

Personalized Machine Learning for Sleep and Recovery Assessment Using Continuous Photoplethysmographic Monitoring

Keren Chopra

Department of Computer Science, George Mason University, Fairfax, VA, USA.
karan.chopra@gmu.edu

Vishal Lyons

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
vishal.lyons@uab.edu

Rainer C. Murray

School of Computing, Clemson University, Clemson, SC, USA.
rainermurray05@clemson.edu

Abstract

The increasing availability of wearable photoplethysmographic sensors has created new opportunities for longitudinal sleep and recovery assessment outside clinical sleep laboratories. However, translating raw optical pulse signals into reliable, individualized estimates of sleep architecture and recovery status remains a systems-level challenge. This paper examines personalized machine learning approaches for sleep and recovery assessment using continuous photoplethysmographic monitoring. Rather than focusing narrowly on model accuracy, the discussion integrates sensor data acquisition, signal quality management, personalization architectures, computational infrastructure, governance, fairness, robustness, and sustainability. The analysis emphasizes that photoplethysmography-derived sleep assessment is not simply a signal processing problem but an infrastructure and policy problem involving heterogeneous devices, noisy ambulatory environments, individual physiological variability, and evolving regulatory expectations. Personalized models must adapt to inter-individual differences in cardiovascular dynamics, sleep physiology, and activity patterns while preserving privacy and maintaining clinical credibility. The paper reviews structural trade-offs between centralized, edge-based, and federated learning architectures, considers the role of domain adaptation and incremental personalization, and evaluates the implications of data drift, algorithmic bias, and long-term system maintenance. A central argument is that sustainable deployment requires balancing model expressiveness with interpretability, local responsiveness with global knowledge transfer, and innovation with transparent governance. The conclusion outlines future directions for resilient, fair, and clinically meaningful personalized sleep and recovery systems based on continuous photoplethysmography.

Keywords

personalized machine learning; sleep assessment; recovery; photoplethysmography; wearable systems; federated learning; health governance; signal quality.

1. Introduction

Sleep is a fundamental physiological process that supports metabolic regulation, immune function, cognitive performance, and emotional resilience. Public health research has increasingly identified inadequate sleep as a population-level risk factor with broad social and economic consequences [1]. At the same time, recovery from physical and cognitive demands is not adequately captured by isolated subjective ratings or occasional clinical assessments. Continuous physiological monitoring offers the possibility of observing recovery trajectories over time, but this promise depends on robust sensing, modeling, and deployment infrastructures. Photoplethysmography has become one of the most widely available modalities for ambulatory cardiovascular monitoring because it can be embedded in wrist-worn devices, finger sensors, and earbuds without requiring complex electrode placement [2]. Wearable photoplethysmography can now provide long-duration pulse waveforms, enabling estimation of heart rate, heart rate variability, oxygen saturation, and derived sleep-related parameters in everyday settings [3]. These capabilities create a foundation for personalized sleep and recovery assessment, yet significant system-level questions remain unresolved.

The complexity of sleep and recovery assessment arises partly from the interaction between physiological signals, environmental noise, behavioral context, and individual differences. Sleep disruption has well-documented effects on glucose metabolism, appetite regulation, and cardiovascular function, which suggests that recovery cannot be represented by a single scalar score [4]. A meaningful assessment system must integrate multiple temporal scales, ranging from beat-to-beat cardiovascular dynamics to nightly sleep architecture and multi-day recovery trends. Personalized machine learning offers a pathway toward such integration by adapting models to individual baselines, lifestyle factors, and health conditions. However, personalization also introduces challenges related to data scarcity, label noise, model overfitting, and the need for continuous updating. This paper approaches these challenges from a systems perspective, emphasizing architectural decisions, infrastructure requirements, governance mechanisms, and long-term sustainability rather than treating personalized machine learning primarily as a methodological exercise.

2. Background and Related Work

Conventional sleep assessment has historically relied on polysomnography in controlled laboratory settings, which provides rich electrophysiological information but is costly, intrusive, and limited in ecological validity. Actigraphy has been widely used as a more practical ambulatory alternative, using accelerometry to estimate sleep and wake patterns based on movement [5]. Although actigraphy has proven useful in large-scale studies, its movement-based assumptions can misclassify quiet wakefulness and sleep in individuals with fragmented sleep or low motor activity. Photoplethysmography adds cardiovascular information that is complementary to movement, allowing the estimation of autonomic changes associated with sleep stages and recovery. Validation studies have demonstrated that photoplethysmography-based sleep staging can approximate some aspects of polysomnography, particularly when combined with motion data and advanced machine learning models [6]. Deep learning models applied to heart rate variability features have further improved the capacity to infer sleep stages from ambulatory cardiovascular signals [7].

Existing machine learning research on sleep staging has produced a wide range of model architectures, including convolutional networks, recurrent networks, and hybrid attention-based systems [8]. The clinical reference standard for sleep stage scoring remains the framework established by the American Academy of Sleep Medicine, which labels discrete stages such as wake, rapid eye movement sleep, and non-rapid eye movement sleep [9].

Translating this categorical framework into continuous, personalized recovery assessment introduces conceptual tensions. A deep learning model trained on large population datasets may achieve impressive average performance while failing to capture clinically meaningful individual differences. High-performance medicine has emphasized the importance of tailoring artificial intelligence systems to patient-specific contexts rather than relying solely on aggregate accuracy [10]. Wearable photoplethysmography is particularly promising for personalized health monitoring because it captures continuous autonomic fluctuations that reflect stress, recovery, and circadian regulation. Nevertheless, moving from population-level proof-of-concept models to reliable personalized systems requires careful attention to data quality, concept drift, fairness, and deployment architecture.

3. System Architecture for Continuous Photoplethysmographic Monitoring

A deployable personalized sleep and recovery assessment system spans multiple physical and computational layers. At the sensing layer, photoplethysmographic sensors acquire optical pulse signals at rates ranging from tens to hundreds of hertz, depending on device capabilities and energy constraints. These raw waveforms are often processed on-device to reduce storage and communication burdens, but aggressive on-device compression can remove subtle waveform features that are informative for sleep and recovery estimation. The edge layer may perform initial artifact detection, motion filtering, and feature extraction, while higher-level personalization can reside in a cloud environment or in a distributed learning framework. Federated learning has been proposed as a way to train models across many wearable devices without centralizing raw physiological data [11]. This architecture is appealing from a privacy perspective, but it introduces additional complexity in model aggregation, communication efficiency, and data heterogeneity.

The choice between centralized, federated, and fully on-device personalization involves fundamental trade-offs. Centralized architectures benefit from abundant computational resources and easier model governance but raise privacy concerns because sensitive physiological data must be transmitted and stored. Fully on-device personalization offers strong privacy and low latency, yet it is constrained by memory, battery, and the limited availability of labeled sleep data on individual devices. Federated approaches occupy a middle ground by sharing model updates rather than raw data, but they require mechanisms to handle non-identically distributed data across users and devices. Responsible machine learning frameworks also demand that these architectural decisions incorporate auditability, consent, and the ability to explain model outputs to users and clinicians [12]. The system architecture must therefore be designed not only for predictive performance but also for accountability, energy efficiency, and resilience under real-world operating conditions.

4. Personalized Modeling Paradigms and Data Representation

Personalization is necessary because photoplethysmographic waveform morphology, heart rate variability, and sleep-related autonomic patterns vary considerably across individuals. Population-level models may perform inconsistently when deployed on heterogeneous user groups, particularly among individuals with cardiovascular conditions, medication effects, or atypical sleep schedules. Studies of wrist-worn monitoring in clinical settings have shown that device accuracy can vary across patient subgroups, highlighting the importance of individualized calibration and context-aware interpretation [13]. Personalization can be achieved through several conceptual strategies, including model fine-tuning on user-specific data, domain adaptation from population-level representations, and the use of user embeddings that condition model predictions on individual characteristics. The choice of

personalization strategy influences data requirements, update frequency, and the risk of overfitting to short-term physiological fluctuations.

Data representation is equally important. Recovery and sleep assessment models may operate on raw waveform segments, handcrafted time-domain and frequency-domain features, or learned representations extracted by deep neural networks. Raw waveform approaches preserve the most information but require robust artifact handling and large computational capacity. Feature-based approaches are often more interpretable and computationally efficient, but they may discard subtle waveform signatures relevant to sleep depth or autonomic recovery. Learned representations occupy an intermediate position, offering flexibility while requiring careful regularization to avoid spurious associations. The system-level implication is that data representation choices interact with deployment constraints, because a model that performs well in a high-performance cloud environment may not be suitable for continuous edge deployment with limited memory and intermittent connectivity.

5. Signal Quality, Artifact Rejection, and Preprocessing Infrastructure

Continuous photoplethysmographic monitoring in ambulatory contexts is challenged by motion artifacts, sensor displacement, ambient light changes, and variations in tissue perfusion. These disturbances can corrupt the pulse waveform and produce misleading cardiovascular features. Robust preprocessing is therefore a prerequisite for credible sleep and recovery assessment. Adaptive notch-filtration techniques have been proposed to recover photoplethysmographic signals during physical activity, illustrating the importance of dynamic artifact rejection rather than static filtering [14]. This type of artifact management must be integrated into the broader preprocessing infrastructure because physical activity and sleep are often separated in time but interconnected in recovery modeling. The system needs to distinguish between periods of usable signal, periods requiring reconstruction, and periods that should be excluded from analysis to avoid propagating noise into personalized models.

Signal quality assessment should be treated as a continuous and contextual process rather than a binary classification task. The system must account for sensor type, wearing location, skin temperature, ambient light, and user movement. Preprocessing pipelines that perform well during nighttime sleep may be insufficient for daytime recovery monitoring, where motion and activity levels are higher. A layered quality management architecture can include on-device signal quality indices, server-side quality audits, and model-level uncertainty estimation. By evaluating signal quality before feature extraction and again before model prediction, the system can reduce the risk of overconfident inferences from corrupted data. This infrastructure also supports transparent reporting, allowing users and clinicians to understand which portions of the recording contributed to the final sleep and recovery assessment.

6. Sleep and Recovery Estimation from Personalized Models

Sleep and recovery estimation can be formulated at multiple levels of abstraction. At the nocturnal level, the system may predict sleep stage sequences, total sleep time, sleep efficiency, and the timing of sleep onset and offset. At the recovery level, the system may estimate autonomic restoration, readiness for physical activity, or cognitive alertness based on trends in heart rate variability and cardiovascular response. Personalization is especially valuable at the recovery level because recovery baselines differ widely. Federated learning methods offer one route to learning shared representations while preserving individual adaptations [15]. However, personalized models must also be capable of handling

nonstationary data, such as changes in fitness, illness, travel, or medication. Incremental updating strategies can adjust model parameters gradually as new user data arrive, but they require safeguards against catastrophic forgetting and against overreacting to transient physiological events.

Gradient-boosted tree models and deep learning models each offer distinct advantages in this context. Tree-based models can be efficient for structured features and provide interpretable feature importance, which is useful for clinical communication [16]. Deep learning models can capture complex temporal dependencies and learn representations directly from signals but often require larger datasets and more careful regularization [17]. A pragmatic systems design may combine multiple model families within a layered architecture, using deep representations for waveform interpretation and interpretable models for high-level recovery scoring. The choice of model class should be guided not only by predictive accuracy but also by computational cost, update frequency, and the need for transparent decision support.

7. Infrastructure and Deployment Considerations

Deploying personalized sleep and recovery assessment systems at scale requires a resilient data infrastructure that can handle heterogeneous devices, irregular sampling, missing data, and firmware variability. Cloud-based model training may be appropriate for developing initial representations, but continuous personalization often requires edge or hybrid strategies. The system must balance the benefits of large-scale data aggregation against the risks of privacy erosion and centralized control. Methodological research in medical machine learning has shown that high internal validation accuracy does not guarantee robust external performance, particularly when training data are not representative of deployment populations [18]. This lesson applies directly to sleep and recovery models, which may be trained on healthy volunteers but deployed among users with sleep disorders, cardiovascular disease, or shift work schedules.

A sustainable deployment pipeline should include model registry systems, version control for personalization artifacts, monitoring for data drift, and automated evaluation on held-out user cohorts. The pipeline should also support graceful degradation when connectivity is unavailable or when sensor quality deteriorates. For example, if the photoplethysmographic signal is unreliable during a period of intense motion, the system might rely on accelerometry and prior user history to provide a conservative recovery estimate while flagging reduced confidence. Such context-aware fallback behavior is essential for real-world trust. A system that produces highly confident but incorrect recovery scores during periods of poor signal quality can undermine user adherence and clinical credibility.

8. Governance, Privacy, and Ethical Implications

Governance structures for personalized sleep and recovery assessment must address the sensitivity of longitudinal physiological data. Continuous photoplethysmographic monitoring can reveal information about sleep patterns, stress, cardiovascular function, and health conditions that extend beyond the stated purpose of recovery assessment. The collection of these data raises questions about consent, secondary use, data retention, and user autonomy. Federated architectures reduce some privacy risks by keeping raw signals on local devices, but they do not eliminate the possibility of model inversion or membership inference attacks. Responsible governance requires data minimization, purpose limitation, and transparent communication with users about how their data contribute to personalization and model improvement [12].

Fairness is another dimension of governance. Machine learning models can encode biases that arise from unrepresentative training data, uneven device performance, or differential access to wearable technology. Biases in sleep and recovery assessment may disproportionately affect populations whose physiological patterns differ from those of the training majority. A fairness-oriented governance framework should require subgroup performance evaluation, bias audits, and the inclusion of diverse populations in model validation. Regulatory oversight must also evolve to address adaptive models that change continuously after deployment. Unlike static medical devices, personalized models may update their parameters with each new night of sleep, raising questions about how to validate a system that is never truly in a fixed state.

9. Fairness, Robustness, and Clinical Validation

Fairness in personalized sleep and recovery assessment cannot be reduced to demographic parity, because physiological measurements are not affected by social categories in a simple or uniform way. Instead, fairness must be understood as consistent performance across clinically meaningful subgroups, including age, sex, cardiovascular status, medication use, and shift work patterns. Existing surveys of bias in machine learning emphasize that fairness definitions must be selected with attention to context and harm [19]. In sleep and recovery monitoring, an unfair model may systematically underestimate recovery in users with high resting heart rates or overestimate sleep efficiency in users with autonomic disorders. Such errors can have direct consequences for training recommendations, occupational health decisions, and clinical referrals.

Robustness requires more than high average accuracy on retrospective test sets. The system must be evaluated under realistic conditions that include motion artifacts, sensor rotation, daylight exposure, and changes in wearing behavior. It must also be robust to distribution shift across seasons, time zones, and health states. Clinical validation for personalized models should include prospective studies that compare photoplethysmography-derived estimates with established reference standards in relevant populations. Validation protocols should report performance by subgroup, signal quality stratum, and degree of personalization. This level of transparency supports regulatory review and helps clinicians interpret model outputs appropriately. Recent discussions of artificial intelligence in health and medicine reinforce the need for rigorous prospective evaluation before broad clinical deployment [20].

10. Sustainability and Long-Term Maintenance

Sustainable deployment of personalized sleep and recovery systems requires attention to model maintenance, energy consumption, and organizational capacity. Continuous photoplethysmographic monitoring generates large volumes of data over months and years, and personalization models must be updated as devices change, operating systems evolve, and user populations shift. A model that is trained once and never updated will degrade over time due to sensor drift, changes in data pipelines, and shifts in user behavior. The maintenance burden can be reduced through automated monitoring, periodic retraining, and modular system design that isolates personalization components from core signal processing modules. Energy consumption is another sustainability concern, because continuous sensing and on-device computation can shorten battery life and reduce user adherence. The design should therefore match computational effort to informational value, reserving intensive processing for periods of high recovery relevance.

Organizational sustainability depends on clear ownership of system components, defined processes for model updates, and mechanisms for responding to adverse events. A distributed system involving device manufacturers, cloud providers, healthcare institutions, and users requires governance structures that assign responsibility without creating excessive administrative burden. The long-term success of personalized machine learning for sleep and recovery assessment depends not only on algorithmic innovation but also on building durable institutional structures that can maintain data quality, monitor model performance, and adapt to new evidence over time.

11. Future Directions

Future research should explore more effective ways to combine photoplethysmography with other ambient and wearable signals, such as accelerometry, skin temperature, and environmental light exposure. Multimodal fusion may improve the robustness of sleep and recovery estimates, but it also increases system complexity and privacy sensitivity. Another important direction is the development of uncertainty-aware personalization methods that can express confidence in each recovery score based on signal quality, model familiarity, and the stability of recent physiological patterns. Such methods would support more responsible use in settings where decisions have real consequences for health and performance. Further work is also needed on evaluation frameworks that capture long-term recovery trajectories rather than isolated nightly predictions. The field would benefit from open benchmarks that include diverse populations, realistic artifact profiles, and transparent reporting of subgroup performance. Finally, governance research should examine how adaptive personalized health models can be regulated without stifling innovation, particularly when models are updated continuously in distributed environments.

12. Conclusion

Personalized machine learning for sleep and recovery assessment using continuous photoplethysmographic monitoring represents a promising intersection of wearable sensing, physiological modeling, and health infrastructure. The potential to observe individual recovery trajectories outside clinical settings is substantial, but the realization of this potential depends on resolving systemic challenges related to signal quality, personalization, deployment architecture, fairness, and governance. This paper has argued that sleep and recovery assessment should be treated as a socio-technical system rather than a standalone algorithmic task. Architectural choices about where computation occurs, how models are updated, and how data are governed shape the reliability and acceptability of the resulting system. Signal quality infrastructure must be robust and context aware, while personalization strategies must balance individual adaptation with generalization and stability. Fairness and clinical validation require attention to subgroup performance and prospective evaluation under realistic conditions. By adopting a systems-level perspective, researchers and practitioners can develop personalized photoplethysmography-based sleep and recovery assessment systems that are not only technically advanced but also trustworthy, sustainable, and aligned with broader health policy goals.

References

1. Barnes, C. M., & Drake, C. L. (2015). Prioritizing sleep health: Public health policy recommendations. *Perspectives on Psychological Science*, 10(6), 733–737.
<https://doi.org/10.1177/1745691615598509>

2. Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3), R1–R39. <https://doi.org/10.1088/0967-3334/28/3/R01>
3. Charlton, P. H., Kyriacou, P. A., Mant, J., Marozas, V., Chowienczyk, P., & Alastruey, J. (2020). Wearable photoplethysmography for cardiovascular monitoring. *Proceedings of the IEEE*, 108(2), 245–261. <https://doi.org/10.1109/JPROC.2019.2955842>
4. Depner, C. M., Stothard, E. R., & Wright, K. P. (2014). Metabolic consequences of sleep and circadian disruption. *Current Diabetes Reports*, 14(7), 507. <https://doi.org/10.1007/s11892-014-0507-z>
5. Sadeh, A. (2011). The role and validity of actigraphy in sleep medicine: An update. *Sleep Medicine Reviews*, 15(4), 259–267. <https://doi.org/10.1016/j.smrv.2010.10.001>
6. Fonseca, P., Weysen, T., Goelema, M. S., Møst, E. I. S., Radha, M., Lunsingh Scheurleer, R., van der Heijden, L., Raymann, R. J. E. M., & Aarts, R. M. (2017). Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep*, 40(7), zsx097. <https://doi.org/10.1093/sleep/zsx097>
7. Radha, M., Fonseca, P., Moreau, A., Ross, M., Cerny, A., Anderer, P., Long, X., & Aarts, R. M. (2019). Sleep stage classification from heart-rate variability using long short-term memory neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(6), 2467–2476. <https://doi.org/10.1109/JBHI.2018.2881425>
8. Faust, O., Razaghi, H., Barika, R., Ciaccio, E. J., & Acharya, U. R. (2019). A review of automated sleep stage scoring based on machine learning techniques. *Computers in Biology and Medicine*, 102, 288–299. <https://doi.org/10.1016/j.combiomed.2018.08.005>
9. Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C. L., & Vaughn, B. V. (2012). The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications, version 2.0. American Academy of Sleep Medicine.
10. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
11. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119. <https://doi.org/10.1038/s41746-020-00323-1>
12. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340. <https://doi.org/10.1038/s41591-019-0548-6>
13. Kroll, R. R., Boyd, J. G., & Maslove, D. M. (2020). Accuracy of a wrist-worn wearable device for monitoring heart rates in hospital inpatients: A prospective observational study. *Journal of Medical Internet Research*, 22(2), e16274. <https://doi.org/10.2196/16274>

14. Zheng, X., Dwyer, V. M., Barrett, L. A., Derakhshani, M., & Hu, S. (2022). Adaptive notch-filtration to effectively recover photoplethysmographic signals during physical activity. *Biomedical Signal Processing and Control*, 72, 103303.
15. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50–60. <https://doi.org/10.1109/MSP.2020.2975749>
16. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
17. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
18. Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine*, 5, 48. <https://doi.org/10.1038/s41746-022-00592-y>
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>
20. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38. <https://doi.org/10.1038/s41591-021-01614-0>