between the pretrained and target models. Specifically, the pretrained EEGNetv4 was trained on 19-channel ERP data with an input window of 128 time points per epoch, whereas our target dataset consisted of 61 EEG channels with 5120 time points per epoch. Consequently, only parameters whose dimensions matched between the pretrained and target models were transferred. Parameters in layers that depended on the input channel number or input length were randomly initialized. To preserve the learned temporal and spatial feature representations while adapting to the target dataset, we froze all layers for which pretrained weights were successfully loaded and trained the remaining randomly initialized layers. This strategy ensured that low-level convolutional filters, which capture generic EEG signal features, remained intact, while higher-level layers were fine-tuned to accommodate dataset-specific characteristics. The classification layer was trained from scratch, as the output dimensionality differed between source and target tasks.

Experimental Settings

Model training and evaluation were conducted within a five-fold subject-independent cross-validation framework. For each fold, approximately 80% of the subjects were allocated to the training set and the remaining 20% to the test set, thereby preserving strict subject-level separation and mitigating data leakage.

All models were optimized using the AdamW optimizer with a learning rate set to 1×10^{-4}. Model training employed the Binary Cross-Entropy with Logits loss function to handle the binary classification task, with updates performed on mini-batches of size 64. The models were implemented in Python and executed using the PyTorch deep learning library (Paszke et al. 2019). The models were trained on an NVIDIA RTX A6000 GPU.

Model performance in each fold was quantified by computing the classification accuracy on the held-out test set. Final results were reported as the mean and STD of accuracy across the five folds, providing a robust estimate of generalization performance under subject-level cross-validation.

Model	Accuracy (%)	STD (pp)
LSTM	61.70	3.20
Transformer	63.35	2.78
EEGNetv4	65.75	5.48
EEGNetv4 (Transfer Learning)	**70.79**	4.17

Table 1: Mean classification accuracy and standard deviation (STD) across five subject-independent folds. Accuracy values are reported in percentages, and STD values are reported as percentage points (pp).

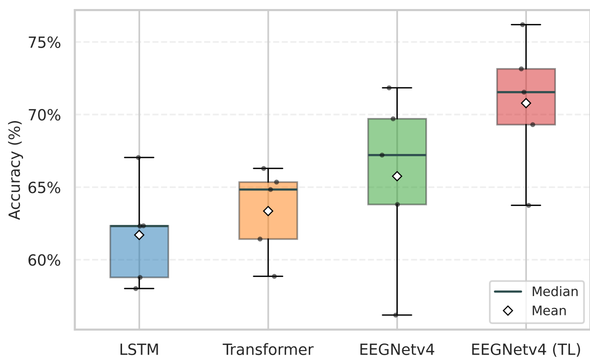

Figure 2: Model accuracy comparison under five-fold subject-independent cross-validation.

Results

Table 1 summarizes the classification performance of the evaluated models under five-fold subject-independent cross-validation. The LSTM and Transformer baselines achieved mean accuracies of 61.70% (STD = 3.20 percentage points (pp)) and 63.35% (STD = 2.78 pp), respectively. EEGNetv4 trained from scratch outperformed both baselines, yielding an accuracy of 65.75% (STD = 5.48 pp). The best performance was obtained with the transfer learning approach, where initializing EEGNetv4 with pretrained weights from the ERP dataset improved accuracy to 70.79% (STD = 4.17 pp). The five-fold accuracies along with their mean and median are shown in Figure 2.

These results demonstrate that transfer learning provided a clear advantage over both conventional sequence models and the same CNN trained from scratch. Notably, the performance gain relative to the non-transfer EEGNetv4 baseline indicates that pretrained temporal and spatial filters generalized effectively to the SD classification task, thereby enhancing model robustness and generalization across subjects.

Conclusion

This study investigated the utility of TL for the classification of SD versus well-rested wakefulness from resting-state EEG. Using EEGNetv4 as the backbone model, we demonstrated that initializing with pretrained weights significantly improved classification accuracy compared to training the same architecture from scratch. These findings highlight the potential of TL to mitigate the limitations of dataset size, a common constraint in neurophysiological research, by leveraging generic spatio-temporal feature representations learned from large external datasets. With TL, the model achieved enhanced generalization across subjects, supporting the broader application of pretrained EEG models to novel clinical and experimental contexts.

Future work will focus on extending this approach to larger and more diverse cohorts, incorporating multimodal data (e.g., behavioral and physiological markers), and evaluating advanced fine-tuning strategies such as layer-wise adaptation or domain-specific pretraining. Such directions hold promise for improving the sensitivity and reliability of EEG-based biomarkers for sleep deprivation, with potential translation into real-world cognitive monitoring and fatigue management systems.

References

Aristimunha, B.; Carrara, I.; Guetschel, P.; Sedlar, S.; Rodrigues, P.; Sosulski, J.; Narayanan, D.; Bjareholt, E.; Barthelemy, Q.; Schirrmeister, R. T.; Kobler, R.; Kalunga, E.; Darmet, L.; Gregoire, C.; Abdul Hussain, A.; Gatti, R.; Goncharenko, V.; Thielen, J.; Moreau, T.; Roy, Y.; Jayaram, V.; Barachant, A.; and Chevallier, S. 2025. Mother of all BCI Benchmarks.

Azab, A. M.; Toth, J. M.; Mihaylova, L. S.; and Arvaneh, M. S. 2018. *A review on transfer learning approaches in brain–computer interface*, chapter Chapter 5, 81–101. The Institution of Engineering and Technology.

Azunre, P. 2021. *Transfer learning for natural language processing*. Simon and Schuster.

Baygin, N. 2025. Melatonin Pattern: A new method for machine learning-based classification of sleep deprivation. *Diagnostics*, 15(3): 379.

Brodzicki, A.; Piekarski, M.; Kucharski, D.; Jaworek-Korjakowska, J.; and Gorgon, M. 2020. Transfer learning methods as a new approach in computer vision tasks with small datasets. *Foundations of Computing and Decision Sciences*, 45(3): 179–193.

Chambon, S.; Galtier, M. N.; Arnal, P. J.; Wainrib, G.; and Gramfort, A. 2018. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4): 758 – 769.

Craik, A.; He, Y.; and Contreras-Vidal, J. L. 2019. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3): 031001.

Delorme, A.; and Makeig, S. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1): 9–21.

Dong, B. T.; Lin, H. Y.; and Chang, C. C. 2022. Driver Fatigue and Distracted Driving Detection Using Random Forest and Convolutional Neural Network. *Applied Sciences*, 12(17).

guido151. 2024. EEGNetv4 Model Repository. https://huggingface.co/guido151/EEGNetv4. Accessed: 2025-08-21.

He, Z.; Tang, M.; Wang, P.; Du, L.; Chen, X.; Cheng, G.; and Fang, Z. 2023. Cross-scenario automatic sleep stage classification using transfer learning and single-channel EEG. *Biomedical Signal Processing and Control*, 81: 104501.

Klimesch, W. 1999. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2–3): 169–195.

Krause, A. J.; Ben Simon, E.; Mander, B. A.; Greer, S. M.; Saletin, J. M.; Goldstein-Piekarski, A. N.; and Walker, M. P. 2017. The sleep-deprived human brain. *Nature Reviews Neuroscience*, 18(7): 404–418.

Landrigan, C. P.; Rothschild, J. M.; Cronin, J. W.; Kaushal, R.; Burdick, E.; Katz, J. T.; Lilly, C. M.; Stone, P. H.; Lockley, S. W.; Bates, D. W.; and Czeisler, C. A. 2004. Effect of Reducing Interns' Work Hours on Serious Medical Errors in Intensive Care Units. *The New England Journal of Medicine*, 351(18): 1838–1848.

Lashgari, E.; Liang, D.; and Maoz, U. 2020. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, 346: 108885.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5): 056013.

Lee, M. H.; Kwon, O. Y.; Kim, Y. J.; Kim, H. K.; Lee, Y. E.; Williamson, J.; Fazli, S.; and Lee, S. W. 2019. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience*, 8(5): giz002.

Lim, J.; and Dinges, D. F. 2010. A Meta-Analysis of the Impact of Short-Term Sleep Deprivation on Cognitive Variables. *Psychological Bulletin*, 136(3): 375–389.

Makeig, S.; Debener, S.; Onton, J.; and Delorme, A. 2004. Mining event-related brain dynamics. *Trends in Cognitive Sciences*, 8(5): 204–210.

Mousavi, S.; Afghah, F.; and Acharya, U. R. 2019. SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLOS ONE*, 14(5): 1–15.

Olesen, A. N.; Jennum, P.; Mignot, E.; and Sorensen, H. B. D. 2020. Deep transfer learning for improving single-EEG arousal detection. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 99–103.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Red Hook, NY, USA: Curran Associates Inc.

Phan, H.; Chén, O. Y.; Koch, P.; Lu, Z.; McLoughlin, I.; Mertins, A.; and De Vos, M. 2021. Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning. *IEEE Transactions on Biomedical Engineering*, 68(6): 1787–1798.

Ren, Z.; Li, R.; Chen, B.; Zhang, H.; Ma, Y.; Wang, C.; Lin, Y.; and Zhang, Y. 2021. EEG-Based Driving Fatigue Detection Using a Two-Level Learning Hierarchy Radial Basis Function. *Frontiers in Neurorobotics*, 15: 618408.

Roy, Y.; Banville, H.; Albuquerque, I.; Gramfort, A.; Falk, T. H.; and Faubert, J. 2019. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5): 051001.

Schirrmeister, R. T.; Springenberg, J. T.; Fiederer, L. D. J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; and Ball, T. 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11): 5391–5420.

Supratak, A.; Dong, H.; Wu, C.; and Guo, Y. 2017. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11): 1998 – 2008.

The MathWorks, Inc. 2024. *MATLAB Release R2024b*. The MathWorks, Inc., Natick, MA.

Van Dongen, H. P. A.; Maislin, G.; Mullington, J. M.; and Dinges, D. F. 2003. The cumulative cost of additional wakefulness: dose–response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep*, 26(2): 117–126.

Wan Masri, W. N. N.; Zulkifli, N. Z. A.; Kamaruzzaman, M. A. A.; and Mohamad Zulkufli, N. L. 2024. EEG-based Sleep Deprivation Classification: A Performance Analysis of Channel Selection on Classifier Accuracy. *International Journal on Perceptive and Cognitive Computing*, 10(2): 67–73.

Wang, H.; Guo, H.; Zhang, K.; Gao, L.; and Zheng, J. 2022. Automatic sleep staging method of EEG signal based on transfer learning and fusion network. *Neurocomputing*, 488: 183–193.

Williamson, A.; Lombardi, D. A.; Folkard, S.; Stutts, J.; Courtney, T. K.; and Connor, J. L. 2011. The link between fatigue and safety. *Accident Analysis & Prevention*, 43(2): 498–515.

Xiang, C.; Fan, X.; Bai, D.; et al. 2024. A resting-state EEG dataset for sleep deprivation. *Scientific Data*, 11(427).

Zhang, X. Z.; Zheng, W. L.; and Lu, B. L. 2017. EEG-Based Sleep Quality Evaluation with Deep Transfer Learning. In *Neural Information Processing*, 543–552. Cham: Springer International Publishing. ISBN 978-3-319-70093-9.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2021. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, 109(1): 43–76.