npj | digital medicine

Editorial

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01086-9

Challenges and opportunities of deep learning for wearable-based objective sleep assessment

In recent years the intersection of wearable technologies and machine learning (ML) based deep learning (DL) approaches have highlighted their potential in sleep research. Yet, a recent study published in NPJ Digital Medicine highlights the generalization limitations of DL models in sleep-wake classification using actigraphy data. Here, this article discusses some of the challenges and opportunities presented by domain adaptation and self-supervised learning (SSL), innovative methodologies that use large-scale unlabeled data to bolster the generalizability of DL models in sleep assessment. These approaches not only improve sleep-wake classification but also hold promise for extending to more comprehensive sleep stage classification, potentially advancing the field of automated sleep assessment through efficient and user-friendly wearable monitoring systems.

Deep learning (DL, Table 1), a subset of machine learning (ML), has significantly impacted the field of automated sleep assessment, especially through the analysis of polysomnography (PSG) data. PSG is the most accurate objective sleep measurement method because it simultaneously assesses multiple physiological parameters, including overnight brain activity, and can classify sleep into distinct stages[1]. DL models trained on clinical PSG data have attained performance levels comparable to human experts, providing clinicians with valuable tools for automated and comprehensive sleep stage analysis[2–5], across a range of clinical datasets e.g., MESA[6], SHHS[7]. However, PSG's suitability for long-term, at-home sleep monitoring is limited due to its intrusive nature. Even headband devices like Dreem™, though less intrusive than traditional PSG technology for brain wave-based sensing, can be cumbersome/uncomfortable during extended wear[8].

Recent developments in wearable and nearable technologies have made it feasible to monitor sleep in home settings[3,9–11]. Despite advancements, the effectiveness of adopting wearable devices and DL methods for sleep analysis is often hindered by data scarcity, leading to model overfitting[12]. For instance, to estimate sleep parameters, a recent study published in NPJ Digital Medicine by Patterson et al.[13] evaluated DL models based on actigraphy data in cross-dataset settings and found that those models often struggle with considerable domain discrepancies[13], which poses challenges for effectively deploying DL models across varied settings and devices. Many wrist-worn devices now feature photoplethysmography (PPG) sensors, alongside actigraphy, indicating their potential for classifying sleep stages[3,4,14]. Nonetheless, many investigations have been conducted on small datasets, yielding limited performance outcomes. Conversely, fields such as natural language processing use abundantly availability datasets to aid the development of sophisticated DL models, such as ChatGPT[15]. That disparity highlights the potential benefits of using large volumes of unlabeled data to enhance sleep monitoring technologies.

Challenges: Wearable sensing and deep learning

The frequent implementation of DL in various fields is remarkable, yet it encounters two key challenges when applied to sleep assessment through wearable sensing-based methodologies. Namely, (i) small-labelled dataset problem (i.e., data scarcity), and (ii) the balancing act between achieving a high signal-to-noise ratio (SNR, a method that compares the level of a desired signal to the level of background noise) in wearables and maintaining user acceptance for long-term usage.

Data Scarcity: Annotation and patient availability

In sleep medicine, especially with wearable computing, the development of supervised learning models is impeded by a lack of richly annotated datasets. Obtaining unlabeled data from wrist-worn wearable devices is feasible and pragmatic. However, annotating those data for sleep classification requires simultaneous electroencephalography (EEG) collection and expert medical annotation. That contrasts with fields such as computer vision, where the annotation is more straightforward (i.e., requires less expertise), underscoring the unique difficulties in assembling annotated sleep-based datasets for supervised DL wearable-based algorithms[16,17].

Furthermore, limited research resources, patient scarcity, and the challenge of recruiting a diverse patient population with varying disease severities exacerbate data imbalances, making models easily overfit to the training dataset, affecting generalizability on unseen populations (i.e., participants were out of the distribution/heterogeneity of the training dataset). That phenomenon is evidenced in the evaluation outcomes presented by Patterson et al., demonstrating that when the training and test datasets originate from the same distribution, the performance of the DL model surpasses that of conventional methods. Assessments based on the proxy signals, such as those from cardiorespiratory signals, reveal distinct patterns in individuals with conditions like sleep apnea[18], underscoring the need for more diverse data to improve model generalizability.

Signal to noise ratio: Adequate hardware

The quest for high SNR wearables persists, capable of precisely gauging brain activity with minimal intrusion and optimal comfort[19]. Approaches based on wrist movement and

Table 1 | Terminology and descriptors used in this editorial

Terminology	Description
Actigraphy	Actigraphy uses a non-invasive wearable device to track rest and activity through movement.
Deep learning (DL)	Deep learning is a subset of machine learning (ML) that uses neural networks with many layers to analyze complex patterns in large amounts of data.
Domain adaptation	Domain adaptation is a technique in ML that aims to improve model performance on a target domain by leveraging knowledge from a related but different source domain.
Model overfitting	Model overfitting in ML occurs when a model fits too closely to the training dataset and cannot generalize to new/unseen data
Nearables	A type of smart object that can enhance the interaction with e.g., people and other smart objects. One notable example is a smartphone that can improve the usability and experience of wearing a smart watch.
Photoplethysmography (PPG)	PPG measures capillary blood volume changes by detecting light variations, used for heart rate monitoring.
Polysomnography (PSG)	PSG is the most accurate objective sleep measurement method because it simultaneously assesses multiple physiological parameters, including overnight brain activity, and can classify sleep into distinct stages
Self-supervised learning (SSL)	Self-supervised learning trains models on tasks using the data itself to generate supervisory signals for training on a task without relying on human-provided labels
Signal-to-noise-ratio (SNR)	Signal-to-noise-ratio quantifies the clarity of a signal in a system by comparing its power to that of the background noise, with a higher SNR indicating a clearer signal.

cardiac sensing data may reach a ceiling effect, as peripheral signals might not precisely reflect sleep stages[20]. Traditional scalp and forehead skin-based sensing methods are less perturbed by physiological activities other than the brain[18,21–25]. The advancements made using DL models with PSG data for automated sleep staging analysis highlighted the significant potential of soft textile-based EEG sleep detection devices, such as MUSE™[9,21]. The trade-off between usability and performance remains crucial in developing wearables aimed at sleep stage classification[19]. Moreover, the persistent data scarcity issue remains challenging, necessitating exploration into ML paradigms like self-supervised learning (SSL) and transfer learning as potential avenues to bolster model generalization and adaptability to new tasks.

Opportunities: Self-supervised machine learning and domain adaptation

In automated sleep analysis, SSL is combined with domain adaptation to become a key strategy for enhancing model generalization[26,27]. Domain adaptation refines models developed in one domain of sleep research (e.g., laboratory sleep patterns) to be applicable in another (e.g., free-living conditions sleep patterns). It overcomes disparities in data volume or quality by discarding irrelevant features and capturing universally recognized patterns, making it a valuable tool for advancing sleep assessment methodologies with limited data. SSL represents a paradigm shift in automated sleep analysis, enabling models to learn from large volumes of unlabelled data through the identification of inherent patterns. This approach is analogous to inferential learning in humans, where understanding is developed through observation rather than explicit instruction (e.g., learning the differences between sleep epochs and similar sleep epochs at different times). By employing *pretext* tasks, such as predicting the next sequence in a series of data points, SSL models can learn general features and patterns relevant to sleep, contributing to the robustness and accuracy of downstream supervised learning tasks classification[28,29].

The great promise of SSL has been observed across a range of domains in computer vision[30], natural language processing[31–33], and speech processing[34]. In automated sleep analysis, with the widespread proliferation of miniature sleep sensing technologies, accumulating substantial quantities of unlabeled data has become increasingly feasible. This development holds the potential to furnish extensive datasets for the training of SSL models, which are frequently structured around an encoder-decoder architecture. The encoders transform raw data into a compact representation, and decoders reconstruct the original data from this representation to learn meaningful patterns without explicit labels. What does that mean? Consider a *pre-train-then-fine-tune* paradigm, the encoder is initially trained to acquire useful representations (features) for downstream sleep-related tasks, such as sleep stage classification and sleep spindle recognition. Subsequently, those learned encoders are frozen, and trained/fine-tuned task-specific classification layers are updated to categorize specific events of interest within a smaller expert-annotated dataset. That approach aims to capture fundamental signal characteristics by learning to discern high-level semantics (e.g., different patterns in sleep data indicate sleep stages, quality, or disturbances) to facilitate effective representation learning. Of further interest is the use of SSL with domain adaptation in integrating those techniques with existing frameworks, potentially enhancing the adaptability and effectiveness of sleep stage classification algorithms across varied data sources and environments.

Harnessing existing SSL approaches

Various existing framework methodologies like SimCLR[35], MoCo[36], SimSiam[37], and Barlow Twins[38], offer universally adaptable frameworks that could seamlessly extend into sleep monitoring, warranting investigations into their efficacy. For instance, a recent study using accelerometer data alone from over 96,000 UK Biobank participants has shown the effectiveness of SSL for three-stage sleep classification and achieved an F1 score of 0.573 ± 0.12, representing a 7.1% improvement over the baseline model that did not incorporate SSL pre-training, as validated through internal evaluations[39]. This outcome challenges previous assumptions regarding the feasibility of sleep stage classification using accelerometer data only. That method, crucial in a domain with limited labelled data, emphasizes the effectiveness of general representations learned through SSL for sleep stage classification.

In conclusion, the study by Patterson et al. highlighted the vulnerability of basic DL models to overfitting, particularly when applied to specific datasets, data preprocessing methodologies, and PSG annotation styles as demonstrated through a single cross-dataset evaluation. The effort to accumulate large-scale datasets of sleep stages, annotated by experts from raw data gathered through wearable devices continues to present a significant challenge. Nonetheless, DL has shown considerable promise in a single dataset setting. Hence, using vast amounts of

unlabeled raw data from wearables and exploring sophisticated model architectures to improve generalizability, like integrating SSL and domain adaptation, offers a promising path for advancing long-term sleep assessment.

Bing Zhai[1], Greg J. Elder[2] & Alan Godfrey[1]
[1]Department of Computer and Information Sciences, Northumbria University, Newcastle, UK. [2]Northumbria Sleep Research, Department of Psychology, Northumbria University, Newcastle upon Tyne, UK.
e-mail: alan.godfrey@northumbria.ac.uk

Received: 12 February 2024; Accepted: 22 March
2024Published online: 04 April 2024

References

1. Murray, B. J. Subjective and objective assessment of hypersomno- lence. *Sleep Medicine Clinics* **15**, 167–176 (2020).
2. Phan, H., Andreotti, F., Cooray, N., Ch´en, O. Y. & De Vos, M. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehab. Eng.* **27**, 400–410 (2019).
3. Zhai, B., Perez-Pozuelo, I., Clifton, E. A. D., Palotti, J. & Guan, Y. Making sense of sleep: multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.* **4**, 1–33 (2020).
4. Fonseca, P. et al. Sleep stage classification with ecg and respi- ratory effort. *Physiol. Measur.* **36**, 2027 (2015).
5. Perslev, M. et al. U-sleep: resilient high-frequency sleep staging. *NPJ Digit. Med.* **4**, 72 (2021).
6. Xiaoli Chen, R. et al. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep* **38**, 877–888 (2015).
7. Quan, S. F. et al. The sleep heart health study: design, rationale, and methods. *Sleep* **20**, 1077–1085 (1997).
8. Graeber, J. et al. Technology acceptance of digital devices for home use: Qualitative results of a mixed methods study. *Digital Health* **9**, 20552076231181239 (2023).
9. Arnal, P. J. et al. The dreem headband com- pared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep* **43**, zsaa097 (2020).
10. Hsu, C.-Y. et al. Zero-effort in-home sleep and insomnia monitoring using radio signals. *Proceedings of the ACM on Interactive, Mobile, Wear- Able Ubiquitous Technol.* **1**, 1–18 (2017).
11. Yu, B. et al. Wifi-sleep: sleep stage monitoring using commodity wi-fi devices. *IEEE Internet Things J.* **8**, 13900–13913 (2021).
12. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (MIT Press, 2016).
13. Patterson, M. R. et al. 40 years of actigra- phy in sleep medicine and current state of the art algorithms. *NPJ Digital Med.* **6**, 51 (2023).
14. Zhai, B., Guan, Y., Catt, M. & Pl¨otz, T. Ubi-sleepnet: Advanced multimodal fusion techniques for three-stage sleep classification using ubiquitous sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technol.* **5**, 1–33 (2021).
15. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inform. Process. Syst.* **35**, 27730–27744 (2022).
16. Phan, H. & Mikkelsen, K. Automatic sleep staging of eeg signals: recent development, challenges, and future directions. *Physiol. Measur.* **43**, 04TR01 (2022).
17. Roy, Y. et al. Deep learning-based electroen- cephalography analysis: a systematic review. *J. Neural Eng.* **16**, 051001 (2019).
18. Tobaldini, E. et al. Heart rate variability in normal and pathological sleep. *Front. Physiol.* **4**, 1–11 (2013).
19. Perez-Pozuelo, I. et al. The future of sleep health: a data-driven revolution in sleep science and medicine. *NPJ Digit. Med.* **3**, 42 (2020).
20. Lujan, M. R., Perez-Pozuelo, I. & Grandner, M. A. Past, present, and future of multisensory wearable technology to monitor sleep and circadian rhythms. *Front. Dig. Health* **3**, 721919 (2021).
21. Kwon, S., Kim, H. & Yeo, W-H. Recent advances in wearable sensors and portable electronics for sleep monitoring. *Iscience*, 24 (2021).
22. Trinder, J. et al. Autonomic activity during human sleep as a function of time and sleep stage. *J. Sleep Res.* **10**, 253–264 (2001).
23. Vanoli, E. et al. Heart rate variability during specific sleep stages. *Circulation* **91**, 1918–1922 (1995).
24. Stein, P. K. & Pu, Y. Heart rate variability, sleep and sleep disorders. *Sleep Med. Rev.* **16**, 47–66 (2012).
25. Boudreau, P., Yeh, W. H., Dumont, G. A. & Boivin, D. B. Circadian variation of heart rate variability across sleep stages. *Sleep* **36**, 1919–1928 (2013).
26. Zhao, M., Yue, S., Katabi, D., Jaakkola, T. S. & Bianchi, M. T. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*, pages 4100–4109. (PMLR, 2017).
27. Heremans, E. R. M. et al. From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep stag- ing. *Journal of Neural Engineering* **19**, 036044 (2022).
28. Xiao, Q. et al. Self-supervised learning for sleep stage classification with predictive and discriminative contrastive coding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1290–1294 (IEEE, 2021).
29. Banville, H., Chehab, O., Hyv¨arinen, A., En- gemann, D.-A. & Gramfort, A. Uncovering the structure of clinical eeg signals with self-supervised learning. *J. Neural Eng.* **18**, 046020 (2021).
30. Gidaris, S., Singh, P. & Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* (2018).
31. Devlin, J., Chang, M-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
32. Radford, A. et al. Improving language understanding by generative pre-training. (2018).
33. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
34. Liu, X. et al. Self-supervised learning: generative or contrastive. *IEEE Trans. Knowledge Data Eng.* **35**, 857–876 (2021).
35. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607 (PMLR, 2020).
36. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Mo- mentum contrast for unsupervised visual representation learning. In *Pro- ceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738 (2020).
37. Chen, X. & He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758 (2021).
38. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & St´ephane, D. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320 (PMLR, 2021).
39. Hang Yuan, T. et al. Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. *medRxiv* (2023).

Author contributions

The first draft was written by B.Z., G.J.E. and A.G. provided critical revisions and approved the final draft.

Competing interests

A.G. is a Deputy Editor of npj Digital Medicine and played no role in the internal review or decision to publish this Editorial. The remaining authors declare no competing interests.