

Uncertainty-Aware Deep Learning for Reliable Cardiovascular Monitoring Under Physical Activity

Pavel Bergman

Department of Computer Science, University of Houston, Houston, TX, USA.

pavel.bergman@uh.edu

Rishi Varma

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.

contactrishi@ku.edu

Abstract

Deep learning has transformed cardiovascular monitoring by enabling automated interpretation of electrocardiogram and photoplethysmogram signals in ambulatory settings. However, physical activity introduces substantial nonstationary motion artifacts, electrode displacement, hemodynamic variation, and environmental interference that challenge conventional deterministic models. This paper examines uncertainty-aware deep learning as a systems-level response to the reliability gap between laboratory performance and real-world cardiovascular monitoring. It discusses architectural strategies that integrate predictive uncertainty estimation with multimodal wearable sensing, motion context, and physiological prior knowledge. Particular attention is given to the structural trade-offs among Bayesian approximations, deep ensembles, calibration layers, and attention-based fusion mechanisms. The paper situates these methods within broader infrastructure requirements including data provenance, annotation quality, device heterogeneity, edge-cloud orchestration, and continuous deployment under distribution shift. It further analyzes robustness, fairness, and governance implications when uncertainty-aware systems are translated into clinical and consumer workflows. Rather than treating uncertainty estimation as a narrow model-level add-on, the paper argues that reliability must be designed across the entire monitoring ecosystem, from sensor acquisition and preprocessing to clinical decision support. Policy frameworks, audit mechanisms, and lifecycle sustainability are discussed as necessary conditions for trustworthy deployment. The analysis highlights that uncertainty awareness can reduce false alarms, support abstention in degraded signal conditions, and enable safer human oversight, but only if institutional and technical safeguards align.

Keywords

cardiovascular monitoring; deep learning; uncertainty quantification; physical activity; wearable sensors; reliability; fairness; clinical governance.

1. Introduction

Recent advances in deep learning have shown that automated interpretation of ambulatory electrocardiogram signals can approach specialist-level performance in controlled datasets [1]. Simultaneously, wrist-worn photoplethysmography has become a scalable sensing modality for continuous cardiovascular surveillance because it does not require adhesive electrodes or chest straps [2]. However, the transition from controlled benchmarks to free-living physical activity exposes a persistent reliability gap. Physical motion perturbs optical coupling, alters

tissue perfusion, and introduces nonstationary noise whose statistical properties differ from training distributions [3]. The central thesis of this paper is that deterministic deep learning architectures, no matter how accurate in nominal testing, remain insufficient for high-stakes cardiovascular monitoring unless they explicitly represent uncertainty.

Uncertainty-aware deep learning changes the system design problem from point prediction to risk-aware inference. In this framing, the model not only estimates heart rate, arrhythmia probability, or oxygen saturation but also quantifies confidence, detects distributional novelty, and signals when sensor data are too degraded for reliable interpretation. This paper provides a system-level analysis rather than a narrow model-centric evaluation. It examines how uncertainty estimation interacts with multimodal sensor fusion, edge-cloud orchestration, annotation pipelines, regulatory audit, fairness, and long-term sustainability. The objective is to articulate a research and deployment agenda in which reliability is treated as an emergent property of the entire monitoring ecosystem, not as an isolated performance metric.

2. Problem Context and Clinical Need

Cardiovascular disease remains a leading source of morbidity and mortality, and ambulatory monitoring is central to early detection of arrhythmia, decompensation, and hypertension-related end-organ stress. Wearable biosensors now support continuous acquisition of electrocardiogram, photoplethysmogram, accelerometry, and temperature signals outside the clinic [4]. This expansion offers the opportunity to detect transient abnormalities that would escape periodic hospital-based assessment. Yet the physiological value of these data is frequently obscured by motion artifacts during walking, exercise, occupational tasks, and sleep transitions. The result is a high false-positive rate that contributes to alarm fatigue and limits clinical adoption.

The technical challenge is not only signal noise but also physiological coupling between activity and cardiovascular state. Physical exertion legitimately changes heart rate, stroke volume, peripheral resistance, and respiratory dynamics. A reliable monitoring systemA reliable monitoring system must therefore separate expected physiological responses to exertion from pathological deviations, while also recognizing when the signal quality is too poor to support any confident interpretation. In practice, a patient climbing stairs may produce heart rate acceleration that is entirely benign, but an algorithm trained primarily on resting data may classify this as tachycardia or artifact. Conversely, a subtle ST-segment change during moderate exertion may carry diagnostic importance but can be masked by low-frequency motion. This dual challenge of physiological context and signal degradation demands more than improved signal processing; it requires inference systems that can reason about their own confidence as conditions change.

The clinical consequences of unreliable monitoring are not distributed equally across settings. In hospital telemetry, false alarms already contribute to alarm fatigue, with nurses and physicians learning to ignore frequent nonactionable warnings. In ambulatory and consumer devices, overalerting creates anxiety, unnecessary emergency visits, and discontinuation of device use, while underalerting may miss dangerous events. Existing regulatory frameworks often assume that a cleared algorithm will maintain performance in similar populations and conditions. Yet free-living physical activity violates this assumption because input distributions shift continuously with posture, motion intensity, skin temperature, and sensor pressure. A robust system must therefore embed distributional awareness at every stage, from signal acquisition to decision presentation.

Reliability in cardiovascular monitoring is not a property of a single model but of a layered sociotechnical infrastructure. At the sensor layer, photoplethysmography is vulnerable to changes in capillary blood volume during muscle contraction, whereas electrocardiogram electrodes suffer from impedance variation and motion-induced baseline wander. At the feature layer, heart rate estimation algorithms may produce plausible but incorrect readings when signal quality is low, and downstream arrhythmia classifiers may inherit these errors without knowing their uncertainty. At the decision layer, clinicians and users need not only predictions but also calibrated probabilities, abstention flags, and explanations that distinguish artifact from physiology. These layers must be designed together, because optimizing any one component independently can produce brittle pipelines.

Uncertainty-aware deep learning addresses a critical missing link in this layered architecture. Instead of treating every input as equally interpretable, the system learns to produce predictive distributions whose spread reflects data quality and model limitations. Two distinct forms of uncertainty are relevant. Aleatoric uncertainty captures irreducible noise in the input, such as optical corruption from movement. Epistemic uncertainty captures model ignorance, such as encountering a rare arrhythmia pattern underrepresented in training. Recent work in computer vision and natural language processing has formalized these uncertainty types and demonstrated their practical value in selective prediction, active learning, and out-of-distribution detection [5]. The translation of these concepts to wearable cardiovascular monitoring, however, requires careful attention to physiological priors and sensor failure modes.

3. Deterministic Deep Learning and Its Limitations

Convolutional and recurrent neural networks have become standard building blocks for classifying arrhythmias from single-lead electrocardiograms and for estimating heart rate from photoplethysmographic signals. These models typically optimize categorical cross-entropy or regression loss on curated datasets. They output point estimates or softmax probabilities that are often interpreted as confidence, even though deep networks are known to be poorly calibrated and overconfident on shifted inputs [8]. The problem is particularly acute in wearable monitoring because the training data frequently excludes heavy exercise, yet deployment occurs precisely when users are physically active. A network trained on clean waveforms may assign high posterior probability to an artifact pattern if that pattern shares low-level features with a known arrhythmia.

Several infrastructure-related constraints amplify this limitation. First, ground truth labels for ambulatory electrocardiogram are expensive and often derived from retrospective clinician review, which itself is subject to interobserver variability. Second, public benchmark datasets such as PhysioNet and MIMIC-III have enabled methodological advances but reflect hospital-based acquisition and may not represent free-living optical noise [11], [12]. Third, the preprocessing pipelines used during development may include aggressive filtering or artifact rejection that is unavailable on low-power wearable processors. This concern is magnified in medical contexts, where the need for uncertainty quantification in machine-assisted decision making has been explicitly articulated [9]. Consequently, models encounter a wider range of inputs in deployment than during validation, and deterministic architectures provide no systematic mechanism to flag these distribution shifts.

A further concern is that deterministic models encourage a false separation between signal quality estimation and diagnosis. In many deployed systems, a separate signal quality index is computed before the deep learning classifier, often using hand-crafted features such as

kurtosis or template correlation. If the quality index is poorly calibrated, the classifier may still be applied to low-quality segments. Uncertainty-aware models offer an alternative by jointly learning from raw or minimally processed signals and producing confidence estimates that reflect both input quality and prediction ambiguity. This joint formulation reduces the risk of cascading failures and allows the system to defer ambiguous cases to longer acquisitions or human review. Nevertheless, achieving this behavior requires architectural and training choices that explicitly account for uncertainty, as discussed next.

4. Uncertainty-Aware Architectural Strategies

The design space for uncertainty-aware deep learning includes several families of methods with distinct computational and statistical trade-offs. Bayesian neural networks place distributions over network weights and marginalize over these distributions during inference, yielding principled epistemic uncertainty estimates [6]. Dropout-based approximations use randomly masked units at test time to simulate multiple plausible models, offering a lightweight alternative that can be added to existing networks [6]. Deep ensembles train multiple independent models with different initializations and combine their predictions, producing empirically strong uncertainty estimates at the cost of multiple forward passes and increased memory [7]. Each approach has different implications for wearable deployment. Bayesian inference is theoretically appealing but can be difficult to scale to large multimodal architectures and to implement on constrained edge hardware. Ensembles offer robustness and simplicity but increase energy consumption and may exceed the memory budgets of low-power microcontrollers.

An emerging direction is to separate aleatoric and epistemic uncertainty within a single architecture, allowing the system to distinguish noisy but learnable inputs from genuinely novel conditions [5]. In cardiovascular monitoring, this distinction matters because motion artifact during exercise is often aleatoric, whereas a rare conduction disorder or an unseen sensor hardware revision is more likely epistemic. Aleatoric uncertainty can be learned by predicting the parameters of a probability distribution over the target, such as a heteroscedastic Gaussian for heart rate regression. Epistemic uncertainty can be approximated through ensembles, dropout, or more recent deterministic forms such as distance-aware classifiers. The choice of method should not be driven solely by benchmark accuracy but by the operational requirements of the monitoring application, including inference latency, battery life, update frequency, and the need for calibrated probabilities.

Calibration is a separate but tightly related concern. A model may produce uncertainty estimates that are internally consistent yet miscalibrated relative to observed event frequencies. Temperature scaling and other post-hoc calibration methods can improve expected calibration error, but they require a representative held-out set and can degrade under distribution shift [8]. For wearable monitoring, calibration must be maintained over time as sensor hardware changes, user populations diversify, and clinical protocols evolve. Uncertainty-aware systems should therefore be evaluated not only on accuracy but also on selective prediction curves, which measure how performance improves as low-confidence cases are abstained. Such evaluation aligns with medical decision-making, where deferring a small fraction of ambiguous cases can substantially increase safety.

A systems-level implication is that uncertainty estimation should not be a post-hoc add-on but part of the training objective and deployment pipeline. For example, a network could be trained to minimize a loss that includes both a task-specific term and a calibration or sharpness term, encouraging predictions that are confident when correct and diffuse when

unreliable. Regularization can also penalize overconfidence on synthetically corrupted data, helping the model recognize motion artifacts seen only during training augmentation. These techniques have been shown to improve reliability under dataset shift in general domains [10], and their adaptation to physiological signals is an active research area. The next section examines how multimodal fusion and motion context can strengthen uncertainty estimates and reduce false alarms during physical activity.

5. Multimodal Fusion and Motion Context

A single photoplethysmographic or electrocardiographic channel provides limited information about whether a detected change is physiological or artifactual. Accelerometer and gyroscope signals from the same wearable can reveal the intensity and type of motion, which are strongly predictive of signal corruption. When a user is running, sudden heart rate spikes are more likely to be noise than when the user is resting. Multimodal fusion can exploit this relationship by conditioning cardiovascular inference on motion context. Attention mechanisms have been proposed to let the network focus on relevant parts of the physiological signal while suppressing motion-corrupted segments [16]. Such approaches can improve robustness during physical activity by learning which time-frequency regions remain trustworthy under different activity states.

However, multimodal fusion introduces structural trade-offs. Motion signals themselves can be noisy, sensor-specific, and affected by wearing position. A thigh-mounted accelerometer may not capture wrist movements during cycling, while a wrist accelerometer may record large accelerations during typing or cooking that do not compromise photoplethysmography. The fusion architecture must learn when motion context is informative and when it should be ignored. Cross-attention between physiological and motion streams can be computationally expensive, and low-power edge devices may not support complex attention modules. Hierarchical fusion, in which features are combined at multiple scales, can offer a compromise between accuracy and efficiency, but it requires careful co-design with hardware accelerators and memory constraints.

Uncertainty estimation benefits from multimodal fusion because it provides additional evidence for abstention. If the physiological signal is noisy and the motion context indicates vigorous exercise, the model can assign high aleatoric uncertainty and suppress alarms. If the motion context is calm but the physiological signal remains anomalous, the model may elevate epistemic uncertainty and recommend a longer acquisition or user confirmation. This synergy is not automatic; it requires training data that include synchronized physiological and motion streams across diverse activities, ages, skin tones, and cardiovascular conditions. Public datasets have begun to include such multidomain recordings, but significant gaps remain for underrepresented populations and extreme activity levels [13]. The required referencing underscores the relevance of joint attention learning for opto-physiological monitoring during physical activity [16], indicating that architectural innovations can directly target motion-induced uncertainty.

Another consideration is the temporal structure of cardiovascular signals. Heart rate variability, arrhythmia onset, and recovery from exercise unfold over seconds to minutes. Uncertainty-aware models should condition on temporal context rather than making independent predictions for each short window. Recurrent networks, temporal convolutions, and transformer-based architectures can capture these dependencies, but they also complicate uncertainty estimation because errors propagate across time steps. Modeling uncertainty at the sequence level may require predicting a distribution over latent states or using ensembles of

temporal models. This is distinct from static image classification and demands careful evaluation on longitudinal wearable data. A monitoring system that correctly identifies high uncertainty during a transient motion artifact but then fails to recover confidence quickly may be perceived as unreliable by users.

6. Data, Infrastructure, and Edge-Cloud Governance

Reliable uncertainty-aware monitoring depends on the quality and provenance of training data as much as on model architecture. Data collection for ambulatory cardiovascular monitoring is expensive, privacy-sensitive, and subject to selection bias. Monitoring studies often recruit participants who are healthy, younger, or less burdened by chronic illness, whereas the intended clinical population includes older adults, people with darker skin tones, and individuals with implanted devices. Photoplethysmography is known to be affected by melanin concentration and tissue composition, and a model trained primarily on lighter skin tones may exhibit systematic uncertainty miscalibration on darker skin tones [14]. This is both a technical and an equity challenge. Data audit mechanisms should record device type, sensor placement, skin tone, activity context, and annotation procedure so that uncertainty performance can be stratified across subgroups.

Federated learning offers a governance pathway that can improve diversity without centralizing sensitive physiological records. In a federated system, models are trained across local devices and only aggregated updates are shared, reducing privacy risks and enabling data from geographically distributed populations [18]. However, federated training introduces additional sources of uncertainty because local data distributions are heterogeneous and communication constraints limit model synchronization. Personalized uncertainty calibration can help tailor models to individual users while preserving global knowledge, but it must be implemented without embedding biases or creating separate performance tracks. The governance question is not only whether federated learning is technically feasible but whether institutions can audit model updates, detect malicious participation, and ensure that underrepresented groups benefit from the shared model.

Edge-cloud orchestration further shapes uncertainty-aware design. Wearable devices often have limited processing power, memory, and battery capacity, making on-device inference attractive for latency and privacy but restrictive for large ensembles or Bayesian networks. Cloud-based inference allows more complex models but introduces dependency on connectivity and raises concerns about data transmission. A hybrid architecture can allocate lightweight uncertainty screening on device and dispatch ambiguous cases to the cloud for deeper analysis. This split requires consistent communication protocols, fallback behavior during outages, and careful management of version skew between edge models and cloud services. Model updates must be delivered without disrupting real-time monitoring, and uncertainty calibration must be revalidated after every update. Infrastructure decisions are therefore inseparable from the epistemic properties of the deployed model.

Sustainability also deserves attention. Training large multimodal uncertainty-aware systems consumes significant energy, and frequent retraining on continuously growing ambulatory data amplifies this cost. At the same time, false alarms and unnecessary clinical visits generate environmental and economic burdens. A system that reduces false positives can be sustainable in clinical practice even if its training cost is higher. Lifecycle assessment should include data storage, sensor production, network transmission, and model update frequency. Uncertainty estimation can support sustainability by enabling active learning, in which only informative or uncertain cases are transmitted for human annotation or cloud processing. This

reduces bandwidth and labeling costs while maintaining reliability. Governance frameworks should encourage such efficiency gains without sacrificing auditability.

7. Robustness, Fairness, and Policy Implications

Robustness to distribution shift is a central motivation for uncertainty-aware monitoring. Ambient temperature changes, sensor replacement, user medication changes, and seasonal activity patterns can all shift input statistics. A model that produces calibrated uncertainty can detect these shifts and request recalibration or abstain more frequently until performance is restored. However, robustness must be evaluated under realistic perturbations, including adversarial noise, sensor dropout, and systematic biases such as pulse oximetry inaccuracies across skin tones. Existing deep learning models have been shown to fail silently on out-of-distribution data, producing confident but incorrect predictions [10]. Uncertainty-aware systems are not immune to this failure; they merely provide a signal that can be used to mitigate harm. If the uncertainty signal is ignored by downstream decision-makers or if it is miscalibrated due to distribution shift, the system remains unsafe.

Fairness analysis must move beyond aggregate accuracy to examine uncertainty patterns across demographic and clinical subgroups. If a model is systematically more uncertain for older adults or people with darker skin tones, then a policy of abstaining on high-uncertainty cases may lead to differential denial of automated monitoring and increase health disparities. It is essential to distinguish between uncertainty that reflects genuine physiological ambiguity and uncertainty that reflects underrepresentation or measurement bias [19]. Bias audits should include subgroup calibration, abstention rates, and downstream outcomes, not only overall sensitivity and specificity. Regulatory submissions should require evidence that uncertainty-aware safety mechanisms do not disproportionately burden protected groups. This aligns with broader ethical frameworks for responsible machine learning in health care [15].

Policy implications extend to approval pathways and postmarket surveillance. Regulatory agencies traditionally evaluate fixed algorithms under predetermined protocols. Uncertainty-aware systems challenge this model because their behavior changes with input confidence and may include abstention or cloud escalation. Regulators need new standards for evaluating calibration, abstention thresholds, and version updates in adaptive or continuously learning systems. Postmarket surveillance should collect real-world uncertainty logs to detect degradation and bias after deployment. Transparency mechanisms such as model cards and datasheets can document intended use, training data characteristics, and known limitations [20]. Without such documentation, clinicians cannot interpret uncertainty outputs meaningfully, and patients cannot contest decisions that affect their care.

Human factors are equally important. Clinicians and lay users may misinterpret probabilistic outputs as absolute risk or may distrust systems that frequently abstain. Interface design should convey uncertainty in ways that support decision-making without inducing anxiety or automation bias. For example, a smartwatch could display a low-confidence heart rate with a visual qualifier and recommend repositioning the sensor or reducing motion, rather than simply hiding the reading. In clinical dashboards, uncertainty could be represented as confidence intervals or alert priority levels, with clear explanations of why a measurement is flagged. The design of such interfaces must be informed by user studies and clinical workflow analysis, not solely by model performance metrics [21]. Uncertainty-aware deep learning can improve reliability only if the humans in the loop understand and act appropriately on the uncertainty information.

Finally, the academic community must resist the tendency to report only best-case results on cleaned datasets. Real-world evaluation of uncertainty-aware cardiovascular monitoring should include prospective studies, longitudinal deployment, and heterogeneous populations. Reporting standards should emphasize calibration error, selective prediction, abstention rates, and fairness metrics alongside traditional accuracy. Open benchmarks and challenge datasets can accelerate progress, but they must be curated with attention to motion, device diversity, and annotation ambiguity. The required reference illustrates a methodological direction in which joint attention learning improves opto-physiological monitoring during physical activity [16], yet such innovations need to be embedded in broader evaluation and governance ecosystems to achieve reliable deployment. Explainability methods can also help users understand why uncertainty was triggered, but they must be evaluated against clinical usefulness rather than visual plausibility [22]. Edge inference acceleration and lightweight model design will further determine whether uncertainty mechanisms can be deployed on consumer wearables at scale [23]. Uncertainty estimates can also be leveraged to predict segmentation and detection quality in adjacent medical imaging tasks, suggesting cross-domain benefits [24].

8. Conclusion

Uncertainty-aware deep learning represents a necessary but insufficient condition for reliable cardiovascular monitoring under physical activity. The transition from deterministic point estimation to predictive uncertainty modeling addresses the fundamental distribution shift introduced by motion, sensor variability, and physiological heterogeneity. However, the reliability of such systems depends on a web of structural choices, including multimodal sensor fusion, edge-cloud orchestration, federated data governance, subgroup calibration, and human-centered interface design. This paper has argued that uncertainty estimation must be treated as a systems property rather than a model-level feature.

The architectural space offers viable options, from Bayesian approximations and deep ensembles to attention-based multimodal fusion. Each involves trade-offs in computational cost, calibration stability, and interpretability. No single method dominates across all deployment contexts, and the selection must be informed by clinical workflow, energy budgets, and regulatory requirements. At the same time, advances in uncertainty modeling will yield limited benefit if training data remain unrepresentative or if postdeployment monitoring ignores distribution shift. The inclusion of motion context and joint attention mechanisms can help suppress artifacts, but these approaches require diverse and carefully annotated data to generalize across populations.

Governance, fairness, and policy are not external constraints but integral components of the monitoring system. Uncertainty outputs must be audited for subgroup disparities, communicated transparently, and integrated into clinical decision pathways in ways that enhance rather than undermine trust. Postmarket surveillance should continuously evaluate calibration and abstention behavior under real-world conditions. As wearable cardiovascular monitoring expands, the research community has an opportunity to build infrastructures that are both technically robust and socially accountable. Achieving this goal will require sustained interdisciplinary collaboration among signal processing experts, deep learning researchers, clinicians, regulators, and patient communities. The path forward is not merely to improve model accuracy but to design monitoring systems that know when they do not know and act accordingly.

References

1. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65–69.
2. Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3), R1–R39.
3. Charlton, P. H., Marozas, V., & Zou, L. (2019). Wearable photoplethysmography for cardiovascular monitoring. *Proceedings of the IEEE*, 107(5), 936–954.
4. Zhang, Z., Pi, Z., & Liu, B. (2015). TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on Biomedical Engineering*, 62(2), 522–531.
5. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 5574–5584.
6. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 1050–1059.
7. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 6402–6413.
8. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.
9. Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1), 20–23.
10. Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J., Ren, J., & Nado, Z. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 13991–14002.
11. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035, 160035.
12. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), e215–e220.
13. Reiss, A., Indlekofer, I., Schmidt, P., & Van Laerhoven, K. (2019). Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14), 3079.
14. Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, 383(25), 2477–2478.

15. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981–983.
16. Zheng, X., Hu, S., Dwyer, V., Barrett, L., & Derakhshani, M. (2026). Joint attention mechanism learning to facilitate opto-physiological monitoring during physical activity. *Biomedical Signal Processing and Control*, 113, 108949.
17. Islam, S. M., Balasubramanian, V. N., & Khan, S. (2020). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297.
18. 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.
19. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4, 123–144.
20. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
21. Yang, Q., Steinfeld, A., & Zimmerman, J. (2019). Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11.
22. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, 359–380.
23. David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., Regev, S., Rhodes, R., Wang, T., & Warden, P. (2021). TensorFlow Lite Micro: Embedded machine learning for TinyML systems. *Proceedings of Machine Learning and Systems*, 3, 800–811.
24. Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59, 101557.