

Multimodal Deep Learning for Joint Estimation of Heart Rate, Respiratory Rate, and Activity States

Chetan M. Sinha

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
sinha2003@buffalo.edu

Mateo M. May

School of Computing, Clemson University, Clemson, SC, USA.
may.mateo95@clemson.edu

Yiran Jin

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
jiny@uc.edu

Abstract

Multimodal deep learning is transforming physiological monitoring by integrating signals from electrocardiography, photoplethysmography, accelerometry, and contextual sensing to jointly estimate heart rate, respiratory rate, and activity states. Single-sensor approaches remain vulnerable to motion artifacts, sensor displacement, and physiological confounding, particularly during ambulatory use. This paper presents a system-level analysis of multimodal deep learning architectures for joint physiological and behavioral inference. We examine early, late, and hybrid fusion strategies, recurrent and attention-based temporal modeling, and their trade-offs in accuracy, latency, energy consumption, and interpretability. The discussion extends beyond model architecture to data acquisition, preprocessing infrastructure, annotation governance, bias management, and deployment across wearable and clinical edge computing platforms. We analyze robustness under distribution shift, fairness across demographic and clinical subpopulations, and regulatory expectations for continuous monitoring systems. A comparative perspective across consumer wearables, hospital telemetry, and remote patient monitoring illustrates structural differences in reliability, privacy, and sustainment. The paper argues that joint estimation should be treated as a socio-technical infrastructure problem, not solely as a pattern recognition task, and offers governance-oriented design principles for developing equitable, robust, and operationally sustainable multimodal monitoring systems.

Keywords

multimodal deep learning; physiological monitoring; heart rate; respiratory rate; activity recognition; wearable systems; algorithmic fairness; edge deployment.

1. Introduction

The convergence of wearable sensing, edge computing, and deep learning has created new possibilities for continuous physiological monitoring outside clinical environments. Heart rate, respiratory rate, and activity state are tightly coupled physiological and behavioral variables that jointly reflect cardiovascular demand, pulmonary function, and physical effort. Although these quantities can be estimated separately using electrocardiography, photoplethysmography, accelerometry, or respiratory inductance plethysmography, their

clinical utility depends on reliable joint inference during movement and across heterogeneous users. Multimodal machine learning offers conceptual and technical tools for learning shared representations across heterogeneous sensor streams [1]. This paper examines multimodal deep learning for joint estimation of heart rate, respiratory rate, and activity states from the perspective of large-scale socio-technical systems rather than isolated algorithmic design.

Single-sensor photoplethysmography has been widely used for heart rate monitoring, but ambulatory measurements are corrupted by motion artifacts, changes in tissue perfusion, and sensor displacement. The TROIKA framework established an important baseline by combining signal decomposition, sparse signal recovery, and spectral tracking to reduce motion artifact in wrist-type photoplethysmography during intensive exercise [2]. Subsequent work demonstrated that learning-based approaches could improve heart rate estimation during physical exercise by modeling the relationship between corrupted photoplethysmography and accelerometer signals [3]. These advances motivate deeper integration of multiple physiological and motion channels for joint estimation.

The central argument of this paper is that joint estimation should be understood as an infrastructure-level challenge. It involves not only model architectures, but also data governance, annotation protocols, edge-cloud resource allocation, robustness under distribution shift, fairness across demographic and clinical populations, and regulatory compliance. We therefore examine structural trade-offs in multimodal fusion, temporal modeling, deployment, and sustainability. In doing so, we connect wearable sensing research with broader conversations about trustworthy artificial intelligence, algorithmic fairness, and resilient health data systems.

2. Background and Related Work

Deep learning approaches for physiological monitoring have evolved from hand-engineered signal processing pipelines to data-driven architectures. Large-scale convolutional neural networks have been applied to photoplethysmography for heart rate estimation across diverse recording conditions, showing that learned representations can reduce sensitivity to noise and varied sensor placement [4]. The CorNET framework further demonstrated that convolutional and recurrent layers can estimate heart rate and biometric identity from ambulatory photoplethysmography in resource-constrained settings [5]. These systems treat motion corruption as a learnable interaction between physiological signals and accelerometry rather than as a purely statistical artifact.

Respiratory rate estimation from electrocardiography and photoplethysmography has a long clinical history and remains challenging during activity. Respiratory modulation affects the amplitude and timing of cardiac-derived signals, but the interference between respiratory and cardiovascular rhythms is difficult to disentangle in ambulatory conditions [6]. Multimodal approaches can exploit thoracic movement, accelerometer-derived respiratory effort, and pulse oximetry features, but they require careful alignment and artifact handling.

Activity recognition has similarly moved toward multimodal deep learning. Wearable activity recognition systems often combine convolutional feature extractors with recurrent networks to capture local movement patterns and temporal dependencies [7]. These architectures provide a useful reference for joint physiological estimation because activity context can serve both as a conditioning variable for artifact suppression and as an outcome that shares latent structure with cardiac and respiratory dynamics.

3. System Architecture and Multimodal Fusion Design

A central design decision in multimodal deep learning is the stage at which heterogeneous sensor streams are combined. Early fusion concatenates raw or lightly processed signals before shared representation learning, late fusion aggregates modality-specific predictions, and hybrid fusion introduces cross-modal interactions at multiple levels. Attention mechanisms provide a more flexible form of hybrid fusion because they allow the network to weight channels and time steps according to contextual relevance [8]. For joint heart rate, respiratory rate, and activity state estimation, attention can suppress accelerometer segments that are not coupled to physiological changes while emphasizing photoplethysmography intervals with high signal quality.

Temporal modeling is equally important. Long short-term memory networks remain widely used because they can retain information over variable physiological cycles and movement sequences [9]. Convolutional layers operating on time-frequency representations provide efficient local feature extraction and have been foundational in deep learning across domains [10]. In a joint estimation system, a convolutional front end can extract spectral and morphological features from electrocardiography and photoplethysmography, while recurrent or attention layers model longer-range dependencies in respiration and activity. Such architectures must balance representational capacity against latency and energy consumption on wearable devices.

Training stability and optimization also shape deployment viability. Adaptive optimization methods such as Adam have become standard because they reduce sensitivity to learning rate selection and accelerate convergence on noisy physiological data [11]. Deep neural networks originally developed for electrocardiogram classification illustrate how large annotated datasets can support clinically meaningful performance for cardiac monitoring [12]. Extending these results to joint estimation requires multitask loss balancing, uncertainty weighting, and careful treatment of missing modalities.

Respiratory rate estimation from pulse oximetry has benefited from robust algorithms that infer respiratory modulation from photoplethysmographic morphology and beat-to-beat variation [13]. In multimodal systems, these respiratory features can be combined with activity context to reduce false fluctuations caused by movement. Multimodal activity and context recognition research has shown that combining inertial, physiological, and environmental signals improves robustness in mobile settings [14]. The same principle applies to joint physiological monitoring: activity state is both an output and a context signal that informs the interpretation of physiological channels.

4. Data Acquisition and Preprocessing Infrastructure

Data infrastructure determines the ceiling of model performance. Multimodal physiological datasets require synchronized acquisition of electrocardiography, photoplethysmography, accelerometry, and sometimes respiratory reference signals. Sampling rates, sensor placement, device form factor, and ambient conditions introduce systematic differences that complicate cross-dataset generalization. Preprocessing must address unit conversion, timestamp alignment, artifact detection, missing data, and dynamic range normalization. These operations are not neutral; choices about filtering and resampling can remove diagnostically relevant features or amplify bias in downstream learning [15]. The governance of preprocessing pipelines therefore deserves the same scrutiny as model architecture.

Because deep supervised learning depends on labels, annotation protocols for heart rate, respiratory rate, and activity states must be reproducible. Reference heart rate can be obtained

from electrocardiography, but respiratory reference signals are often invasive or obtrusive, and activity labels may rely on video annotation or self-report. Label noise and interrater variability propagate through the system and disproportionately affect rare activity states or physiological patterns. A system-level response includes curated benchmark datasets, transparent annotation manuals, and systematic audits of missingness across groups and environmental conditions.

Data provenance and representativeness are also critical. Wearable datasets often overrepresent younger, healthier, and more active populations, while clinical validation sets may include older adults with arrhythmias, reduced perfusion, or respiratory disease. If training distributions do not reflect deployment populations, accuracy metrics from laboratory studies may not transfer to home monitoring or hospital step-down units. Multimodal fusion can mitigate some distribution shift by exploiting redundant signals, but redundancy is not guaranteed when sensor failure correlates with clinical condition.

5. Model Governance, Fairness, and Robustness

Algorithmic fairness in physiological monitoring is a structural concern because measurement errors can differ systematically by skin tone, body mass, age, and disease status. Fairness cannot be reduced to equal accuracy across groups when physiological signals themselves differ in morphology and dynamic range. Bias mitigation requires a combination of representative data collection, subgroup performance audits, and model-level constraints that account for clinically meaningful uncertainty [16]. Governance frameworks for machine learning systems emphasize the need for documented accountability, transparency, and institutional oversight when algorithms inform health decisions [17].

Attention-based multimodal architectures can contribute to robustness by making the system more selective about which modalities and time segments to trust. A joint attention mechanism for opto-physiological monitoring during physical activity has been proposed to align physiological and motion channels under challenging conditions [18]. Such mechanisms are promising because they can learn dynamically from sensor quality and activity context, but they also introduce new governance needs. Attention weights are not inherently interpretable, and high attention on a corrupted signal segment may still produce an incorrect estimate if the training objective is misaligned with clinical risk.

Robustness must be tested under adversarial and natural distribution shifts. Sensor dropout, battery depletion, loose device fit, extreme temperatures, and skin pigmentation changes are common in field deployments. A system that performs well on curated test sets may fail in real-world settings where multiple modalities are absent or delayed. Multitask learning can improve robustness by sharing representations across heart rate, respiratory rate, and activity state, but it can also create negative transfer if one task dominates optimization. Model governance should therefore include stress testing, worst-case subgroup analysis, and continuous postmarket surveillance.

6. Deployment and Sustainability Considerations

Deployment of multimodal deep learning systems involves a spectrum from cloud-based inference to fully on-device execution. Cloud inference enables complex models and frequent updates but raises privacy, latency, and connectivity concerns. On-device inference improves privacy and latency but imposes constraints on memory, energy, and thermal dissipation. Hybrid architectures can split feature extraction and temporal modeling across wearable,

gateway, and cloud tiers. Such decisions are not purely technical; they affect battery life, user adherence, data minimization, and the ability to audit model behavior in clinical settings.

Explainability is particularly important when physiological estimates are used for triage or remote patient monitoring. Clinicians and users need to understand why a system reports elevated heart rate or unstable activity state, especially when sensor quality is poor. Explainability methods for healthcare artificial intelligence can support trust, but they must be integrated with clinical workflows rather than treated as post hoc visualizations [19]. For joint estimation, explanations should indicate which modalities contributed to a prediction, whether the estimate is reliable, and what missing data might have influenced the result.

Privacy and data protection regulations shape deployment architectures. In many jurisdictions, physiological data are special categories of personal data subject to stringent safeguards, risk assessments, and user rights [20]. International health bodies have encouraged the use of digital technologies for public health while emphasizing interoperability, equity, and governance [21]. Sustainable deployment therefore requires standardized data formats, secure device onboarding, consent management, and mechanisms for deleting or correcting personal data without disrupting model monitoring.

The clinical history of photoplethysmography demonstrates that physiological measurement technology must be validated across skin types, perfusion states, and pathological conditions before widespread adoption [22]. Multimodal systems inherit these validation obligations and add new ones related to sensor fusion. A failure in one modality may be compensated by another, but compensation must be visible and auditable. Without such visibility, users may overtrust a system that is silently operating near the boundary of its reliability envelope.

7. Policy, Ethics, and Future Directions

The policy landscape for multimodal physiological monitoring is fragmented across consumer wellness devices, medical devices, and research systems. Consumer devices are often marketed for wellness and are not subject to the same regulatory evidence requirements as medical devices, even when their outputs influence clinical decisions. This creates ambiguity for joint estimation systems that may be used across both contexts. Policy frameworks should distinguish between low-risk activity tracking and higher-risk physiological monitoring for arrhythmia, respiratory decompensation, or rehabilitation. Regulatory pathways should accommodate continuous learning systems while requiring version control, audit trails, and postmarket performance reporting.

Ethically, joint estimation raises questions about continuous surveillance, user autonomy, and the meaning of physiological normality. Heart rate and respiratory rate are not fixed values; they fluctuate with sleep, stress, medication, and environment. Algorithms that reduce this variability to a single risk score may pathologize normal physiological variation or obscure clinically significant changes. Multimodal systems should therefore be designed with uncertainty communication, contextual awareness, and user-facing explanations that respect the interpretive complexity of physiological data.

Future research directions include self-supervised pretraining on large unlabeled wearable datasets, federated learning across institutions to preserve privacy, and causal representation learning to separate physiological causes from measurement artifacts. Cross-domain lessons from industrial machine health monitoring and environmental sensing suggest that failure prediction, calibration, and maintenance scheduling are as important as model accuracy. In physiological monitoring, this translates to algorithms that predict their own reliability, flag

sensor degradation, and recommend maintenance or reacquisition. Such capabilities will be essential as joint estimation systems move from short-term validation studies to continuous use in homes, hospitals, and public health programs.

8. Conclusion

This paper has examined multimodal deep learning for joint estimation of heart rate, respiratory rate, and activity states as a system-level problem spanning architecture, data infrastructure, governance, deployment, and policy. We argued that fusion strategies and attention mechanisms can improve robustness, but only when embedded within transparent data pipelines, representative datasets, and accountable deployment structures. The structural trade-offs between accuracy, latency, energy, interpretability, and fairness require interdisciplinary design and continuous evaluation. As wearable sensors become more capable and physiological monitoring extends beyond hospitals, the central challenge will not be merely learning better representations, but building resilient and equitable systems that maintain trust under real-world variability. Future advances in self-supervised learning, federated inference, and causal modeling must be accompanied by governance mechanisms that protect individuals while enabling the health benefits of continuous multimodal monitoring.

References

1. Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>
2. 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>
3. Temko, A. (2017). Accurate heart rate monitoring during physical exercises using PPG. *IEEE Transactions on Biomedical Engineering*, 64(9), 2016–2024. <https://doi.org/10.1109/TBME.2017.2676243>
4. 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. <https://doi.org/10.3390/s19143079>
5. Biswas, D., Everson, L., Liu, M., Panwar, M., Verhoef, B. E., Patki, S., Kim, C. H., Acharyya, A., Van Hoof, C., Konijnenburg, M., & Van Helleputte, N. (2019). CorNET: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment. *IEEE Transactions on Biomedical Circuits and Systems*, 13(2), 282–291. <https://doi.org/10.1109/TBCAS.2019.2892297>
6. Charlton, P. H., Birrenkott, D. A., Bonnici, T., Pimentel, M. A. F., Johnson, A. E. W., Alastruey, J., Tarassenko, L., Watkinson, P. J., Beale, R., & Clifton, D. A. (2018). Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. *IEEE Reviews in Biomedical Engineering*, 11, 2–20. <https://doi.org/10.1109/RBME.2017.2763681>

7. Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 115.
<https://doi.org/10.3390/s16010115>
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).
9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
10. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
11. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.
arXiv:1412.6980
12. 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.
13. Karlen, W., Raman, S., Ansermino, J. M., & Dumont, G. A. (2013). Multiparameter respiratory rate estimation from the photoplethysmogram. *IEEE Transactions on Biomedical Engineering*, 60(7), 1946–1953.
14. Radu, V., Tong, C., Bhattacharya, S., Lane, N. D., Mascolo, C., Marina, M. K., & Kawsar, F. (2018). Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4), 157.
15. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
16. Sjöding, 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.
17. Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507.
18. 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.
19. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134*.
20. Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer.
21. World Health Organization. (2021). *Global strategy on digital health 2020–2025*. World Health Organization.

22. Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3), R1–R39.