

Explainable Artificial Intelligence for Activity-Aware Cardiovascular State Estimation from Wearable Signals

Rowan R. Gonzalez

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
rowan556@colostate.edu

Zhantian He

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
hellozhan@missouri.edu

Akshay R. Gandhi

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.
akshay2004@uab.edu

Abstract

Wearable sensing now enables continuous, non-invasive observation of cardiovascular dynamics through photoplethysmography, electrocardiography, and inertial measurement. However, translating these signals into clinically and behaviorally meaningful cardiovascular state estimates under free-living conditions remains challenging because physical activity introduces nonstationary artifacts, alters autonomic context, and shifts sensor coupling. This paper presents a systems-oriented analysis of explainable artificial intelligence for activity-aware cardiovascular state estimation from wearable signals. It examines how signal processing, representation learning, temporal modeling, and explanation layers can be integrated into an infrastructure that is simultaneously accurate, interpretable, and robust outside controlled environments. Rather than promoting a single model, the paper considers structural trade-offs among gradient-boosted ensembles, recurrent and attention-based deep networks, and interpretable parameterized models. It argues that explanation in wearable cardiovascular monitoring must address multiple audiences, including clinicians, patients, developers, and regulators, and must account for activity context as a first-class element rather than a nuisance variable. The discussion covers motion artifact recovery, adaptive filtering, feature attribution, causal and social dimensions of explanation, fairness across demographic and physiological subgroups, deployment governance, energy constraints, privacy, and long-term sustainability. Case illustrations and cross-domain comparisons highlight how explanation methods developed for computer vision or static tabular data require careful adaptation for streaming physiological time series. The paper concludes that explainability should be designed as a systemic property of the monitoring infrastructure, not appended as a post hoc visualization, and that activity-aware cardiovascular state estimation will benefit from tighter integration between sensing hardware, physiological priors, and transparent machine learning workflows.

Keywords

explainable artificial intelligence; wearable sensors; cardiovascular state estimation; photoplethysmography; physical activity; physiological time series; responsible deployment.

1. Introduction

The rapid expansion of wearable health technologies has transformed cardiovascular monitoring from episodic clinical measurements into a continuous, free-living observational practice. Wrist-worn and chest-worn devices now routinely acquire photoplethysmographic, electrocardiographic, accelerometric, and temperature signals at sampling rates that were previously achievable only in specialized laboratory environments. This transformation creates substantial opportunities for early detection of cardiac abnormalities, longitudinal tracking of autonomic regulation, and individualized health management. However, the physiological meaning of wearable signals is highly context-dependent. Heart rate, heart rate variability, pulse wave morphology, and derived cardiovascular indices change not only because of underlying cardiac state but also because of postural transitions, locomotion, sleep-wake state, emotional stress, thermal stress, and sensor displacement. An activity-aware cardiovascular state estimator must therefore disentangle physiological regulation from behavioral context while remaining transparent enough for clinical trust and regulatory acceptance.

The computational landscape for this task has shifted decisively toward data-driven modeling. Large-scale models such as gradient-boosted tree ensembles have become common for structured feature extraction and classification tasks in wearable analytics [1]. These models are attractive because they handle heterogeneous features, missingness, and nonlinear interactions without requiring exhaustive physiological modeling. At the same time, explainability methods have expanded from local linear approximations to additive feature attribution frameworks that assign importance values to model inputs [2]. Model-agnostic explanation systems have further enabled surrogate interpretability for black-box predictors [3]. In parallel, gradient-based saliency methods originally developed for image classification have been adapted to highlight influential regions of physiological time series [4]. These technical developments support the analysis of model behavior, but they do not automatically answer the clinical question of whether a model relies on physiologically plausible evidence. For wearable cardiovascular state estimation, this distinction is particularly important because motion-related waveform distortions can masquerade as pathological changes if the model is not explicitly activity-aware.

Deep learning has further altered the performance frontier. Cardiologist-level arrhythmia detection from ambulatory electrocardiographic recordings has been demonstrated with deep convolutional architectures [5], and semi-supervised sequence learning has been applied to cardiovascular risk prediction from wearable heart rate and step data [6]. These models can automatically learn hierarchical temporal representations that are difficult to specify manually. Multi-task learning offers another architectural advantage by sharing representations across related physiological prediction tasks, thereby improving sample efficiency and reducing overfitting in domains where labeled clinical events are rare [7]. Nevertheless, as model capacity increases, the relationship between learned representations and established cardiovascular physiology becomes less direct. This opacity creates barriers for clinical adoption, patient comprehension, and regulatory validation. The central challenge addressed in this paper is therefore not simply to improve predictive accuracy, but to design an integrated system in which activity context, physiological prior knowledge, and explainability are treated as co-equal requirements across the full deployment lifecycle.

2. Wearable Cardiovascular Sensing and Activity-Aware State Estimation

Wearable cardiovascular monitoring depends fundamentally on the quality and interpretability of the underlying physiological signals. Photoplethysmography is especially attractive because it can be acquired continuously from the wrist with low power consumption and without the need for adhesive electrodes. However, photoplethysmographic signals are highly susceptible to motion artifacts caused by arm swing, muscle contraction, changes in probe pressure, and ambient light variation. These artifacts often share frequency content with the physiological pulse waveform, making simple band-pass filtering insufficient. Activity awareness therefore begins at the signal conditioning stage. Adaptive notch filtration strategies have been developed to recover photoplethysmographic signals during physical activity [8]. By tracking motion-related spectral components and suppressing them in a time-varying manner, such methods can preserve pulse morphology while reducing distortion from locomotion. This is a representative case in which signal processing and machine learning must operate jointly; the downstream model cannot be expected to correct for artifacts that have already been folded into misleading feature values.

Several research systems have demonstrated the importance of combining spectral signal processing with temporal verification for wearable heart rate monitoring during intensive physical exercise [9]. These approaches decompose the raw signal into spectral components, identify likely pulse frequencies using accelerometer-derived motion references, and then verify the estimated heart rate trajectory over time. Subsequent work has refined this task by modeling the relationship between motion intensity and photoplethysmographic corruption more explicitly, leading to more accurate heart rate monitoring during exercise [10]. Reviews of wearable photoplethysmography sensors repeatedly emphasize that sensor positioning, skin tone, tissue perfusion, and ambient conditions introduce systematic variability that cannot be ignored in deployment [11]. The activity-aware problem is therefore not a narrow artifact removal problem. It involves a broader representation of context that may include exercise type, movement intensity, circadian phase, and environmental state.

From a systems perspective, the signal processing layer must satisfy multiple constraints simultaneously. It must operate within finite power and memory budgets, adapt to inter-individual differences in skin and vasculature, and preserve enough signal detail for downstream explainability. Overly aggressive filtering can remove physiologically meaningful waveform features such as systolic peak shape, dicrotic notch, and pulse transit characteristics. Under-filtering contaminates learned representations with motion-driven variation that may subsequently be misinterpreted as cardiovascular instability. The design trade-off is structural: a compact onboard signal processor can reduce data transmission and protect privacy, but it can also destroy information that would otherwise support retrospective audit and model explanation. A full activity-aware system should therefore treat signal processing decisions as an explainability-relevant component of the overall architecture, not merely as a pre-processing convenience.

3. Explainable Artificial Intelligence Approaches for Physiological Time Series

Explainability in wearable cardiovascular monitoring must account for the temporal, multivariate, and context-dependent nature of the data. Approaches imported from computer vision often produce saliency maps over input pixels or image regions, but cardiovascular signals require explanations that align with physiological time markers such as heartbeat onset, systolic peak, and respiratory modulation. Fairness considerations intersect with explainability because model explanations can reveal which subpopulations are being

represented poorly [12]. If explanations differ systematically across age groups, skin tones, or activity levels, the explanation layer itself becomes a diagnostic instrument for auditing model behavior. At the same time, explanation is not only a computational artifact. It is also a social interaction in which the explainer selects evidence, the recipient evaluates plausibility, and the broader context determines whether the explanation is actionable [13]. In clinical and consumer health environments, this social dimension is especially relevant because users interpret cardiovascular estimates through their own health beliefs, emotional state, and prior experiences.

Causability offers a further conceptual refinement by distinguishing between the technical output of an explainability algorithm and the human capacity to understand and act upon that output [14]. A feature attribution map may be technically correct but causally uninformative if it highlights a segment of the photoplethysmographic waveform that a clinician cannot relate to underlying cardiac physiology. Interpretable models can reduce this gap by expressing cardiovascular state estimation in terms of explicit physiological parameters such as heart rate, amplitude modulation, and motion-induced uncertainty. In high-stakes settings, there is a strong argument for using inherently interpretable models rather than relying solely on post hoc explanations of complex black-box systems [15]. However, building fully interpretable models for free-living cardiovascular monitoring is difficult because physiological dynamics often involve higher-order interactions that do not have straightforward parameterized forms. The practical compromise is likely a hybrid architecture in which interpretable physiological models provide a transparent baseline, while more complex learned components are constrained by domain knowledge and accompanied by targeted explanations when they deviate from the baseline.

The challenge of explaining temporal models is further compounded by the contextual nature of activity. A heart rate increase can be explained as a normal response to brisk walking or as a sign of cardiac stress, depending on activity context. An explainability system that simply highlights the accelerometer magnitude as an important feature has not fully explained the cardiovascular state. It must also articulate why a given combination of motion and heart rate response is considered normative or atypical. This requires moving beyond input attribution toward counterfactual reasoning and case-based comparison. For example, a useful explanation might indicate that the current heart rate trajectory is similar to previously observed episodes of moderate walking but unusual for the same individual at a lower activity intensity. Such explanations depend on longitudinal personal baselines, contextual similarity metrics, and uncertainty communication, all of which must be embedded in the system rather than treated as external commentary.

4. Architectural and Infrastructural Trade-Offs

The design of an activity-aware cardiovascular state estimation architecture must reconcile several competing objectives: predictive accuracy, interpretability, energy efficiency, privacy, latency, and adaptability. Modern temporal architectures such as attention-based transformers can model long-range dependencies and variable-length physiological sequences, but they often require substantial memory and specialized hardware [16]. Recurrent networks with long short-term memory cells remain a robust alternative for learning temporal dependencies in streaming physiological data [17]. In many wearable deployments, the computational constraints of edge devices favor hybrid architectures that use lightweight feature extraction on the sensor followed by more expressive modeling in the cloud. Deep learning more generally offers powerful representation learning capabilities, but these representations are

not automatically aligned with physiological categories or clinical vocabularies [18]. The architectural decision is therefore not solely about model family. It is about where abstraction occurs, how information is compressed, and whether the resulting representations preserve the semantic structure needed for explanation.

A central infrastructural tension arises between onboard inference and remote inference. Onboard inference reduces data transmission, improves privacy, and can lower latency for time-critical alerts. However, onboard models are constrained by battery life, memory, and computational throughput, which often forces the use of simplified networks or reduced sampling rates. Remote inference can support more sophisticated explanations, longitudinal personalization, and cohort-level comparisons, but it introduces connectivity dependencies, data governance risks, and potential delays. A well-designed system may use tiered inference, in which a compact onboard model performs continuous screening and a more capable cloud model performs detailed analysis when an anomaly or ambiguous state is detected. This tiered design also has implications for explanation: onboard explanations must be simple and immediate, while cloud-based explanations can be more detailed and contextual. The infrastructure must support both levels without creating a situation in which the deployed model and the audited model diverge.

Another trade-off concerns the temporal granularity of state estimation. Beat-to-beat analysis can reveal short-term autonomic dynamics such as respiratory sinus arrhythmia and postural responses, but it is computationally sensitive and vulnerable to artifact. Minute-level aggregates are more stable and energy-efficient, but they obscure transient events that may be clinically important. Activity-aware estimation should ideally operate at multiple timescales, using fast signal processing for immediate safety-relevant events and slower trend analysis for long-term cardiovascular health characterization. This multi-timescale structure has direct implications for explainability. Explanations at the beat level might refer to wave morphology and sensor motion, while explanations at the monthly trend level might refer to changes in resting heart rate, sleep quality, and physical activity patterns. A coherent infrastructure must maintain traceability across these timescales so that aggregate conclusions can be linked back to supporting signal evidence and activity context.

5. Robustness, Fairness, and Governance in Deployment

Robustness in wearable cardiovascular monitoring is challenged by distribution shift across devices, sensor revision cycles, population subgroups, and environmental conditions. A model trained on controlled treadmill protocols may degrade when applied to cycling, household chores, or sleep. Explainability can support robustness monitoring by revealing whether the model is relying on artifact-related features rather than stable physiological signals. If explanations are unstable under minor input perturbations, or if they change abruptly across sensor versions, this can serve as an early warning that the model has not learned a genuinely physiological mapping. Robustness is therefore not a separate property that can be verified only through aggregate accuracy. It must be assessed through the behavior of explanations, uncertainty estimates, and contextual validity checks.

Fairness is of particular concern in wearable cardiovascular monitoring because signal quality can vary systematically with skin tone, vascular properties, body composition, and activity patterns. Fairness audits in machine learning have shown that differential model performance across demographic groups can arise even when the training data appear balanced [12]. In cardiovascular sensing, photoplethysmographic signal attenuation is influenced by melanin concentration, tissue thickness, and peripheral perfusion, all of which can introduce structured

differences in signal-to-noise ratio. If these differences are not explicitly modeled, the system may allocate higher uncertainty or lower accuracy to certain groups. Explainable activity-aware systems can help detect such disparities by examining whether model explanations for equivalent cardiovascular states differ across groups in ways that are not physiologically justifiable. Fairness cannot be reduced to a single metric; it requires ongoing dialogue among model developers, clinicians, community representatives, and regulators.

Governance frameworks for machine learning in health care emphasize the need to align model behavior with clinical values, to document training data and failure modes, and to establish clear lines of responsibility [19]. International AI ethics guidelines similarly converge on principles such as transparency, accountability, privacy, and beneficence, although their operationalization differs across regions [20]. In wearable cardiovascular monitoring, governance must address not only the algorithm but also the sensor platform, the user interface, and the data management ecosystem. A responsible deployment will specify how explanations are presented, when they must be escalated to a clinician, and how users can contest or opt out of automated inferences. Regulatory pathways such as software as a medical device require evidence that the system performs safely under intended use conditions, including free-living activity. Explainability should be integrated into this evidence base as a mechanism for demonstrating that the system uses clinically relevant information, not incidental correlations.

6. Sustainability and Policy Implications

Sustainability in wearable cardiovascular monitoring encompasses environmental, operational, and social dimensions. Energy consumption is a direct operational constraint because frequent sensing, model inference, and wireless transmission deplete batteries and increase electronic waste over the lifetime of a device. Explainability methods themselves can be computationally expensive, especially when they require repeated perturbation or gradient calculation. A sustainable system must balance the carbon and energy footprint of model training against the marginal benefit of additional accuracy, and must avoid implementing explanation modules that exceed the energy budget of the edge device. One possible approach is to use sparse explanation queries that run only when the model is uncertain, when a clinical threshold is approached, or when a user explicitly requests an explanation. This selective deployment of explainability reduces overhead while preserving auditability where it matters most.

Long-term sustainability also depends on data governance and interoperability. Wearable cardiovascular datasets are often fragmented across proprietary platforms, making it difficult to compare model performance, reproduce findings, or conduct independent fairness audits. Policy mechanisms that encourage secure data sharing, federated learning, and standardized metadata can improve the external validity of activity-aware models without requiring raw physiological data to leave the device. Federated learning architectures can update model parameters across populations while keeping personal data decentralized, but they also require careful explanation infrastructure because the global model may behave differently on local subgroups. Interoperability standards for activity labels, signal quality annotations, and explanation artifacts would allow researchers and regulators to evaluate models consistently.

From a policy perspective, activity-aware cardiovascular state estimation occupies a complex space between consumer wellness and medical diagnosis. The same sensor stream can be used for fitness coaching, stress management, arrhythmia screening, or heart failure monitoring, and the appropriate regulatory posture depends on the intended use. Explainability can

support this differentiation by making the evidentiary basis of each claim explicit. For example, a wellness application might explain a heart rate reserve trend in terms of recent exercise load, while a medical screening tool must explain why a rhythm irregularity is flagged for clinician review. Policy frameworks should require that explanation methods are validated not only for accuracy but also for comprehension by the intended user population. This includes patients with varying health literacy, clinicians with time constraints, and regulators who need standardized documentation. Without such policy-level integration, explainability risks becoming a rhetorical feature rather than a substantive safeguard.

7. Conclusion

Explainable artificial intelligence for activity-aware cardiovascular state estimation from wearable signals is best understood as an infrastructural challenge rather than a model refinement problem. The system must integrate adaptive signal conditioning, contextual activity modeling, temporal representation learning, uncertainty quantification, and multi-audience explanation across constrained edge devices and more capable remote services. Motion artifacts must be addressed at the signal processing layer so that downstream models do not mistake motion-induced variability for cardiovascular change. Explanation methods must move beyond static feature attribution toward contextual, counterfactual, and physiologically grounded accounts that consider activity intensity, personal baseline, and longitudinal trend. Governance and fairness must be embedded from the beginning, because wearable sensing data reflect systematic differences in skin properties, perfusion, sensor fit, and movement patterns that can otherwise become hidden sources of inequity. The architectural trade-offs between onboard inference, cloud analysis, interpretable baselines, and deep learned representations should be treated as explicit design decisions with consequences for sustainability, privacy, and auditability. Future systems will require shared protocols for signal quality annotation, activity context labeling, explanation validation, and federated model updates. Only by treating explainability as a systemic property can activity-aware cardiovascular state estimation become trustworthy enough for both individual health insight and broader clinical translation.

References

1. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
2. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30, 4765–4774.
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
4. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
5. 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. <https://doi.org/10.1038/s41591-018-0268-3>
6. Ballinger, B., Hsieh, J., Singh, A., Sohoni, N., Wang, J., Tison, G. H., Marcus, G. M., Sanchez, J. M., Maguire, C., Olgin, J. E., & Pletcher, M. J. (2018). DeepHeart: Semi-supervised sequence learning for cardiovascular risk prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2079–2086. <https://doi.org/10.1609/aaai.v32i1.11891>
 7. Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12), 5586–5609. <https://doi.org/10.1109/TKDE.2021.3070203>
 8. 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.
 9. 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. <https://doi.org/10.1109/TBME.2014.2359372>
 10. Temko, A. (2017). Accurate heart rate monitoring during physical exercises using PPG. *IEEE Transactions on Biomedical Engineering*, 64(9), 2016–2024.
 11. Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., & Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics*, 4(4), 195–202.
 12. 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>
 13. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
 14. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312. <https://doi.org/10.1002/widm.1312>
 15. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
 16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30, 5998–6008.
 17. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
 18. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
 19. 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., Thadaneys-Israeli, 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>
20. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>